Symposium Proceedings

DRAFT

52nd International Symposium on Operational Research

www.symopis2025.fon.bg.ac.rs

SYM-OP-IS 2025

Palić, 7– 10 September 2025.

SYMPOSIUM PROCEEDINGS

(DRAFT VERSION)

DRAFT VERSION

Editors:

Prof. dr Marija Kuzmanović
Prof. dr Dragana Makajić-Nikolić
Prof. dr Bisera Andrić Gušavac

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

EXECUTIVE ORGANIZER

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

MAIN ORGANIZER

DOPIS
Operational Research Society of Serbia

CO-ORGANIZERS

Faculty of Transport and Traffic Engineering, Belgrade

Faculty of Mathematics, Belgrade

Mathematical Institute, Belgrade

Faculty of Economics and Business, Belgrade

Ministry of defence Republic of Serbia

The Serbian Armed Forces

Faculty of Technical Sciences, Novi Sad

University of Banja Luka

Faculty of Mining and Geology, Belgrade

School of Civil Engineering and Geodesy, Belgrade

Institute Mihailo Pupin, Belgrade

Economics Institute

SUPPORTED BY

Република Србија
Министарство просвете,
науке и технолошког развоја

АКАДЕМИЈА
ИНЖЕЊЕРСКИХ НАУКА
СРБИЈЕ

DRAFT VERSION

PROGRAMME COMMITTEE

Kuzmanović Marija, PC Chair, University of Belgrade, Faculty of Organizational Sciences, Serbia
Makajić-Nikolić Dragana, Vice PC Chair, University of Belgrade, Faculty of Organizational Sciences, Serbia
Aleksić Vule, University of Belgrade, College of Civil Engineering and Geodesy, Serbia
Banković Radoje, University of Belgrade, Military Geographic Institute, Serbia
Bjelić Nenad, University of Belgrade, Faculty of Transport and Traffic Engineering, Serbia
Ćirović Goran, University of Novi Sad, Faculty of Technical Sciences, Serbia
Davidović Tatjana, University of Belgrade, Mathematical Institute SANU, Serbia
Delibašić Boris, University of Belgrade, Faculty of Organizational Sciences, Serbia
Dimitrijević Branka, University of Belgrade, Faculty of Transport and Traffic Engineering, Serbia
Dražić Milan, University of Belgrade, Faculty of Mathematics, Serbia
Dugošija Đorđe, State University of Novi Pazar, Serbia
Đorović Boban, University of Defense, Serbia
Janković Irena, University of Belgrade, Faculty of Economics, Serbia
Kočović Jelena, University of Belgrade, Faculty of Economics, Serbia
Kratica Jozef, University of Belgrade, Mathematical Institute SANU, Serbia
Kutlača Đuro, University of Belgrade, Mihajlo Pupin Institute, Serbia
Martić Milan, University of Belgrade, Faculty of Organizational Sciences, Serbia
Miljanović Igor, University of Belgrade, Faculty of Mining and Geology, Serbia
Mladenović Zorica, University of Belgrade, Faculty of Economics, Serbia
Mučenski Vladimir, University of Novi Sad, Faculty of Technical Sciences, Serbia
Netjasov Feđa, University of Belgrade, Faculty of Transport and Traffic Engineering, Serbia
Nikolić Dragan, University of Belgrade, College of Civil Engineering and Geodesy, Serbia
Nikolić Miloš, University of Belgrade, Faculty of Transport and Traffic Engineering, Serbia
Pamučar Dragan, University of Belgrade, Faculty of Organizational Sciences, Serbia
Papuga Saša, University of Banja Luka, Faculty of Technology, Bosnia and Herzegovina
Peško Igor, University of Novi Sad, Faculty of Technical Sciences, Serbia
Petrović Dalibor, Ministry of Defense, Serbia
Petrović Nataša, University of Belgrade, Faculty of Organizational Sciences, Serbia
Praštalo Željko, University of Belgrade, Mining Institute, Serbia
Preradović Ljubiša, University of Banja Luka, Faculty of Architecture, Civil Engineering and Geodesy, BiH
Radojević Dragan, University of Belgrade, Mihajlo Pupin Institute, Serbia
Savić Aleksandar, University of Belgrade, Faculty of Mathematics, Serbia
Savić Gordana, University of Belgrade, Faculty of Organizational Sciences, Serbia
Simić Dejan, University of Belgrade, Faculty of Organizational Sciences, Serbia
Stamenković Mladen, University of Belgrade, Faculty of Economics, Serbia
Stanić Stanko, University of Banja Luka, Faculty of Economics, Bosnia and Herzegovina
Stanimirović Zorica, University of Belgrade, Faculty of Mathematics, Serbia
Stanojević Milan, University of Belgrade, Faculty of Organizational Sciences, Serbia
Stojković Dejan, Ministry of Defense, Serbia
Suknović Milija, University of Belgrade, Faculty of Organizational Sciences, Serbia
Šelmić Milica, University of Belgrade, Faculty of Transport and Traffic Engineering, Serbia
Urošević Dragan, University of Belgrade, Mathematical Institute of SANU, Serbia
Vujošević Mirko, University of Belgrade, Faculty of Organizational Sciences, Serbia
Aloise Daniel, Federal University of Rio Grande do Norte, Brazil
Artiba Abdelhakim, University of Valenciennes, France

DRAFT VERSION

Carrizosa Emilio, University of Seville, Spain
Duarte Abraham, King Juan Carlos University, Spain
Eremeev Anton, Omsk State University, Russia
Escudero Laureano, Miguel Hernández University, Spain
Hanafi Saïd, University of Valenciennes, France
Hudec Miroslav, University of Economics in Bratislava, Slovakia
Kochetov Yuri, Sobolev Institute of Mathematics, Novosibirsk, Russia
Labbé Martine, Free University of Brussels, Belgium
Melián Batista Belén, University of La Laguna, Spain
Migdalas Athanasios, Aristotle University of Thessaloniki, Greece
Moreno-Perez José A., University of La Laguna, Spain
Papageorgiou Markos, Technical University of Crete, Greece
Raidl Guinter, Vienna University of Technology, Austria
Salhi Said, University of Kent, United Kingdom
Sevaux Marc, University of Southern Brittany, France
Sifaleras Angelo, University of Macedonia, Greece
Sörensen Kenneth, University of Antwerp, Belgium
Velkovski Trajče, Ss. Cyril and Methodius University, Skopje, North Macedonia

DRAFT VERSION

HONORARY PROGRAMME COMMITTEE

Andrejić Marko, University of Defence, Military Academy, Serbia
Backović Marko, University of Belgrade, Faculty of Economics, Serbia
Batanović Vladan, University of Belgrade, Mihajlo Pupin Institute, Serbia
Borović Siniša, University of Belgrade, Faculty of Medical Sciences and Sports, Serbia
Vujić Slobodan, University of Belgrade, Mining Institute, Serbia
Vuleta Jovo, University of Belgrade, Faculty of Economics, Serbia
Đorđević Branislav, University of Belgrade, Faculty of Civil Engineering, Serbia
Guberinić Slobodan, University of Belgrade, Mihajlo Pupin Institute, Serbia
Ilić Aleksandar, University of Defence, Ministry of Defence, Serbia
Kovač Mitar, University of Defence, Ministry of Defence, Serbia
Kovačević-Vujčić Vera, University of Belgrade, Faculty of Organizational Sciences, Serbia
Mesaroš Katalin, University of Novi Sad, Faculty of Economics, Serbia
Mihaljević Miodrag, University of Belgrade, Mathematical Institute of SANU, Serbia
Milovanović Gradimir, Serbian Academy of Sciences and Arts (SANU), Serbia
Mučibabić Spasoje, University of Defence, Ministry of Defence, Serbia
Ognjanović Zoran, University of Belgrade, Mathematical Institute of SANU, Serbia
Opricović Serafim, University of Belgrade, Faculty of Civil Engineering, Serbia
Pap Endre, Singidunum University, Serbia
Petrović Slavica, University of Kragujevac, Faculty of Economics, Serbia
Rakić Milan, University of Belgrade, Mihajlo Pupin Institute, Serbia
Stanojević Milorad, University of Belgrade, Faculty of Transport and Traffic Engineering, Serbia
Starčević Dušan, University of Belgrade, Faculty of Organizational Sciences, Serbia
Cvetković Dragoš, Serbian Academy of Sciences and Arts (SANU), Serbia
Čabarkapa Obrad, University of Defence, Ministry of Defence, Serbia
Čangalović Mirjana, University of Belgrade, Faculty of Organizational Sciences, Serbia
Zečević Tomislav, University of Belgrade, Faculty of Economics, Serbia

ORGANIZING COMMITTEE

Bisera Andrić Gušavac, OC Chair, University of Belgrade, Faculty of Organizational Sciences, Serbia
Biljana Panić, Vice OC Chair, University of Belgrade, Faculty of Organizational Sciences, Serbia
Ognjen Nikolić, Technical Secretary, University of Belgrade, Faculty of Organizational Sciences, Serbia
Andrijana Džamić, University of Belgrade, Faculty of Organizational Sciences, Serbia
Isidora Gaćeša, University of Belgrade, Faculty of Organizational Sciences, Serbia
Pavle Milošević, University of Belgrade, Faculty of Organizational Sciences, Serbia
Nenad Medić, University of Novi Sad, Faculty of Technical Sciences, Serbia
Ivana Jovanović, University of Belgrade, Faculty of Transport and Traffic Engineering, Serbia

DRAFT VERSION

TABLE OF CONTENT

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

APPLICATIONS OF OR IN DEFENSE 1

APPLICATION OF THE AHP METHOD FOR SELECTING THE PROCEDURE FOR REMOVING POST-DETONATION REMNANTS OF EXPLOSIVES FROM THE ENVIRONMENT

Zoran Bajić, Boško Vujković, Vladimir Ignjatović, Jovica Bogdanov 2

METHODOLOGY FOR DEFINING CRITERIA FOR ANALYZING THE ROLE OF THE MILITARY AND THE MILITARY PROFESSION IN CONTEMPORARY SOCIETY: THE FUZZY EXTENSION OF THE AHP METHOD

Dragan Bojanić, Marina Bojanić 8

THE SIGNIFICANCE OF PSYCHO-PHYSICAL CHARACTERISTICS OF DRIVER CANDIDATES FOR PREDICTING SUCCESS IN DRIVER TRAINING

Srđan Dimić, Jelena Vicanović, Srđan Ljubojević, Momčilo Đorđević 14

ARTIFICIAL INTELLIGENCE 20

PRODUCT PROFESSIONALS' PERSPECTIVES ON TRADITIONAL AND AI-ENHANCED PRODUCT BACKLOG PRIORITIZATION APPROACHES

Miloš Belčević, Dragan Pamučar 21

SHAPE-BASED IMAGE CLASSIFICATION USING HU AND ZERNIKE MOMENTS WITH BEE COLONY OPTIMIZATION FOR TUNING SVM PARAMETERS

Dušan Denčić, Nataša Milosavljević, Tatjana Davidović 26

LEVERAGING ARTIFICIAL INTELLIGENCE FOR ENHANCED E-COMMERCE PERFORMANCE

Petar Ilić, Jelica Stanojević, Marija Bogićević Sretenović 32

RED TEAMING GENERATIVE AI APPLICATIONS: THREAT MODELING AND MITIGATION STRATEGIES

Matea Lukić, Ana Poledica, Pavle Milošević 38

AI-BASED DEMAND FORECASTING FOR INVENTORY MANAGEMENT

Dušan Mitrović, Miroslav Minović 44

AUTONOMOUS AI SCIENTISTS: A NEW ERA OF SCIENTIFIC DISCOVERY

Jelena Novaković 50

DL-SPP: A DEEP LEARNING SYSTEM FOR SPARE PARTS PREDICTION IN FIELD SERVICE OPERATIONS

Alessio Schiavo, Alessandro Renda, Pietro Ducange, Francesco Marcelloni 56

DRAFT VERSION

CHALLENGES OF MEASURING CONVERSION RATE IN CONDITIONS OF AI GENERATED PRODUCT SEARCH
Sonja Vučenović, Goran Vukmirović, Nikola Macura 62

EXTENDING THE INTELLIGENT ENTERPRISE MODEL (IEM) FOR SMALL AND MEDIUM-SIZED ENTERPRISES BY INTEGRATING GENERATIVE ARTIFICIAL INTELLIGENCE
Dragan Vukmirović, Željko Bolbotinović, Tijana Čomić, Ranka Popovac 67

DATA SCIENCE **73**

A MODIFIED DBSCAN ALGORITHM FOR TRAFFIC ACCIDENT CLUSTERING
Jordan Atanasijević, Danijela Protić, Dimitrije Kolašinac, Stefan Ivanović 74

CLUSTER ANALYSIS OF CONSUMER ATTITUDES TOWARDS ECO-FRIENDLY PRODUCTS IN WESTERN BALKANS: AN OPERATIONS RESEARCH APPROACH
Bojan Kitanovikj, Violeta Cvetkoska, Alek Pachemski 80

E-BUSINESS **86**

THE ROLE OF S-COMMERCE IN HIGHER EDUCATION: ANALYZING THE PRESENCE OF SERBIAN FACULTIES ON TIKTOK
Sara Kostić, Jelena Spajić, Bojana Milić, Dunja Bošković, Danijela Lalić 87

SMART RAILWAY AS A SMART CITY SERVICE
Danijela Stojanović, Nenad Stanisavljević, Petar Lukovac 93

ECONOMIC MODELS AND ECONOMETRICS **99**

REVISITING THE TOURISM-LED ECONOMIC GROWTH HYPOTHESIS: EMPIRICAL INSIGHTS FROM SERBIA
Aleksandra Anić, Branislava Hristov Stančić, Jelena Rašković 100

THE CRISIS IN THE EUROPEAN AUTOMOTIVE INDUSTRY AND ITS IMPACT ON SERBIA'S ECONOMY
Mirjana Gligorić Matić, Dušan Marković 106

DEMAND RATIONING RULES: BEYOND EFFICIENCY AND PROPORTIONALITY
Bojan Ristić, Nikola Njegovan 112

DRAFT VERSION

ENVIRONMENTAL MANAGEMENT AND MANAGEMENT OF NATURAL RESOURCES 118

TREND ANALYSIS OF PRECIPITATION AND AIR TEMPERATURE OBSERVED AT THE METEOROLOGICAL STATION BELGRADE
Jelisaveta Vakanjac, Aleksandar Tanasković, Jugoslav Nikolić, Boris Vakanjac, Vesna Ristić Vakanjac *119*

ECOLOGICAL RESPONSIBILITY AND ENVIRONMENTAL ACTIVISM IN KRUŠEVAC: RESIDENTS' ATTITUDES
Milan Zajić , Dušan Garabinović *125*

FINANCE AND BANKING 131

ECONOMICS OF HAPPINESS: THE RELATIONSHIP BETWEEN GDP AND SUBJECTIVE WELL-BEING
Dragana Draganac, Ivana Ivković *132*

GAME THEORY 138

MORAL HAZARD IN THE 2008 FINANCIAL CRISIS
Đorđe Mirković, Ognjen Nikolić, Marija Kuzmanović *139*

GEOINFORMATION SYSTEMS 145

ESTABLISHMENT OF A BURIAL SITE REGISTRY
Dragan Đurđević, Biljana Kuridža, Dragana Skorup, Miroslav Vujasinović, Goran Marinković *146*

GEOSPATIAL ANALYSIS OF RADIOACTIVITY DISTRIBUTION IN ALLUVIAL SOILS OF THE KOLUBARA RIVER BASIN AT 20 CM DEPTH
Vuk Gajić, Boris Vakanjac, Nikola Milić *152*

DRONE SURVEYING FOR THE RESTORATION OF HISTORIC BUILDINGS
Nenad Galjak, Jelena Kolorogić, Sladjana Stanišić, Miroslav Vujasinović, Miodrag Regodić *157*

OPTIMIZATION OF PRECISE EDM GEODETIC CALIBRATION BASELINE
Siniša Drobnjak, Srđan Stojaković, Vojkan Stanojević, Radoje Banković *163*

DETERMINATION OF THE POSITION SOLAR MODULES USING UAV PHOTOGRAMMETRY AND COMPUTER VISION
Nikola Santrač, Tatjana Budimirov, Đuro Krnić, Mehmed Batilović, Vladimir Bulatović *169*

DRAFT VERSION

SPATIAL PLANNING IN THE REPUBLIC OF CROATIA
Bojan Linardić, Vladimir Baričević, Luka Babić, Ivan Landek, Robert Paj *175*

PYTHON-BASED AUTOMATION OF GEOPROCESSING IN DIGITAL TOPOGRAPHIC MAP PRODUCTION: CONTENT ALIGNMENT BETWEEN MAP SHEET BORDERS
Marko Simić, Mihajlo Bobar, Saša Bakrač *181*

REMOTE MONITORING OF WATER SURFACE POLLUTION
Marko Stojanović, Mirjana Kokotović, Slađana Stanišić, Miodrag Regodić *187*

GRAPHS AND NETWORKS 193

VORONOI DIAGRAM AS A TOOL FOR ANALYSING DEVELOPMENT IN RAILWAY AND AIR TRANSPORT INFRASTRUCTURE: A EUROPEAN PERSPECTIVE
Martin Fale, Tea Vizinger, Bojan Rupnik, Tomaž Kramberger *194*

SIGNED TOTAL ROMAN DOMINATION OF COMPLETE SPLIT GRAPHS
Aleksandar Savić, Zoran Maksimović *200*

DISTANCE-EDGE-MONITORING NUMBER OF COMPLETE SPLIT GRAPHS
Milica Milivojević Danas, Jozef Kratica *205*

HEURISTICS 209

MULTI-START LOCAL SEARCH FOR SUGAR BEET TRANSPORTATION PROBLEM WITH EQUITY REGULATIONS FOR GROWERS
Ana Anokić, Đorđe Stakić, Dragana Drenovac, Tatjana Davidović *210*

VNS APPROACH FOR TOTAL TARDINESS MINIMIZATION IN A SINGLE MACHINE SCHEDULING PROBLEM WITH PERIODIC RESOURCE CONSTRAINTS
Maja Ristić, Zorica Dražić *216*

VARIABLE NEIGHBORHOOD SEARCH FOR THE P-NEXT CENTER PROBLEM WITH DISCOUNT FACTOR
Jelena Tasić, Zorica Stanimirović, Zorica Dražić *222*

COMPARATIVE STUDY OF A BASIC AND MODIFIED MEMETIC ALGORITHM FOR THE ELECTRIC VEHICLE ROUTING PROBLEM
Luka Matijević *228*

DRAFT VERSION

INFORMATION SYSTEMS AND TECHNOLOGY 234

MODELING MONO AND MULTI-REPOSITORY DILEMMA FOR MICROSERVICE SOFTWARE ARCHITECTURE DEVELOPMENT
Radovan Čikić, Miloš Milić, Siniša Vlajić 235

DIRECTION OF ARRIVAL ESTIMATION IN FMCW RADAR USING THE MUSIC-2D ALGORITHM
Borko Đaković, Slobodan Simić, Danijela Protić, Kristina Živanović 241

DETERMINING OPTIMAL REQUEST RATE TO MAINTAIN THERMAL EQUILIBRIUM IN SERVICE-ORIENTED SOFTWARE ARCHITECTURE
Miloš Milić, Nebojša Nikolić, Dragana Makajić-Nikolić 247

A HYBRID APPROACH TO DATA MANAGEMENT: INTEGRATING JSON DOCUMENTS INTO RELATIONAL DATABASES
Mihajlo Pavlović, Tatjana Stojanović, Saša Lazarević 253

A SYSTEMATIC REVIEW OF VECTOR DATABASE USE IN RETRIEVAL-AUGMENTED GENERATION FOR LLM-BASED EDUCATIONAL PLATFORMS
Filip Stamenković, Jelica Stanojević, Miroslav Minović 259

LOGISTICS AND SUPPLY CHAINS 265

PRODUCT-FORM SOLUTION OF A QUEUEING INVENTORY WITH TWO MODES OF PRODUCTION
Abdul Rof V, Achyutha Krishnamoorthy 266

QUADRATIC MODEL FOR MULTI-DEPOT VEHICLE ROUTING PROBLEM WITH TIME WINDOWS IN DISASTER MANAGEMENT
Nemanja Golubović, Nenad Bjelić, Branislava Ratković 272

OPTIMIZING THE LOCATION OF MOBILE MAMMOGRAPHY UNITS IN RURAL COMMUNITIES
Predrag Grozdanović, Milica Šelmić, Miloš Nikolić, Dušan Teodorović 278

MANAGEMENT 284

ANALYSES OF THE CONVERGENT TELCO-INSURANCE BUSINESS MODEL: MANAGERIAL PERSPECTIVES IN NORTH MACEDONIA
Metodija Mirchev, Violeta Cvetkoska, Igor Ivanovski 285

CROWDFUNDING ART PROJECTS: A SYSTEMATIC LITERATURE REVIEW
Dragana Pantić, Dragan Bjelica 290

DRAFT VERSION

MATHEMATICAL PROGRAMMING 296

THE CAPACITATED SINGLE ALLOCATION HUB MAXIMAL COVERING PROBLEM WITH HUB INSTALLATION COSTS
Zorica Stanimirović, Olivera Stančić *297*

MINING, GEOLOGY AND ENERGY 303

ANALYSIS OF COAL EXPLOITATION TRENDS
Dragan Milošeić, Branka Jovanović *304*

MULTICRITERIA ANALYSIS AND OPTIMIZATION 310

OPTIMIZING THE TRAJECTORY FOR IMPROVING QUALITY EDUCATION USING SLACK-BASED DATA ENVELOPMENT ANALYSIS
Ana Krstić Srejović, Predrag Mimović *311*

APPLICATION OF THE ANALYTICAL HIERARCHY PROCESS FOR SELECTING A MOTOR VEHICLE FOR TRANING CATEGORY ''C'' DRIVER CANDIDATES IN THE MINISTRY OF DEFENCE AND THE SERBIAN ARMED FORCES
Vesko Lukovac, Luka Ivezić, Pavle Gladović, Vladimir Popović, Milan Stanković *318*

ANYTIME ALGORITHM TO GENERATE NON-DOMINATED VECTORS TO BI-OBJECTIVE INTEGER PROGRAMMING PROBLEMS
Nada Mladenović, Bogdana Stanojević *324*

ADAPTIVE MULTI-CRITERIA OPTIMIZATION OF EV CHARGING DECISIONS USING XGBOOST AND TOPSIS
Jelena Velimirović, Lazar Velimirović, Petar Vranić *329*

OR APPLICATIONS IN CONSTRUCTION 336

FUZZY ADAPTIVE CPM-PERT METHOD
Nataša Popović-Miletić, Goran Ćirović, Dragan Pamučar *337*

THE USE OF DESCRIPTIVE STATISTICS IN ANALYZING THE PERCEPTION OF OCCUPATIONAL RISKS: THE CASE OF SILICA DUST EXPOSURE AMONG CONSTRUCTION WORKERS
Željana Kužet, Vladimir Mučenski, Jovana Topalić *343*

DRAFT VERSION

PERFORMANCE ANALYTICS 349

ADVANTAGES OF THE SUPER EFFICIENCY SBM DEA MODEL
Jelena Novaković 350

PERFORMANCE MANAGEMENT IN SERBIA'S PUBLIC SECTOR: INSIGHTS FROM QUANTITATIVE REGRESSION MODELING
Aleksandra Pavlović, Andrea Ivanišević, Katarina Stojanović 356

RESEARCH & DEVELOPMENT 362

100+ YEARS OF PATENTING IN SERBIA
Đuro Kutlača 363

OPERATIONS RESEARCH IN CIVIL ENGINEERING AND GEODESY ACADEMIC STUDIES IN FORMER YUGOSLAV REPUBLICS
Jovana Topalić, Tiana Milović, Vladimir Mučenski, Mehmed Batilović 369

SIMULATION AND STOCHASTIC MODELS 374

SIMULATION MODEL OF ROBOT MOVEMENT ALONG A DEFINED PATH DURING PRE-FLIGHT INSPECTION OF AN A320 AIRCRAFT
Dunja Savić, Marko Đogatović, Ljubiša Vasov 375

SOFT COMPUTING 381

VARIANTS OF THE MINIMUM ERROR THRESHOLDING METHOD FOR IMAGE SEGMENTATION
Dejan Cebić, Nebojsa Ralević, Nataša Milosavljević 382

LOGIC AGGREGATION BY QUANTIFIER 'MOST OF' FOR ENTITIES EVALUATION
Miroslav Hudec, František Papala, Hanna Kristín Skaftadóttir 390

LEVERAGING GENERATIVE ADVERSARIAL NETWORKS IN PROJECT MANAGEMENT FOR TASK ESTIMATION
Nikola Vojtek, Bojan Smudja, Pavle Milošević, Ivana Dragović 396

APPLICATION OF COMPUTATIONAL INTELLIGENCE BASED CLUSTERING ALGORITHMS ON EDUCATIONAL DATA
Milica Zukanović, Pavle Milošević, Ragnar Már Vilhjálmsson, Haraldur Dadi Ragnarsson, Ana Poledica 402

DRAFT VERSION

STATISTICAL MODELS 408

THEORETICAL AND EMPIRICAL RESULTS ON THE COVERAGE PROBABILITY OF THE STANDARD QUANTILE ESTIMATOR FOR THE BURR DISTRIBUTION

Jelena Stanojević, Marija Koprivica *409*

TRAFFIC, TRANSPORT AND COMMUNICATION 415

FUEL CONSUMPTION AND CO2 EMISSION FORECAST ON DIVERGING DIAMOND INTERCHANGE

Aleksandar Jovanović, Katarina Kukić, Ana Uzelac *416*

SELECTION OF CITY LOGISTICS DEVELOPMENT SCENARIOS BY APPLYING A HYBRID MCDM MODEL

Snežana Tadić, Mladen Krstić, Aleksa Maravić, Miloš Veljović *422*

DRAFT VERSION

APPLICATIONS OF OR IN DEFENSE

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

APPLICATION OF THE AHP METHOD FOR SELECTING THE PROCEDURE FOR REMOVING POST-DETONATION REMNANTS OF EXPLOSIVES FROM THE ENVIRONMENT

ZORAN BAJIĆ[1], BOŠKO VUJKOVIĆ[2], VLADIMIR IGNJATOVIĆ[1,2], JOVICA BOGDANOV[3]

[1] Univerzitet odbrane, Vojna akademija, Beograd, zoran.bajic@va.mod.gov.rs, 0000-0002-8492-3333
[2] Vojska Srbije, Uprava za logistiku, Beograd, bole199555@gmail.com, 0009-0008-2529-2608, wllaxy@yahoo.com 0009-0009-5253-3224
[3] Univerzitet odbrane, Beograd, jovica.bogdanov@va.mod.gov.rs, 0000-0001-7995-3004

Abstract: *The subject and goal of this paper are focused on areas contaminated by military activities, as well as providing answers regarding the selection of the best way for removing explosive material (EM) residues from polluted areas using the AHP multi-criteria decision-making model. Military training grounds play a crucial role in the training and testing of weapons and ammunition, but they are also a significant source of environmental pollution. During military activities, explosive materials enter the environment through the activation of explosive ordnance (XO). This typically occurs due to weapon firing, unexploded XO remnants during detonation, improper storage, and destruction of XO. Explosive materials contaminate soil and groundwater and, through leaching, can further spread into other ecosystems.*

Keywords: *MCDM, AHP method, environment, explosives.*

DRAFT VERSION

1. INTRODUCTION

Military training and testing activities play a crucial role in national defense, but they also contribute to environmental contamination, particularly through the release of explosive materials into soil and groundwater. Explosives such as trinitrotoluene (TNT), cyclotrimethylenetrinitramine (RDX), and octogen (HMX) are widely used in military operations, yet their environmental persistence and toxicity raise significant concerns (Bernstein & Ronen, 2012). The contamination of military ranges occurs due to munitions activation, unexploded ordnance, improper disposal, and degradation of explosive residues, leading to long-term ecological effects (Brannon & Myers, 1997; Juhasz & Naidu, 2007).

The environmental fate of these explosives varies based on their chemical properties, soil composition, and biodegradation potential. Studies have shown that explosive residues can persist in terrestrial and aquatic ecosystems, posing risks to human health and wildlife (Kalderis et al., 2011). Furthermore, improper remediation efforts may exacerbate contamination rather than eliminate it. Various remediation strategies, including chemical treatment, soil excavation, and biodegradation, have been proposed to address the problem (McGrath, 1995). A systematic approach using multi-criteria decision analysis has been introduced to determine optimal removal routes for RDX and HMX in contaminated sites (Bajić et al., 2024).

Given the widespread environmental hazards associated with explosives in military ranges, identifying efficient and sustainable remediation methods is imperative. This study employs the Analytical Hierarchy Process (AHP) model to evaluate the effectiveness of various removal techniques, aiming to mitigate the impact of explosive contaminants and ensure ecosystem resilience (Brannon & Pennington, 2012). By integrating multi-criteria decision-making principles, this research seeks to provide actionable insights for environmental management and pollution mitigation in military training areas.

2. DEVELOPMENT OF MULTICRITERIA DECISION MODEL

The selection of appropriate environmental remediation technologies is fundamentally influenced by various site-specific factors, including geographic location, soil composition, water body characteristics, contaminant type and concentration, as well as local geological and climatic conditions. Additionally, remediation urgency indicators—such as proximity to residential areas, contaminant toxicity, and intended future land use—must be taken into account when planning a remediation project. Prior to implementing any

remediation measures, comprehensive site investigations and urgency assessments should be conducted. The gathered data is then reviewed by experts, who evaluate possible alternatives based on technical explanations and relevant recommendations. Typically, multiple remediation options are available, and determining the optimal approach requires a scientifically developed decision-making model based on suitable MCDM methods.

During expert evaluations, three primary criteria are used to assess the sustainability of recommended remediation strategies: economic, environmental, and technological impacts. Economic considerations include factors such as equipment costs, operational expenses, and costs associated with detection and analysis. Environmental factors encompass risks of secondary contamination, potential hazards to human health, and public acceptance of the remediation approach. The technological aspect is evaluated based on remediation efficiency, required implementation time, and the maturity of the technology (Li et al., 2018, Bajić et al., 2024). Detailed overview of the technological data relevant to explosives as environmental pollutants was given elsewhere (Bajić et al., 2024).

When determining the most effective remediation technology, factors beyond efficiency, technological maturity, and cost must be taken into account, including:

- Physicochemical characteristics of explosive materials
- Extent and intensity of soil and water contamination
- Geological and hydrogeological characteristics of the affected area
- Previous applications of remediation technologies in similar cases
- Simplicity of methods, accessibility, and feasibility of equipment usage
- Potential secondary environmental impacts and mitigation approaches
- Risks to human life and health
- Social acceptability of the remediation approach
- Economic, technical, and time constraints

In addressing complex and large-scale problems, professional experts from relevant fields are frequently consulted to analyze and discuss potential solutions. Their insights serve as valuable references in the decision-making process. However, treating all expert opinions equally without considering variations in expertise—such as quality, knowledge, experience, and ability—can lead to an imbalanced assessment. To ensure a more rational and scientifically grounded decision, it is essential to first carefully evaluate the relative weight of each expert's input (Bajić et al., 2024).

3. RESULTS

The expert competency model was used to rank experts based on the criteria presented elsewhere (Bajić et al., 2024). Their ranks are shown in Table 1.

Table 1: Comparison matrix of experts based on given criteria

Rank	Expert	Weight coefficient of each expert
1	E1	0.3038
2	E2	0.2830
3	E3, E5	0.1415
4	E4	0.1261

During the survey, the most optimal remediation technologies (alternatives) for the given case were proposed, and experts expressed their opinions using quantitative comparisons of criteria and alternatives within the designated software. The criteria included:

- Cost (C1)
- Remediation burden (C2)
- Remediation efficiency (C3)
- Operability (C4)
- Technology maturity (C5)

The proposed alternatives were:

- Photolysis (A1)
- Redox reactions (A2)
- Excavation (A3)

DRAFT VERSION

- Solidification (A4)
- Ultrasonication (A5)
- Incineration (A6)
- Thermal desorption (A7)

Final comparison matrices, based on the opinions of each expert, are presented in Tables 2 to 7.

Table 2: Criteria comparison matrix

Criteria	C1	C2	C3	C4	C5
C1	1	2.6458	0.189	5	3
C2	0.378	1	0.126	1.291	0.2236
C3	5.2915	7.9373	1	7.4833	4.4721
C4	0.2	0.7746	0.1336	1	0.1543
C5	0.3333	4.4721	0.2236	6.4807	1

Table 3: Alternative comparison matrix for criterion C1

C1	A1	A2	A3	A4	A5	A6	A7
A1	1	5.9161	7.4833	1.2247	3.1623	1.5811	1.5811
A2	0.169	1	1.4142	1	1.1547	1.4142	0.4082
A3	0.1336	0.7071	1	0.2887	0.7746	0.866	0.866
A4	0.8165	1	3.4641	1	1	1.1547	0.866
A5	0.3162	0.866	1.291	1	1	1.2247	1
A6	0.6325	0.7071	1.1547	0.866	0.8165	1	0.3536
A7	0.6325	2.4495	1.1547	1.1547	1	2.8284	1

Table 4: Alternative comparison matrix for criterion C2

C2	A1	A2	A3	A4	A5	A6	A7
A1	1	1	6.4807	4.4721	1.1547	1.1547	1.2247
A2	1	1	4.2426	4.4721	0.2582	0.866	1
A3	0.1543	0.2357	1	0.25	0.2582	0.25	0.2236
A4	0.2236	0.2236	4	1	0.2	0.2582	0.2582
A5	0.866	3.873	3.873	5	1	4.4721	4
A6	0.866	1.1547	4	3.873	0.2236	1	0.4472
A7	0.8165	1	4.4721	3.873	0.25	2.2361	1

Table 5: Alternative comparison matrix for criterion C3

C3	A1	A2	A3	A4	A5	A6	A7
A1	1	4.899	2.6458	4.2426	4.5826	3.1623	3.1623
A2	0.2041	1	1.118	1	1.291	1.291	1.118
A3	0.378	0.8944	1	0.3162	0.9129	0.7746	0.4082
A4	0.2357	1	3.1623	1	1	0.3536	0.2887
A5	0.2182	0.7746	1.0954	1	1	1	1
A6	0.3162	0.7746	1.291	2.8284	1	1	0.4472
A7	0.3162	0.8944	2.4495	3.4641	1	2.2361	1

DRAFT VERSION

Table 6: Alternative comparison matrix for criterion C4

C4	A1	A2	A3	A4	A5	A6	A7
A1	1	3.873	4.5826	3.4641	2.4495	2.2361	2.2361
A2	0.2582	1	2.8284	2.2361	0.8165	2.4495	1.7321
A3	0.2182	0.3536	1	0.2887	0.5774	0.2887	0.2887
A4	0.2887	0.4472	3.4641	1	0.7746	2	0.5774
A5	0.4082	1.2247	1.7321	1.291	1	1.291	1.291
A6	0.4472	0.4082	3.4641	0.5	0.7746	1	0.4472
A7	0.4472	0.5774	3.4641	1.7321	0.7746	2.2361	1

Table 7: Alternative comparison matrix for criterion C5

7	A1	A2	A3	A4	A5	A6	A7
A1	1	1.4142	2.6458	3.1623	1.1547	1.7321	1.5811
A2	0.7071	1	4.2426	3.1623	1.291	2.4495	2.2361
A3	0.378	0.2357	1	0.3162	0.8165	0.3162	0.7071
A4	0.3162	0.3162	3.1623	1	2.2361	0.5774	0.5774
A5	0.866	0.7746	1.2247	0.4472	1	1.2247	1.118
A6	0.5774	0.4082	3.1623	1.7321	0.8165	1	0.5774
A7	0.6325	0.4472	1.4142	1.7321	0.8944	1.7321	1

Each comparison matrix of alternatives relative to a given criterion has been normalized, and the weighting coefficients of the alternatives concerning that criterion have been obtained. Based on the criterion comparison matrix, the weighting coefficients for the criteria were obtained (Figure 1), which, when multiplied with the alternative-criterion matrix, provide the final results of the model (Table 8 and Figure 2).

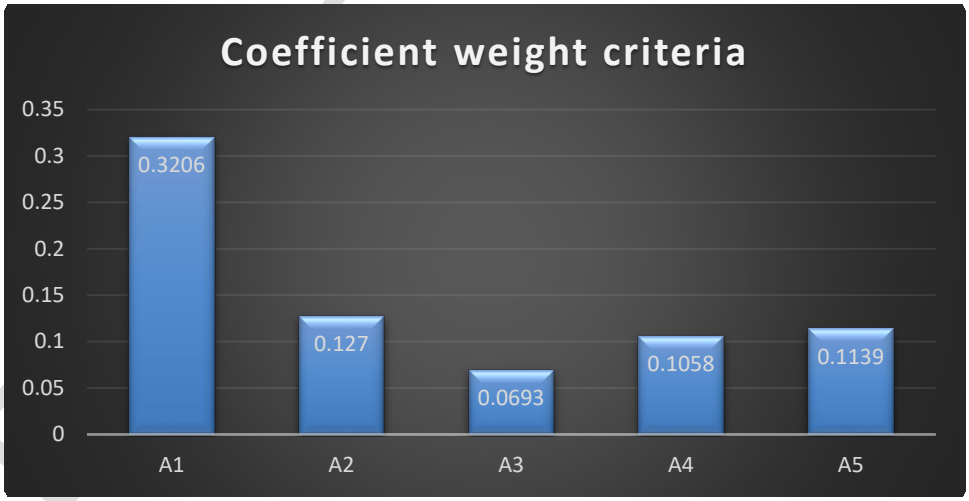


Figure 1: Graphical representation of the obtained values of the weighting coefficients for the criteria

PRE-VERSION

Table 8: Alternative rankings

	C1	C2	C3	C4	C5	Result
A1	0.0635	0.0099	0.2011	0.0128	0.0332	**0.3206**
A2	0.0196	0.007	0.0583	0.0064	0.0358	**0.127**
A3	0.0144	0.0017	0.0418	0.0019	0.0094	**0.0693**
A4	0.0299	0.0026	0.0512	0.0043	0.0177	**0.1058**
A5	0.0219	0.017	0.0511	0.0053	0.0185	**0.1139**
A6	0.0194	0.0061	0.0607	0.0037	0.0185	**0.1085**
A7	0.0333	0.0076	0.0887	0.0056	0.0199	**0.1551**

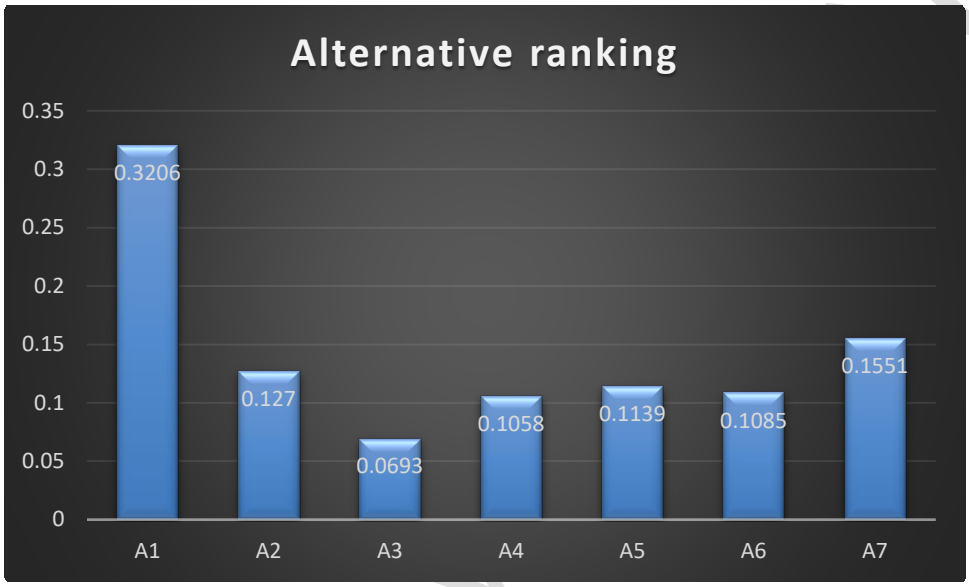


Figure 2: Graphical representation of the alternative ranking

The application of the Analytic Hierarchy Process and the expert competency model led to the clear identification of A1 (photolysis) as the most suitable option for the remediation of EM from the contaminated sites. The application of the expert competency model proved to be appropriate and justified, despite encountering significant challenges—namely, the limited number of available experts and their insufficient experience with issues related to EM contamination, remediation pathways, and relevant techniques. These difficulties are largely due to the fact that the problem has not been clearly addressed in Serbia, and suitable remediation technologies are still not developed or widely implemented in the country. Although photolysis is not a new method, it has yet to be fully confirmed as a successful and widely viable solution. Nevertheless, it ranked highest due to its demonstrated benefits. Correspondingly, the consistency of the pairwise comparisons and the significant margin in priority weights among the alternatives indicate a stable decision outcome.

4. CONCLUSION

This paper describes a multi-criteria decision-making model based on the Analytical Hierarchy Process (AHP) method. The model has been adapted and modified to minimize the possibility of errors when comparing criteria and alternatives. If errors occur, the model identifies them as outlier values and applies appropriate procedures to correct them.

In addition to the multi-criteria decision-making model, an expert competency model has been integrated. This competency model ranks the "weight" (importance) of experts based on specific criteria. These criteria are established by decision-makers and, in this case, pertain to the academic and professional development and performance of experts.

The application of the modified AHP model in the remediation of explosive materials has resulted in the selection of the best remediation technology—photolysis—based on the overall expert consensus. The model has demonstrated its effectiveness in environmental remediation and can also be applied in other fields such as transportation, market analysis, human resources, and more.

DRAFT

Future development of the multi-criteria decision-making model should be accompanied by the implementation of new mathematical models, leveraging its application for decision-making processes that require complex analysis.

ACKNOWLEDGMENT

The authors are grateful for the financial support from University of Defence, Belgrade, VA-TT/1/22-24.

.

LITERATURE

Bajić, Z., Bogdanov, J., & Vujković, B. (2024). The development of a multi-criteria model for the best removal route of post-detonation RDX and HMX explosive remnants in the environment. *SYMOPIS 2024*.

Bernstein, A., & Ronen, Z. (2012). Biodegradation of the explosives TNT, RDX, and HMX. In *Microbial degradation of xenobiotics* (pp. 135-176). Springer.

Brannon, J. M., & Pennington, J. C. (2012). Environmental fate and transport process descriptors for explosives. *ERDC/EL TR-02-10*. U.S. Army Engineer Research and Development Center.

Brannon, M., & Myers, T. E. (1997). Review of fate and transport processes of explosives. *Technical Report IRRP-97-2*. U.S. Army Corps of Engineers.

Juhasz, A. L., & Naidu, R. (2007). Explosives: Fate, dynamics, and ecological impact in terrestrial and marine environments. *Reviews of Environmental Contamination and Toxicology, 191*, 163-215.

Kalderis, D., Juhasz, A. L., Boopathy, R., & Comfort, S. (2011). Soils contaminated with explosives: Environmental fate and evaluation of state-of-the-art remediation processes (IUPAC Technical Report). *Pure and Applied Chemistry, 83*(7), 1407-1484.

Li, X., Li, J., Sui, H., He, L., Cao, X., & Li, Y. (2018). Evaluation and determination of soil remediation schemes using a modified AHP model and its application in a contaminated coking plant. *Journal of Hazardous Materials, 353*, 300-311.

McGrath, C. J. (1995). *Technical Report IRRP-95-2*. U.S. Army Corps of Engineers, Waterways.

Veličković, Z., Bajić, Z., Karkalić, R., Ivanković, N., Gigović, Lj., Kričak, Đ., Vuruna, M. (2016). Uticaj vojnih aktivnosti na zagađenje zemljišta poligona „Pasuljanske livade“, *Energija, ekonomija, ekologija, No. 1-2*, 148–154.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

METHODOLOGY FOR DEFINING CRITERIA FOR ANALYZING THE ROLE OF THE MILITARY AND THE MILITARY PROFESSION IN CONTEMPORARY SOCIETY: THE FUZZY EXTENSION OF THE AHP METHOD

DRAGAN BOJANIĆ[1], MARINA BOJANIĆ[2]

[1]University of Defence, Belgrade, Strategic Research Institute, ,dragan.bojanic@mod.gov.rs, 0000-0003-0426-4479
[2]SEE „Nikola Tesla“, Pančevo, marbojanic@gmail.com, 0009-0009-3400-568X

Abstract: *Social transformations characteristic of contemporary society have inevitably influenced the military—its organization, missions, societal role, and status. These changes necessitate the continuous transformation of national armed forces. The aim of this paper is to present a methodology for defining criteria for analyzing the position and role of the military and the military profession in modern society.This paper proposes a modification of the Analytic Hierarchy Process (AHP) method that incorporates the level of uncertainty in decision-making, allowing for a degree of belief that may be less than absolute certainty. The approach introduces linguistic expressions corresponding to the optimality of the compared criteria. To determine the weights of the criteria and the evaluation of alternatives, fuzzy numbers are employed, enabling a more flexible and realistic representation of expert judgment under uncertainty.*

Keywords:*. military profession, analysis, decision-making, military sciences, fuzzy AHP,*

1. INTRODUCTION

Given the complexity of the issue, which is influenced by a multitude of diverse factors, their identification and definition are not only desirable but virtually essential. The analytical framework provides a structured approach to data collection and interpretation, guiding researchers toward the appropriate selection of indicators and scientific methods. This ensures that the research process remains coherent and predictable. Moreover, such a framework facilitates the operational definition of the research subject, allowing for the recognition of relevant elements—namely, the societal determinants of the military profession, and more broadly, the internal characteristics of the military organization.

The concept of an "analytical framework" should be understood as a general model that serves to "identify, categorize, and organize those factors considered most relevant for understanding a particular phenomenon" (McGinnis 2011, 167). In the field of political science research, methodologists Milosavljević and Radosavljević propose a "typical cognitive model of the research subject" consisting of six factors, which may also serve, in an adapted form, as a useful model for acquiring scientific insight in the context of this research (Milosavljević & Radosavljević 2006, 74–88).

By applying knowledge from operations research, certain social phenomena can also be quantified, which may facilitate more robust conclusions. At this stage, however, the analysis remains primarily qualitative.The problem context involves considering the future consequences of decisions, the presence of data uncertainty, and the collection of expert insights and opinions. This requires the application of procedures that enable overcoming cognitive biases and the complexity of mental processes.

Within this framework, the assessment of the position and role of the military and the military profession represents a scenario in which it is difficult—often impossible—to identify an optimal solution based on all relevant criteria. This situation is characterized by conflicting opinions, a high degree of uncertainty, and internal complexity, highlighting the necessity of applying multi-criteria decision-making (MCDM) methods.The determination of criteria weights represents one of the fundamental challenges in the development and application of multi-criteria decision analysis (MCDA) models. This issue has been the subject of extensive scientific investigation and methodological development over several decades. In general, existing methodologies for determining the relative importance of criteria can be categorized into two primary groups: subjective and objective approaches.

Subjective approaches rely on the input of decision makers or domain experts who are directly involved in the decision-making process.The AHP, developed by Saaty (1980), is one of the most widely applied pairwise comparison methods and employs a nine-point scale to assess the relative importance of criteria.In this paper,

DRAFT VERSION

AHP was modified using fuzzy numbers, resulting in the Fuzzy AHP (FAHP) method, which was then applied to calculate the weight coefficients of the criteria.In order to achieve a higher level of accuracy and to minimize the influence of subjectivity and inconsistency, the significance of the criteria in this study will be evaluated by multiple decision-makers (experts) from the same professional field. Following the initial assessment, the obtained evaluations will be adjusted by taking into account the normalized values of each participant's individual level of competence. The resulting weighted evaluations will then be aggregated and prepared for further analysis using operational research methods.

2. ELEMENTS OF THE ANALYTICAL FRAMEWORK FOR UNDERSTANDING THE MILITARY AND THE MILITARY PROFESSION

The image of contemporary society in the security context, and its transformation from the modern to the postmodern era, has reflected upon the military and its need for adaptation. However, understanding the global social context is still not sufficient for fully grasping the role and position of the military in contemporary society—and by extension, the military profession.

Therefore, among the key determinants of the military and the military profession—which the author of this paper identifies, and which may be considered elements of an analytical framework—are the following: (1) the political system and legislative framework, (2) national culture and social values, (3) historical experience, (4) the economy and the level of technical-technological development, (5) the quality of military education and training, and (6) international military cooperation.

This is, of course, a conditional classification that is not final and depends on the historical period and the region being observed.

The political system and legislative framework (C1) represent one of the fundamental factors in determining the character of the military and the military profession within a given society. National culture and social values (C2) are essential for understanding various phenomena and processes, as well as for comprehending the functioning of institutions and the behavior of prominent individuals and society as a whole. Historical experience (C3) is an indispensable element for perceiving the true role of the military and the military profession in society. Without a profound understanding of the past, it is impossible to fully grasp the present or reliably anticipate future developments. The state of the economy and the level of technical and technological development (C4) significantly influence both the size and the quality of the armed forces (Radojević et al., 2023). In order to assess the role and significance of the military within a society, it is necessary to examine the proportion of national budget allocated to defense. The quality of military education and training (C5) must ensure that the armed forces are composed of officers and soldiers who are not only capable but also willing to defend the state, its citizens, and the shared values of the community (Hornstra at al., 2024). Only such members of the military profession can earn the respect of their society, as well as the recognition of peers from other armed forces—be they allies or adversaries. International military cooperation (C6) is a component that undoubtedly contributes to the reputation of the national armed forces, both in the eyes of partner states and militaries, and within the domestic public sphere (Vukadinović & Milenković, 2024).

In accordance with the outlined criteria, the aim of this research is to formulate new criteria—based on expert opinions and their aggregation—by applying multi-criteria decision-making methods and techniques for determining the weighting coefficients of criteria, including elements of expert competence assessment, while taking into account the specific characteristics of the field of Military Sciences.

3. METODOLOGY

Establishing criteria for decision-making is a complex and responsible task. Today, decision-making almost always involves multi-criteria optimization, that is, making decisions based on more than one criterion. Various mathematical methods have been developed to address such problems.

To enable a more precise evaluation of criteria, a fuzzy linguistic approach was employed. Fuzzy logic and fuzzy sets provide a mathematically formalized means of representing vagueness, uncertainty, and imprecision.

The Analytic Hierarchy Process (AHP) is a method frequently employed in the field of multi-criteria decision-making. The method is based on decomposing a complex problem into a hierarchy, with the overall goal at the top, followed by criteria, sub-criteria, and alternatives organized across different levels and sub-levels of the hierarchy Figure 1.

A detailed explanation of the method will not be provided here, as its fundamental principles are well-documented in an extensive body of literature (see Liu et al., 2020; Bojanic et al., 2018). In the context of this

DRAFT VERSION

research, the method is of particular importance as it will be used to derive the weighting coefficients for the criteria.

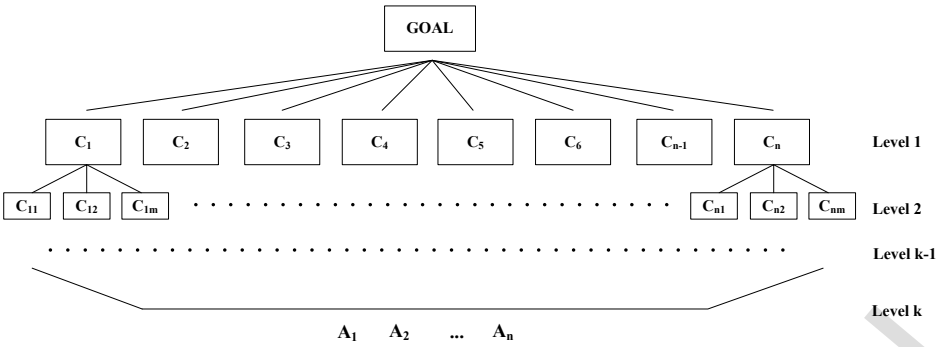


Figure 1. The General Hierarchical Model in the AHP

The methodology for determining the criteria will be applied through the following steps:
Step 1 – Defining the problem and determining the objective.
Step 2 – Identifying the criteria.
Step 3 – Evaluating/comparing the criteria in terms of their importance relative to the objective.
Step 4 – Assessing the consistency of the matrix and calculating the weight vector.
Step 5 – Ranking the criteria.

3.1. Expert evaluation

The first step of the research involved selecting and conducting interviews with experts in order to evaluate the criteria, with the aim of determining the weighting coefficients of the criteria for the application of the fuzzified AHP method. Expert evaluation, as defined by Milićević (2014:18), represents a methodological procedure for obtaining an assessment of a given problem based on the collective (or individual) judgment of subject-matter experts. A critical step in the application of the expert evaluation method is the selection of experts and the assessment of their competence, as it is essential to identify individuals with recognized expertise in the relevant field—experts capable of conducting intuitive-logical analysis of the problem and providing informed judgments. The group of experts was composed of individuals who, through their professional experience, have dealt with the determinants of the military and the military profession. The task of the expert group was to compare the determinants (criteria) in terms of their significance for understanding the role and position of the military—and consequently, the military profession—in contemporary society. A structured interview protocol was developed and applied individually with each expert, following the principles of the Delphi method.

Table 1. Experts qualification

Years of Experience in				**Job Position**					
Description	**Under 10**	**Above 10 and under 20**	**Above 20**	**Military Analyst / Defense Strategist**	**Professor of Military Sciences / Defense Studies**	**Psychologist Specialized Security Institutions**	**International Relations and Security Policy**	**Armed Forces Command**	**Technology and Innovation Expert in the Defense Industry**
Number	1	3	2	1	1	1	1	1	1
Percentage	17%	50%	33%	17%	17%	17%	17%	17%	17%

The competence of the experts was assessed according to the model presented in (Tešić & Božanić, 2024). This model for evaluating expert competence is based on the calculation of the competence coefficient (K). The coefficient encompasses three assessment dimensions: an objective assessment – Kd, an evaluation of the

DRAFT VERSION

source of argumentation – Ka, and a subjective assessment – Ks. The overall competence coefficient is calculated according to the following formula:

$$K = q_1 K_d + q_2 K_\mathrm{a} + q_3 K_s \quad (1)$$

The competence coefficients of the selected experts are presented in Table 2.

Table 2. Competence Coefficients of the Selected Experts

Expert	Assessment Aspects			Competence Coefficient	Normalized Values
	Kd **(0,6)**	***Ka*** **(0,25)**	***Ks*** **(0,15)**		
1.	0,95	0,70	0,8	**0,86**	0,178534572
2.	0,69	0,80	0,8	**0,73**	0,151186791
3.	0,91	0,85	0,9	**0,89**	0,184726522
4.	0,74	0,85	0,9	**0,79**	0,163054696
5.	0,85	0,60	0,8	**0,78**	0,160990712
6.	0,73	0,85	0,9	**0,78**	0,161506708

Since the average competence coefficient of the surveyed experts, both individually and as a group, exceeds 0.5, it can be concluded that the experts are competent (Tešić & Božanić, 2024). The obtained competence coefficients (normalized values) will be used as weighting factors in the evaluation of criteria and alternatives during the application of multi-criteria optimization methods.

4. CALCULATION OF CRITERIA WEIGHTING COEFFICIENTS USING THE FAHP METHOD

Fuzzification of the method was carried out by fuzzifying the scale used for comparing the criteria. The simplest way to fuzzify Saaty's scale is through the application of fuzzy numbers with a predefined confidence interval, that is, with a predetermined left and right distribution of the fuzzy number. In other words, the confidence intervals of the fuzzy numbers are first established, and then pairwise comparisons are performed.

Although various approaches to the fuzzification of Saaty's scale exist, triangular fuzzy numbers are most commonly used in the literature. In our study, we use the model [x–1, x, x+1] for values ranging between "equal importance" and "absolute dominance," where x denotes the value of the fuzzy number with a membership degree equal to 1 (Liu et al., 2020; Ekmekcioğlu et al., 2021). The linguistic terms, values of triangular fuzzy numbers, and their reciprocal values are presented in Table 3.

Table 3. Linguistic Variables and Fuzzified Evaluation Scales

Linguistic Term	Value According to the Classical Saaty Scale	Triangular Fuzzy Number	Fuzzy Reciprocal Value
Equal significance	1	(1,1,1)	(1/1, 1/1, 1/1)
Low dominance	3	(2,3,4)	(1/4, 1/3,1/2)
High dominance	5	(4,5,6)	(1/6, 1/5, 1/4)
Highly dominance	7	(6,7,8)	(1/8, 1/7, 1/6)
Absolute dominance	9	(9,9,9)	(1/9, 1/9, 1/9)
Medium values	2	(1,2,3)	(1/3, 1/2, 1/1)
	4	(3,4,5)	(1/5, 1/4, 1/3)
	6	(5,6,7)	(1/7, 1/6, 1/5)
	8	(7,8,9)	(1/9, 1/8, 1/7)

The task of each expert was to individually compare the determinants (criteria) in terms of their importance for the role of the military and the military profession in contemporary society, using the comparison matrix presented in Table 4.

Table 4. Comparison matrix

	C1	C2	C3	C4	C5	C6
C1	1, 1, 1					

DRAFT VERSION

C2		1, 1, 1				
C3			1, 1, 1			
C4				1, 1, 1		
C5					1, 1, 1	
C6						1, 1, 1

The data collected through interviews were entered for each expert individually. The experts independently evaluated the criteria, respectively. For the purpose of data collection, interviews were conducted with six selected and competent experts.

Determining the Consistency of the Matrix and Calculating the Criteria Weight Vector. Since the obtained comparison matrices must satisfy the consistency condition, two parameters were considered for each comparison matrix provided by the experts: the Consistency Index (CI) and the Consistency Ratio (CR). Both parameters are defined by the following expressions (2 and 3), respectively,

$$CI = \lambda_{max} - n \,/\, \mathrm{n} - 1 \tag{2}$$

$$CR = CI/\mathrm{RI} \tag{3}$$

where λ_{max} is the maximum eigenvalue of the comparison matrix, RI is the Random Index, and n is the number of criteria being compared. The value of the Random Index depends on the size of the matrix and is taken from Table 5.

Table 5. The value of the Random Index

N	1	2	3	4	5	6	7	8	9
RI	0	0	0,58	0,9	1,12	1,24	1,32	1,41	1,45

Table 6. The obtained values of the criteria weights

	C1	**C2**	**…**	**C6**	**Fuzzy geom. mean**	**Wi (fuzzy)**	**Wi (COA)**	**Wi (n)**	**Rank**
C1	1,00 1,00 1,00	1,63 4,53 1,22		1,69 2,63 3,05	1,29 1,88 2,40	0,14 0,29 0,52	0,319	0,277	**1**
C2	0,22 0,27 0,62	1,00 1,00 1,00		0,76 1,10 1,48	0,60 0,82 1,17	0,07 0,13 0,25	0,149	0,130	**4**
C3	0,22 0,27 0,62	0,74 1,14 2,16		1,18 1,75 2,49	0,68 1,00 1,52	0,08 0,15 0,33	0,187	0,163	**3**
C4	0,23 0,29 0,38	0,48 0,63 0,82	….	0,53 0,61 0,83	0,44 0,55 0,75	0,05 0,08 0,16	0,099	0,087	**6**
C5	0,83 1,29 1,92	1,07 1,36 1,94		1,45 2,99 3,83	1,04 1,50 2,06	0,12 0,23 0,45	0,266	0,232	**2**
C6	0,33 0,42 0,59	0,67 0,91 1,32		1,00 1,00 1,00	0,55 0,70 0,98	0,06 0,11 0,21	0,128	0,111	**5**
Σ					4,60 6,45 8,88		1,149		
1/Σ					0,11 0,15 0,21				

If the value of the Consistency Ratio (CR) is less than 0.10, the inconsistency is within acceptable limits; otherwise, it is necessary to repeat the evaluation and the previous steps. In our case, the experts demonstrated consistency in their assessments. These confirmed consistent comparison matrices were then weighted using the normalized values of the experts competence coefficients and subsequently aggregated into a single matrix by summing the values of the fuzzy numbers Table 6.

To rank the criteria, the fuzzy geometric mean of the criteria weight coefficients and the weight coefficients of their left and right distributions was first calculated from the aggregated pairwise comparison matrix. The final fuzzy weight value of each criterion was obtained by multiplying the fuzzy geometric mean value with the reciprocal of the sum of the fuzzy geometric means of all criteria. The crisp weight value was derived using the centroid method (as the arithmetic mean), and the values were subsequently normalized. The final ranking of the criteria is presented in Table 7, obtained by using methods of fuzzy AHP.

Table 7. Rank of Criteria

	Fuzzy Geometric Mean			**Fuzzy Criterion Weight**			**Criterion Weight**	**Normalized Criterion Weight Value**	**Rank**
C1	1,288	1,882	2,396	0,144	0,286	0,513	0,314	**0,274**	**1**

DRAFT VERSION

C2	0,605	0,821	1,167	0,068	0,125	0,250	0,147	**0,129**	**4**
C3	0,743	1,117	1,598	0,083	0,170	0,342	0,198	**0,173**	**3**
C4	0,442	0,549	0,754	0,049	0,084	0,161	0,098	**0,086**	**6**
C5	1,044	1,502	2,065	0,117	0,228	0,442	0,262	**0,229**	**2**
C6	0,551	0,702	0,977	0,061	0,107	0,209	0,126	**0,110**	**5**
Σ	**4.508**	**6.558**	**9.138**				**1,146**	**1**	
1/ Σ	**0.109**	**0.152**	**0.221**						

5. CONCLUSION

The methodology and the analytical framework proposed in this paper is intended to provide a more comprehensive understanding of the role and position of the military and the military profession in contemporary society. As such, it may serve as a solid foundation for future research on this subject. What is certain is that the future of the military and the military profession will be marked by the need for rapid adaptation to dynamic technological, geopolitical, and broader societal changes.

The application of this modified approach to the case of defining criteria for analyzing the position and role of the military and the military profession in modern society demonstrates that the developed model enables the identification of the significance of criteria from a predefined set. Therefore, fuzzy AHP proves to be a valuable tool in the strategic decision-making process in such contexts. The model not only mitigates the cognitive burden on the decision maker but also supports effective decision-making among individuals with limited experience.

Given these advantages, this methodological framework can be extended to other decision-making scenarios, particularly in situations characterized by varying degrees of subjective conviction among decision makers.

6. LITERATURE

Bojanic, D., Kovač, M., Bojanic, M., & Ristic, V. (2018). Multi-criteria decision-making in A defensive operation of the guided anti-tank missile battery: An example of the hybrid model fuzzy AHP - MABAC. Decision Making: Applications in Management and Engineering, 1(1), 51–66. https://doi.org/10.31181/dmame180151b

Ekmekcioğlu, Ö., Koc, K., Özger, M., (2021). Stakeholder perceptions in flood risk assessment: A hybrid fuzzy AHP-TOPSIS approach for Istanbul, Turkey, *International Journal of Disaster Risk Reduction*, Volume 60/102327, https://doi.org/10.1016/j.ijdrr.2021.102327

Hornstra, Steven P. A., Steven J. Durning, Jaap A. Hoogenboezem and Walther N. K. A. van Mook. 2024. "Closing the gap between skills training and academic education at a military academy: An integrated instructional design model". *Ukrainian Journal of Educational Studies and Information Technology, 12*(1): 1–14. DOI 10.32919/uesit.2024.01.01

Liu, Y., Eckert, C. M., & Earl, C. (2020). A review of fuzzy AHP methods for decision-making with subjective judgements. Expert Systems with Applications, https://doi.org/10.1016/j.eswa.2020.113738

Milićević, M. Ekspertsko ocenjivanje,Medijacentar „Odbrana", Beograd, 2014.

Radojević, S. M., Knežević, M. N., & Bojanić, D. J. (2023). Politika neutralnosti i ulaganja u sistem odbrane. Vojno delo, 75(2), 14-31. https://doi.org/10.5937/vojdelo2302014R

Saaty, T. L. The analytic hierarchy process. McGraw-Hill, New York. 1980.

Tešić, D. Z., & Božanić, D. I. (2024). Model for determining competences of experts in the field of Military Science. Vojno delo, 76(1), 1-22. https://doi.org/10.5937/vojdelo2401001T

Vukadinović, M. i Milenković. M (2024), „Vojna profesija u Srbiji kroz prizmu međunarodne vojne saradnje u savremenom geopolitičkom kontekstu" U *Međunarodne integracije kao instrument geopolitike* / ed. Zoran Milošević i Dajana Lazarević, 105–126. Beograd: Institut za političke studije. DOI: 10.18485/ips_integr_geopol.2024.ch9

Zadeh, L. A. Fuzzy sets. *Information and Control*. 1965, 8, 338-353.

Župac, G.*A* Model for Determining and Evaluating Criteria for the Selection of a Medium-Range Surface-to-Air Missile System, Doctoral Dissertation, University of Defence in Belgrade, Military Academy, Belgrade, 2013.

DRAFT VERSION

SYM-OP-IS 2025
52. Simpozijum o operacionim istraživanjima
Palić, 7 - 10.9.2025.

UNIVERZITET U BEOGRADU
FAKULTET ORGANIZACIONIH NAUKA

Original article

THE SIGNIFICANCE OF PSYCHO-PHYSICAL CHARACTERISTICS OF DRIVER CANDIDATES FOR PREDICTING SUCCESS IN DRIVER TRAINING

SRDJAN DIMIĆ[1], JELENA VICANOVIĆ [2], SRDJAN LJUBOJEVIĆ[3], MOMČILO ĐORĐEVIĆ[4]

[1] University of Defence in Belgrade, Military Academy, Belgrade, srdjan.dimic@mod.gov.rs, 0000-0003-0673-4710
[2] University of Belgrade, Faculty of Technology and Metallurgy, Belgrade, jvicanovic@tmf.bg.ac.rs, 0000-0002-8268-5503
[3] University of Defence in Belgrade, Military Academy, Belgrade, srdjan.ljubojevic@va.mod.gov.rs, 0000-0002-2696-3062
[4] University of Defence in Belgrade, Military Academy, Belgrade, momcilo_djordjevic@yahoo.co.uk, 0000-0001-7791-7026

Abstract: *As driver training evolves, the demand for effective and resource-efficient educational frameworks intensifies. Based on experts´ opinion, this paper examines the significance of the psychophysical characteristics of driver candidates that contribute to training success, with the goal of enhancing predictive capacity and individualizing instruction. Grounded in research conducted within the Serbian Armed Forces, this paper presents findings from a project initiated by the University of Defence in Belgrade. Results underscore the relevance of attributes such as logical reasoning, concentration, visual perception, stress tolerance, and psychomotor coordination, while also revealing expert uncertainty in quantifying their relative weight. The findings support the development of adaptive training programs and more efficient resource planning. By anticipating individual training trajectories early in the process, this approach enhances both pedagogical quality and operational efficiency in military and civilian driver education.*

Keywords: *Psycho-Physical characteristics, Vienna Test System, driver candidates, driver training.*

1. INTRODUCTION

Parallel to the technological advancements in the automotive industry, the process of training individuals in the operation of motor vehicles has also undergone significant development. In its early stages—spanning the late 19th and early 20th centuries—driver education was largely informal, relying predominantly on the acquisition of practical skills through hands-on experience, typically facilitated by the vehicle owner or a close acquaintance possessing the requisite knowledge and expertise. However, with the increasing incidence of traffic accidents, the necessity for a more structured and formalized approach to driver training became evident. Consequently, the first theoretical and practical lessons were introduced, along with examinations as prerequisites for obtaining a driver's license. In contemporary practice, driver training is conducted within specialized institutions, commonly referred to as driving schools, although alternative models also exist

Models of driver education vary across countries; however, several general typologies can be distinguished:

- *The classical training model*, which comprises theoretical instruction, practical driving lessons, and a final examination, all conducted within licensed institutions or driving schools;
- *The competency-based model*, which is less formalized but emphasizes the acquisition and development of concrete driving skills and appropriate traffic behavior;
- *The graduated driver licensing model*, in which training is implemented in multiple stages over a period of several years. In this model, the driver's license—and with it, full driving privileges—is acquired progressively through stages such as learner's permits, provisional licenses, and, finally, full licence;
- *Modern training models*, which concentrate on the development of specialized driving competencies, such as eco-driving, safe or defensive driving, and driving under specific conditions. These models extensively incorporate contemporary technological tools, including simulators and other digital educational resources.

Regardless of the specific training model in question, a common characteristic across all is their high demand for resources.

In the Republic of Serbia, driver training is conducted within specialized, licensed institutions (driving schools), predominantly through practice on training grounds and in public traffic, with simulators being

DRAFT VERSION

employed only exceptionally. In addition to civilian driving schools, training is also carried out by the Serbian Armed Forces (SAF), within a unit dedicated to driver education, which boasts a decades-long tradition in this field. Given the institutionalized nature of the training process, its long-standing tradition, as well as the extensive capacities and technical infrastructure—particularly with regard to practical training delivered at the standardized and unique facility of the ”Beranovac” autodrome—the SAF is recognized as a leader in this domain, not only nationally but also in a broader context. What further distinguishes the driver training conducted by the SAF’s specialized unit from other driving schools and training centers is the intensive and complex psychological preparation of candidates, along with a strong emphasis on fostering a sense of responsibility and traffic culture among trainees.

In line with the foregoing, the driver training process requires substantial resources—material, human, temporal, and otherwise—while efforts to optimize their utilization remain ongoing. To this end, the University of Defence in Belgrade has initiated a scientific research project, one of whose subsidiary objectives is the development of a mathematical model capable of reliably predicting candidate success in training, based on psychophysical assessments conducted prior to instruction. This model aims not only to forecast training outcomes with a high degree of confidence but also to identify individual strengths and weaknesses, thereby enabling the customization of instruction by tailoring training components to the specific needs of each candidate.

This paper presents the results of a study examining the relevance of selected psychological and physical attributes of driver candidates for the multi-criteria evaluation of their potential to successfully complete training. The findings lay the groundwork for improved prediction of the resources required for training, particularly in the context of so-called additional training, which is provided to candidates who fail the driving test on their first attempt.

2. PSYCHOLOGICAL AND PHYSICAL CHARACTERISTICS OF DRIVER CANDIDATES RELEVANT TO THE SUCCESSFUL COMPLETION OF DRIVING TRAINING

The success of candidates in driver training is contingent upon a wide array of internal and external factors. Among the internal factors, several categories of psychophysical characteristics can be identified that facilitate the successful completion of training and, subsequently, the efficient and safe operation of a motor vehicle. The most prominent among these are typically considered to be attributes related to psychological stability, concentration and attention, psychomotor coordination, sensory capabilities, cognitive abilities, motivation, and a sense of responsibility (see Porter, 2011).

For the purpose of predicting training success, psychophysical characteristics were selected that can be objectively measured using standardized tests within the framework of the Vienna Test System (VTS - a psychophysical testing platform developed by the Vienna-based company Schuhfried GmbH), which was developed, among other applications, specifically for the assessment of driver capabilities (Kubinger, 2007).

From the available tests within the VTS, a battery of tests was assembled and administered to driver candidates prior to the commencement of training. The battery comprised nine tests through which sixteen variables were measured (a more detailed description of the tests can be found in Schuhfried, 2013):

1. Logical Reasoning Test – designed to assess the candidates’ cognitive abilities, administered in the form of a specially developed adaptive matrices test within the VTS;
2. Cognitron Test – aimed at evaluating candidates’ attention and concentration;
3. Determination Test – used to assess reactive stress tolerance, attention, and reaction speed;
4. Peripheral Perception Test – measures the ability to receive and process peripheral visual information, specifically the candidate’s visual field breadth and capacity for divided attention;
5. Reaction Time Test – evaluates various aspects of reactive behavior, namely, the average reaction time as an indicator of response speed to relevant stimuli, and the average motor time as a measure of movement speed during planned activity;
6. Adaptive Tachistoscopic Traffic Perception Test – assesses visual observation skills, situational awareness, visual orientation, and perceptual speed in driving scenarios;
7. Traffic Risk-Taking Test – measures the candidate’s propensity to take risks in specific traffic situations;
8. Driving-Related Personality Inventory – evaluates personality traits based on self-assessment, including the individual's willingness to take risks, perceived social responsibility, self-control, and psychological stability;

DRAFT VERSION

9. Driving-Related Item Set – used to assess dimensions such as uncritical self-perception, aggressiveness in interpersonal interactions, and attitudes toward traffic norms, commonly referred to as "emotional driving".

The success of candidates on the driving exam is a function of all the aforementioned variables, reflecting a multicriteria nature. For the design of a multicriteria model aimed at predicting candidate success on the exam, one of the essential steps involves determining the significance of each variable (i.e., the candidates' psychophysical characteristics) within the model. This procedure was conducted as part of the research activities under the previously mentioned project of the University of Defence and was implemented through a combination of methods, including expert opinion elicitation, linear regression based on feedback data from the tested candidates' exam performance, multicriteria optimization techniques, and the application of metaheuristic methods.
This paper presents the results obtained through expert opinion collection.

3. DETERMINING THE SIGNIFICANCE OF PSYCHOPHYSICAL CHARACTERISTICS OF DRIVER CANDIDATES FOR PREDICTING TRAINING SUCCESS

In order to determine the significance (i.e., relative weights) of the psychophysical characteristics of driver candidates within a multicriteria model for predicting their training success, experts in the planning, organization, and implementation of training within the Serbian Armed Forces were surveyed using the fuzzy Delphi method.

The task assigned to the experts participating in the weighting process was to assess and linguistically express the impact of each variable from the selected VTS test battery on candidate success in passing the driving exam. Their assessments were provided using a customized fuzzy linguistic scale, represented by triangular fuzzy numbers $A_k = (a_k^l, a_k^m, a_k^u)$, where the parameters of a triangular fuzzy number correspond to the lower or pessimistic value (a^l), the middle or most likely value (a^m), and the upper or optimistic value (a^u) (see Ghazinoory, Zadeh, Memariani, 2007). The index k denotes the k-th characteristic being evaluated for its significance (Figure 1).

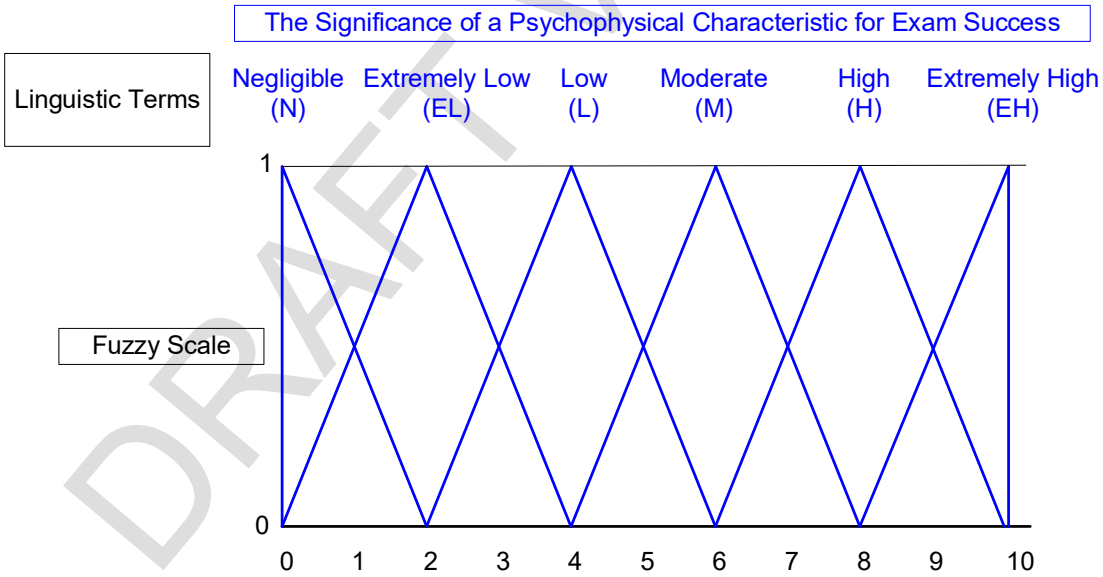


Figure 1: Fuzzy Scale for Assessing the Significance of Psycho-physical Characteristics of Driver Candidates for Exam Success

The aggregated opinion of the expert group for each individual psychophysical characteristic of the candidate, expressed as a fuzzy number $A_{sr} = (a_{sr}^l, a_{sr}^m, a_{sr}^u)$, is obtained by combining the individual assessments of all N experts according to expression (1), representing the average value of the resulting set of fuzzy numbers (opinions) for the *k*-th characteristic

DRAFT VERSION

$$A_{sr} = (a_{sr}^{l}, a_{sr}^{m}, a_{sr}^{u}) = (\frac{1}{N} \cdot \sum_{i=1}^{N} a_i^{l}, \frac{1}{N} \cdot \sum_{i=1}^{N} a_i^{m}, \frac{1}{N} \cdot \sum_{i=1}^{N} a_i^{u}) \quad (1)$$

In this specific case, all experts were assigned equal importance (although this is not necessarily required), by allocating identical weighting coefficients w_i (where $\sum_{1}^{n} w_i = 1$, N – denotes the number of experts).

The final expert assessment of the significance of a given psychophysical characteristic was determined after conducting the required number of iterations within the expert survey process, in accordance with the Delphi method as described in Bojadzijev and Bojadzijev (2007) and Ljubojević (2016). To ensure the stability of expert opinions, a convergence condition was established: the sum of the weighting coefficients assigned to experts who revised their judgments in the final round of the survey (ΔW), compared to the previous round, must be less than w_{gr} (%) of the total sum of weighting coefficients (see Expression 2).

$$\Delta W = \sum_{1}^{n} w_j^* < w_{gr} \text{ , where } w_j^* = \begin{cases} w_j, O_{jp} \neq O_{jp-1} \\ 0, O_{jp} = O_{jp-1} \end{cases} \quad (2)$$

w_{gr} (%) – stability threshold of expert opinions (expressed as a percentage), where *Ojp* denotes the assessment provided by the *j*-th expert in the *p*-th evaluation cycle.

The opinion of the expert group regarding the significance of that characteristic for the overall success of the candidate on the driving exam is represented by a triangular fuzzy number $A_{sr} = \left(a_{sr}^{l}, a_{sr}^{m}, a_{sr}^{u}\right)$, whose membership function takes the following form:

$$\mu_A(x) = \begin{cases} 0, & x \le a^l \\ \dfrac{x - a^l}{a^m - a^l}, & a^l < x \le a^m \\ \dfrac{a^u - x}{a^u - a^m}, & a^m < x < a^u \\ 0, & x \ge a^u \end{cases} \quad (3)$$

A total of 13 experts participated in the present survey, all of whom, in addition to possessing theoretical knowledge, have many years of experience in military driver training. At the end of 2024 and the beginning of 2025, these individuals provided their evaluations across three rounds of surveying, from which the aggregated (consensus) opinion of the expert group was derived.

To determine opinion stability, a convergence criterion was established: the sum of the weighting coefficients assigned to experts who altered their assessments in the final survey round had to be less than 0.20 (ΔW < 0.20), which effectively implies that 80% of the surveyed experts did not revise their initial judgments.

The normalized values of the expert assessments concerning the significance of the observed characteristics of driver candidates for their success on the driving exam are presented in Figure 2 and Table 1.

DRAFT VERSION

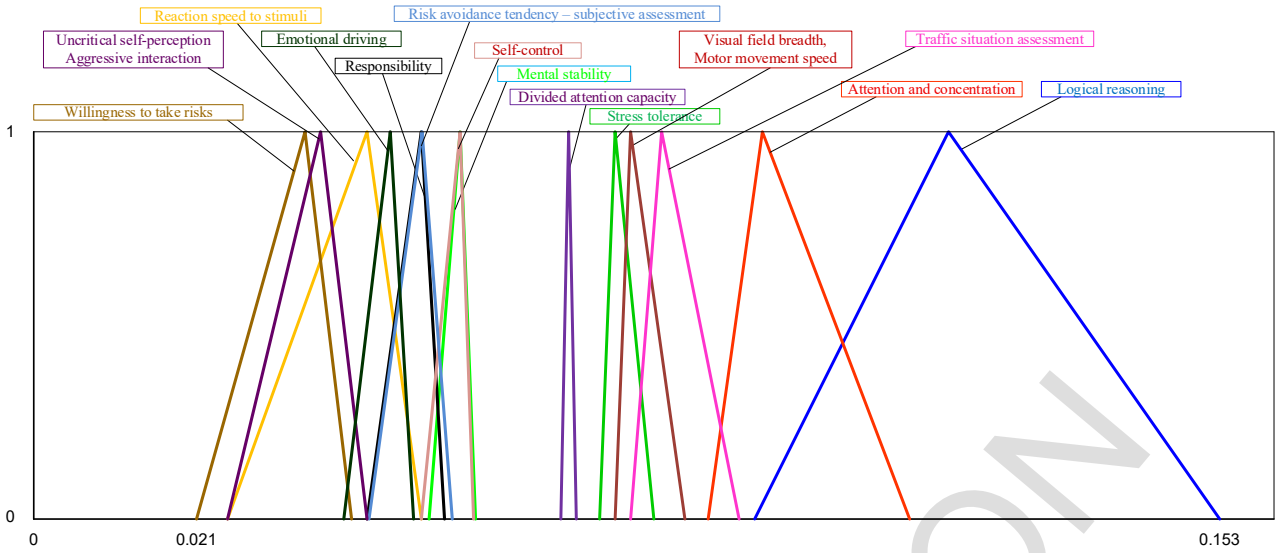


Figure 2: Expert Evaluation of the Significance of the Observed Psychophysical Characteristics Expressed as Fuzzy Numbers

Table 1. Significance of Psychophysical Characteristics of Driver Candidates for Success on the Driving Exam

Psychophysical Characteristic of the Candidate	$A_{sr} = (a_{sr}^{l}, a_{sr}^{m}, a_{sr}^{u})$	**Centroid**	**Dispersion**
Logical reasoning	(0.096, 0.118, 0.153)	0.122	0.057
Attention and concentration	(0.087, 0.094, 0.113)	0.098	0.026
Stress tolerance	(0.073, 0.075, 0.080)	0.076	0.007
Visual field breadth	(0.075, 0.077, 0.084)	0.079	0.009
Divided attention capacity	(0.068, 0.069, 0.070)	0.069	0.002
Reaction speed to stimuli	(0.025, 0.043, 0.050)	0.039	0.025
Motor movement speed	(0.075, 0.077, 0.084)	0.079	0.009
Traffic situation assessment	(0.077, 0.081, 0.091)	0.083	0.014
Mental stability	(0.051, 0.055, 0.057)	0.054	0.006
Responsibility	(0.043, 0.050, 0.053)	0.049	0.010
Self-control	(0.050, 0.055, 0.057)	0.054	0.007
Risk avoidance tendency – subjective assessment	(0.043, 0.050, 0.054)	0.049	0.011
Willingness to take risks	(0.021, 0.035, 0.041)	0.032	0.020
Uncritical self-perception	(0.025, 0.037, 0.043)	0.035	0.018
Aggressive interaction	(0.025, 0.037, 0.043)	0.035	0.018
Emotional driving (attitude toward norms)	(0.040, 0.046, 0.049)	0.045	0.009

Table 1 also presents a comparison of the obtained fuzzy values based on the centroid method (Wang et al., 2006) and the width, that is, the degree of uncertainty (Hajjari, 2011). Among the results displayed, it is evident that the selected group of experts attributes the greatest significance for success on the driving exam to the following candidate characteristics: logical reasoning ability, ability to sustain attention and concentration, capability for traffic situation assessment, visual field breadth, motor movement speed, and stress tolerance. According to the experts, these six psychophysical

DRAFT VERSION

characteristics collectively account for more than 50% of the overall influence on exam success (specifically, 53.7%).

At the same time, this group of characteristics is also associated with the greatest dispersion of expert opinions—that is, the highest degree of uncertainty and disagreement among experts regarding the precise evaluation of their importance. In other words, while the experts agree that these attributes are of paramount importance for training success, they differ to some extent in quantifying the exact weight of their contribution.

4. CONCLUSION

A wide range of factors influence the success of driver candidates during motor vehicle training. By identifying the candidates' psychophysical characteristics and abilities prior to the onset of training, it becomes possible to predict their likelihood of success with a satisfactory degree of reliability. To achieve this, a suitable multicriteria model is required, as well as an understanding of the relative importance of each selected characteristic within that model.

This paper presents the results of an expert assessment of the significance of sixteen psychophysical characteristics of driver candidates, all of which can be objectively measured using an appropriate battery of tests within the Vienna Test System. According to these findings, experts assign greater importance to attributes such as logical reasoning, attention and concentration, traffic situation assessment, visual field breadth, motor response speed, and stress tolerance, as opposed to candidates' willingness to take risks, uncritical self-perception, or aggressive interaction with other road users.

To arrive at a definitive conclusion regarding the influence and relative weight of these characteristics within a mathematical multicriteria model, further analysis using additional methods is necessary. Such a model would allow for the early identification—immediately upon completion of psychophysical testing—of candidates unlikely to pass the driving exam on their first attempt. This would, in turn, facilitate the prediction of the additional material, time, personnel, and other resources required to support supplementary training.

Simultaneously, test results could serve as the foundation for designing individualized training programs tailored to each candidate's specific weaknesses, thereby increasing their chances of success. In this way, both the quality of driver education and the efficiency of resource planning and utilization can be substantially improved.

DRAFT VERSION

ACKNOWLEDGMENT

This paper stems from the scientific research project VA-TT/2/20-25 funded by the Ministry of Defence of the Republic of Serbia.

LITERATURE

Bojadziev, G., Bojadziev, M. (2007). *Fuzzy Logic for Business, Finance and Management*, Word Scientific Publishing, New Jersey, USA

Ghazinoory, S., Zadeh, E. A., Memariani, A. (2007). Fuzzy SWOT analysis, *Journal of Intelligent and Fuzzy Systems*, *18*, 99–108.

Hajjaari, T. (2011). Ranking Fuzzy Numbers Based on Ambiguity Degrees, *Australian Journal of Basic and Applied Sciences*, 5(1), 62-69.

Kubinger, K. D. (2007). Psychological test calibration using the Rasch model: some critical remarks. *Psychology Science, 49(1)*, 49-56. doi:10.4018/s15327574ijt0504_3

Ljubojevic, S. (2016). *Model odlučivanja organa saobraćajne službe u zadacima strategijskog transporta*, Doctoral dissertation, University of Defence in Belgrade, Military Academy, Belgrade, Serbia.

Porter, E. B. (2011). *Handbook of traffic psychology* (1st ed.), Academic Press, USA.

Schuhfried, G. (2013). *Vienna test system: test manual.* Modling, Austria: Schuhfried GmbH.

Wang, Y-M., Yang, J-B., Xu, D-L., Chin, K-S. (2006). On the centroids of fuzzy numbers, *Fuzzy sets and systems*, Vol. 157, Issue 7, 919-926. https://doi.org/10.1016/j.fss.2005.11.006

ARTIFICIAL INTELLIGENCE

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

PRODUCT PROFESSIONALS' PERSPECTIVES ON TRADITIONAL AND AI-ENHANCED PRODUCT BACKLOG PRIORITIZATION APPROACHES

MILOŠ BELČEVIĆ[1], DRAGAN PAMUČAR[2]

[1] University of Belgrade, Faculty of Organizational Sciences, Belgrade, milos.belcevic@gmail.com, 0009-0009-1579-269X
[2] University of Belgrade, Faculty of Organizational Sciences, Belgrade, dragan.pamucar@fon.bg.ac.rs, 0000-0001-8522-1942

Abstract: *Product backlog prioritization is a critical to modern software development, yet the methods used in practice often remain ad hoc and heuristic-driven. While academic research has proposed numerous advanced techniques, empirical evidence on their adoption and effectiveness in real-world settings is limited. This study bridges the gap between theory and practice by combining a systematic literature review (SLR) with a survey of 287 product professionals internationally. The SLR categorizes existing backlog prioritization methods across eight methodological families and highlights a disconnect between the abundance of novel models and their practical adoption. Survey results reveal that traditional techniques are the most commonly used, although satisfaction varies, with some showing high variance and polarized opinions. AI/ML-based approaches are used infrequently, but there is strong openness among practitioners to explore these. The findings emphasize the need for better alignment between academic advancements and industry practices, clearer value propositions for AI/ML methods, and greater support for product teams in adopting new techniques. This research contributes to the field by offering an empirical foundation for understanding backlog prioritization trends, identifying barriers to AI/ML adoption, and outlining directions for future research aimed at improving the effectiveness, scalability, and human-centered usability of product backlog prioritization methods.*

Keywords: *artificial intelligence, product backlog. requirements prioritization*

1. INTRODUCTION

Product backlog prioritization represents the ongoing process of estimating the importance (priority) of product backlog items. Despite the critical role of product backlog prioritization in modern software development, traditional methods used by most product professionals face significant challenges in efficiency, consistency, and adaptability. Numerous algorithmic and data-driven prioritization methods have been explored in academic literature – including clustering-based, genetic algorithm-based, NLP-driven, and fuzzy logic approaches – but their practical adoption remains limited. As a result, backlog prioritization often remains an ad-hoc and heuristic-driven process rather than a systematically optimized one.

The emergence of AI-driven backlog management introduces both opportunities and challenges. AI can theoretically enhance prioritization by identifying hidden patterns, predicting impact, and automating decision-making. However, questions remain regarding the effectiveness, interpretability, and industry adoption of AI-assisted backlog management. Are AI-enhanced approaches genuinely superior to traditional methods in practice? Do product teams find them useful, or do they introduce new challenges such as explainability, trust, and alignment with business goals? Addressing these questions is crucial for understanding the role AI can play in product backlog management and its potential to reshape software development workflows.

1.1. Research gap and research question

There is a significant gap in research of the real-world adoption, perceived effectiveness, and prevalence of product backlog prioritization approaches, both traditional and novel. Existing studies primarily introduce and evaluate new prioritization models, while empirical investigations into the usage and satisfaction levels of established methods – such as MoSCoW, RICE, and WSJF – remain scarce. No research comprehensively examines how frequently product professionals adopt either traditional or novel methods or assesses their practical effectiveness in real-world product management contexts. This study aims to address this gap.

DRAFT VERSION

2. RELATED WORKS

After reviewing and selecting relevant documents, eight main methodological areas of focus can be defined for the research of novel requirements prioritization methods: clustering-based, LLM-based, NLP-based, Machine Learning-based, nature-inspired, fuzzy-based, genetic algorithm-based, and semi-automated methods. Traditional approaches to product backlog prioritization appear to have received limited to no scholarly attention.

Clustering-based prioritization methods aim to enhance the management of large and complex requirement sets by grouping similar items based on various similarity measures (Achimugu et al. (2014); Belsis et al. (2014); Kumar et al. (2022, 2023); Sharma and Kumar (2023); Yang et al. (2023)). NLP-based methods prioritize software requirements by leveraging natural language processing techniques to analyze, rank, and order backlog items (Izhar et al. (2024); Ko et al. (2023); McZara et al. (2015); Shafiq et al. (2021)). LLM-based prioritization methods leverage deep learning models like GPT and BERT to analyze and rank software requirements based on contextual relevance, historical data, and stakeholder preferences (Sami et al. 2024, 2025). Machine learning-based prioritization methods leverage predictive models and adaptive learning to optimize requirement ranking, reducing manual effort and improving accuracy (Avesani et al. (2004, 2005); Perini et al. (2013); Somohano-Murrieta et al. (2021)). Nature-inspired prioritization methods utilize optimization techniques derived from biological systems, such as ant and bee colonies, to enhance software requirements ranking (Alrezaamiri et al. (2020); de Souza et al. (2011); Del Sagrado et al. (2010)). Fuzzy-based prioritization methods leverage fuzzy logic to manage uncertainty and imprecision in software requirements ranking, making them particularly useful for subjective and evolving decision-making scenarios (Ahmad et al. (2017); Gulzar et al. (2017); Mishra et al. (2016); Mougouei et al. (2019); Sadia et al. (2019)). Genetic Algorithm (GA)-based prioritization methods apply evolutionary computing techniques to optimize requirements prioritization in software development by simulating natural selection (Ahuja et al. (2018); Marghny et al. (2017); Tonella et al. (2010)). Semi-automated methods integrate automation with stakeholder preferences and dependency modeling to enhance efficiency and accuracy (Chua et al. (2022); Gupta & Gupta (2018); Shao et al. (2017)).

3. METHODOLOGY

This study employed a mixed-method approach that combined a Systematic Literature Review (SLR) with a survey of product professionals involved in product backlog prioritization. The goal was to explore both theoretical advancements and practical adoption trends in traditional and artificial intelligence-enhanced backlog prioritization methods.

Insights from the SLR and survey were synthesized to develop a comprehensive view of both academic innovation and industry practice in backlog prioritization. The results are discussed in the following sections to derive implications for product managers and to suggest future research directions.

4. RESULTS

Out of the total 307 respondents, 287 qualified as they participated in the backlog prioritization in the last 12 months. The majority work as “Product Manager” (56.1%), smaller percentage as “Product Owner” (9.4%), “Head of Product” (8.4%), “Product Consultant” (7.7%), and “Business Analyst (2.1%). There’s been a significant number of “Other” roles (16.4%), such as “CPO”, "Product Coach", "Director of Product", "Project Manager", "Program Manager" and so on where the majority of individual roles represented less than 1%, and a few around 2% or 3% of the respondents. Unlike roles, company sizes were quite equally distributed, ranging in size from 1-10 (15.7%), 11-50 (16.4%), 51-200 (23.7%), 201-1000 (16.7%), 1001-10000 (13.2%) and over 10000 (14.3%) employees. In terms of the professional experience, different brackets were similarly represented, ranging from 1-3 years (15%), 4-6 years (31%), 7-10 years (26.5%), to more than 10 years (25.4%) with only one underrepresented category of less than 1 year (2.1%). The majority of respondents hasn’t used AI or ML tools in prioritization but is open to trying it (63.4%), 23.3% used it occasionally, while 7.3% used it frequently and 5.9% did not use it and is not interested in trying. 42.5% respondents stated they were somewhat likely to use AI-assisted backlog prioritization tools in the future, 26.1% very likely, while 7.7% were somewhat unlikely and 3.5% very unlikely. 20.2% were neutral.

Our research shows that the most commonly used methods are MoSCoW, RICE, WSJF, Value vs. Effort Matrix, Kano Model, Critical Path, Opportunity Scoring, Eisenhower Matrix, Cost of Delay and custom formulas. Most methods have a median rating of around 3 or 4 (out of 5). Some methods, such as WSJF and Value vs. Effort Matrix have more spread-out ratings, suggesting a mix of strong advocates and critics. We

DRAFT VERSION

can conclude that while most prioritization methods receive generally positive ratings, some have outliers and a wider range of opinions.

Nearly 70% of the respondents haven't used AI/ML-assisted backlog prioritization, with 63,4% being open to trying it, and 5,9% not interested. 23,3% have used it occasionally, and 7,3% frequently. In terms of the satisfaction level with traditional backlog prioritization methods, no significant difference was found between AI/ML-users and non-users, beyond satisfaction with certain methods was wider spread in non-user respondents, compared to AI/ML-users which were more uniformed.

5. DISCUSSION

The findings from both SLR and the survey responses offer a nuanced perspective on the current landscape of product backlog prioritization with several key insights emerging.

Traditional methods are dominant, but satisfaction is varied. Techniques such as MoSCoW, RICE and Value vs. Effort matrices remain most frequently used approaches across company sizes and seniority levels, however satisfaction levels reveal more nuanced picture. While median scores were generally positive, several methods (such as WSJF, Value vs. Effort) showed high variance, with some practitioners rating them very low. This suggests that while these methods are well-known, they may not consistently meet the needs of diverse teams and product contexts.

Novel and AI/ML-assisted approaches show strong potential but low adoption. Despite a growing body of literature proposing novel and AI/ML-enhanced methods, real-world adoption remains limited. Only about 30% of respondents have used AI in backlog prioritization, and only 7.3% do so frequently. The most common stance is openness to experimentation, with over 60% expressing interest in future use. This points to a gap between real-world adoption and potential for novel and AI/ML-enhanced methods. The willingness to adopt new tools, with nearly 70% respondents likely to use AI/ML-based prioritization methods in the future, suggests that the industry is ready for more advanced prioritization support.

There is a gap between academic innovation and practical application. The literature reveals a wealth of novel prioritization techniques, but many remain underutilized in practice.

6. CONCLUSION

This study bridges the gap between academic advancements and industry practices in product backlog prioritization by combining a systematic literature review with a survey of product professionals. It highlights the widespread use of traditional methods like MoSCoW, RICE, and WSJF, while revealing limited adoption of AI/ML-based approaches despite growing interest.

The results underscore a need for better alignment between theoretical models and practical applications, as well as for enhanced guidance, tool usability, and training to support the adoption of AI/ML in backlog management. By providing empirical evidence on the current landscape and perceptions of backlog prioritization methods, this research lays the groundwork for future studies focused on improving the effectiveness, explainability, and accessibility of AI-enhanced prioritization techniques. Ultimately, this work contributes to more informed decision-making for product teams and encourages further exploration into AI's role in optimizing software development processes.

LITERATURE

Achimugu, P., Selemat, A., & Ibrahim, R. (2014). A clustering-based technique for large-scale prioritization during requirements elicitation. *Recent Advances on Soft Computing and Data Mining*, 623–632. https://doi.org/10.1007/978-3-319-07692-8_59

Ahmad, K. S., Ahmad, N., Tahir, H., & Khan, S. (2017). FuzzyMoSCoW: A fuzzy-based MoSCoW method for the prioritization of software requirements. *Proceedings of the 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, 433–437. https://doi.org/10.1109/ICICICT1.2017.8342602

Ahuja, H., Sujata, & Batra, U. (2018). Performance enhancement in requirement prioritization by using least-squares-based random genetic algorithm. In B. Panda, S. Sharma, & U. Batra (Eds.), *Innovations in Computational Intelligence: Best Selected Papers of the Third International Conference on REDSET 2016* (pp. 251–263). Springer.

Alrezaamiri, H., Ebrahimnejad, A., & Motameni, H. (2020). Parallel multi-objective artificial bee colony algorithm for software requirement optimization. *Requirements Engineering, 25*(3), 363–380. https://doi.org/10.1007/s00766-020-00328-y

DRAFT VERSION

Avesani, P., Bazzanella, C., Perini, A., & Susi, A. (2005). Facing scalability issues in requirements prioritization with machine learning techniques. *Proceedings of the 13th IEEE International Conference on Requirements Engineering (RE'05).* https://doi.org/10.1109/RE.2005.30

Avesani, P., Bazzanella, C., & Perini, A. (2004). Supporting the requirements prioritization process: A machine learning approach. *Proceedings of the International Conference on Requirements Engineering*, 306–311.

Belsis, P., Koutoumanos, A., & Sgouropoulou, C. (2014). PBURC: A patterns-based, unsupervised requirements clustering framework for distributed agile software development. *Requirements Engineering, 19*, 213–225. https://doi.org/10.1007/s00766-013-0172-9

Chua, F.-F., Lim, T.-Y., Tajuddin, B., & Yanuarifiani, A. P. (2022). Incorporating semi-automated approach for effective software requirements prioritization: A framework design. Journal of Informatics and Web Engineering, Vol. 1 No. 1 (2022): March 2022. https://doi.org/10.33093/jiwe.2022.1.1.1

Dalton, J. (2019). *Great big agile: An OS for agile leaders.* https://doi.org/10.1007/978-1-4842-4206-3

de Souza, J.T., Maia, C.L.B., Ferreira, T.d.N., Carmo, R.A.F.d., Brasil, M.M.A. (2011). An Ant Colony Optimization Approach to the Software Release Planning with Dependent Requirements. In: Cohen, M.B., Ó Cinnéide, M. (eds) Search Based Software Engineering. SSBSE 2011. Lecture Notes in Computer Science, vol 6956. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23716-4_15

Del Sagrado, J., Del Águila, I. M., & Orellana, F. J. (2010). Ant colony optimization for the next release problem: A comparative study. *Proceedings of the 2nd International Symposium on Search-Based Software Engineering.* https://doi.org/10.1109/SSBSE.2010.18

Gupta, A., & Gupta, C. (2018). CDBR: A semi-automated collaborative execute-before-after dependency-based requirement prioritization approach. *Journal of King Saud University – Computer and Information Sciences 34 (2022)* 421–432

Izhar, R., Cosh, K., & Bhatti, S. N. (2024). Enhancing agile software development: A novel approach to automated requirements prioritization. *Proceedings of the 21st International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 286–293. https://doi.org/10.1109/JCSSE61278.2024.10613648

Ko, T., Lee, J., & Jeong, H. D. (2023). Project requirements prioritization through NLP-driven classification and adjusted work items analysis. *Journal of Construction Engineering and Management, 150*(3). https://doi.org/10.1061/JCEMD4.COENG-13655

Kumar, B., Tiwari, U. K., Dobhal, D. C., & Negi, H. S. (2022). User story clustering using K-Means algorithm in agile requirement engineering. *Proceedings of the 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, 1–5. https://doi.org/10.1109/CISES54857.2022.9844390

Kumar, B., Tiwari, U. K., & Dobhal, D. C. (2023). Machine learning-based approach for user story clustering in agile engineering. *SN Computer Science, 4*, 763. https://doi.org/10.1007/s42979-023-02212-2

Marghny, M. H., El-Hawary, H. M., & Dukhan, W. H. (2017). An effective method of systems requirement optimization based on genetic algorithms. *Information Sciences Letters, 6*(1), 15–28.

McZara, J., Sarkani, S., Holzer, T., & Eveleigh, T. (2015). Software requirements prioritization and selection using linguistic tools and constraint solvers—A controlled experiment. *Empirical Software Engineering, 20*, 1721–1761. https://doi.org/10.1007/s10664-014-9334-8

Mishra, N., Khanum, M. A., & Agrawal, K. (2016). Approach to prioritize the requirements using fuzzy logic. *Proceedings of the ACEIT Conference.*

Mougouei, D., Powers, D. M. W., & Mougouei, E. (2019). A fuzzy framework for prioritization and partial selection of security requirements in software projects. *Information and Software Technology, 37*(2), 2671- 2686. https://doi.org/10.3233/JIFS-182907

Perini, A., Susi, A., & Avesani, P. (2013). A machine learning approach to software requirementsprioritization. *IEEE Transactions on Software Engineering, 39*(4), 445–461. https://doi.org/10.1109/TSE.2012.52

Sami, M. A., Waseem, M., Zhang, Z., Rasheed, Z., Systä, K., & Abrahamsson, P. (2025). Early results of an AI multi-agent system for requirements elicitation and analysis. In D. Pfahl, J. Gonzalez Huerta, J. Klünder, & H. Anwar (Eds.), *Product-focused software process improvement. PROFES 2024*. Lecture

DRAFT VERSION

Notes in Computer Science, vol. 15452. Springer, Cham. https://doi.org/10.1007/978-3-031-78386-9_20

Shao, F., Peng, R., Lai, H., & Wang, B. (2017). DRank: A semi-automated requirements prioritization method based on preferences and dependencies. *Journal of Systems and Software*, 126, 141–156. doi:10.1016/j.jss.2016.09.043

Sharma, S., & Kumar, D. (2023). Product backlog optimization technique in agile software development using clustering algorithm. *Multimedia Tools and Applications*, 82(30), 1–21. https://doi.org/10.1007/s11042-023-15406-w

Tonella, P., Susi, A., & Palma, F. (2010). Using interactive GA for requirements prioritization. *In 2nd International Symposium on Search Based Software Engineering* (pp. 57–66). https://doi.org/10.1109/SSBSE.2010.17

Yang, B., Ma, X., Wang, C., Guo, H., Liu, H., & Jin, Z. (2023). User story clustering in agile development: A framework and an empirical study. *Frontiers of Computer Science, 17*, 176213. https://doi.org/10.1007/s11704-022-8262-9

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

SHAPE-BASED IMAGE CLASSIFICATION USING HU AND ZERNIKE MOMENTS WITH BEE COLONY OPTIMIZATION FOR TUNING SVM PARAMETERS

DUŠAN DENČIĆ[1], NATAŠA MILOSAVLJEVIĆ[2], TATJANA DAVIDOVIĆ[3]

[1] Faculty of Agriculture, University of Belgrade Serbia, dusan.dencic@agrif.bg.ac.rs, 0000-0002-7231-4186
[2] Faculty of Agriculture, University of Belgrade Serbia, natasam@agrif.bg.ac.rs, 0000-0003-4056-089X
[3] Mathematical Institute, Serbian Academy of Science and Arts, tanjad@mi.sanu.ac.rs, 0000-0001-9561-5339

Abstract: *In this paper, we propose a hybrid approach for image classification that combines invariant shape descriptors with metaheuristic hyperparameter tuning. Specifically, Hu and Zernike moments are used to represent geometric and structural properties of objects within images, while a Support Vector Machine (SVM) is employed as the base classifier. To optimize the performance of the SVM, we integrate the Bee Colony Optimization (BCO) algorithm, which efficiently searches for the optimal combination of hyperparameters CCC and γ. Experimental results on publicly* available *Fruit and Cotton* image *datasets from Kaggle show that the proposed system significantly improves the classification performance, particularly in class-imbalanced scenarios. The application of BCO resulted in a stable solution with a weighted F1-score of 0.7933, demonstrating both convergence efficiency and robustness. This framework is adaptable to a wide range of image recognition tasks where shape plays a dominant role.*

Keywords: *computer vision, object segmentation, invariant shape descriptors, hyperparameter selection, metaheuristics.*

DRAFT VERSION

1. INTRODUCTION

The classification of visual objects based on their structural characteristics is a fundamental problem in computer vision and pattern recognition. In applications ranging from waste sorting to biological image analysis, shape-based features often play a dominant role, especially when texture or color is not sufficiently informative. Traditional machine learning models, such as Support Vector Machines (SVM) (Chandra et al. 2021), have demonstrated strong performance in such domains, but their effectiveness critically depends on the appropriate selection of hyperparameters.

Hu moments and Zernike moments (Shen et al. 1999) are well-known shape descriptors that provide geometric invariance to translation, scaling, and rotation. They allow for a compact and discriminative numerical representation of an object's contour and global structure. However, even with strong features, the performance of an SVM can degrade if hyperparameters such as the regularization term CCC and kernel parameter γ are not well-tuned.

To address this challenge, we propose the use of Bee Colony Optimization (BCO) (Davidovic et al. 2015; Teodorovic et al., 2015), a population-based metaheuristic inspired by the foraging behavior of honey bees. BCO is utilized to optimize the SVM hyperparameters with respect to the F1-score on a validation dataset, guiding the search toward the most effective model configuration.
The contributions of this paper are threefold:

1. We utilize established shape descriptors — Hu and Zernike moments — in combination to extract complementary features from grayscale images, enabling effective classification of structurally similar objects.
2. We employ the BCO algorithm to automate the optimization of SVM hyperparameters, improving both performance and convergence stability.
3. We present an empirical analysis showing the robustness and practicality of the proposed approach in class-imbalanced scenarios.

In the following sections, we present a detailed mathematical formulation of the Hu and Zernike moments and how they are used to construct feature vectors for image representation. We then describe the experimental setup, including the structure of the dataset, the preprocessing pipeline, and the implementation of the BCO algorithm for hyperparameter tuning of the SVM classifier. Through empirical results and comparative performance metrics, we demonstrate how the integration of shape-based descriptors with metaheuristic optimization leads to improved classification outcomes, particularly in scenarios involving class-imbalance and structural variability.

2. MATHEMATICAL EXPRESSIONS

This section introduces the mathematical foundations of Hu and Zernike moments, which are used to extract shape-based features from grayscale images. Let $I(x, y)$ denote the intensity value (grayscale level) of the image at the pixel coordinates (x, y), where x and y are integers within the dimensions of the image.

Hu moments are a set of seven invariant moments derived from normalized central moments, used to characterize the shape of an image. The raw (or spatial) moments of order (p, q) is defined as:

$$m_{pq} = \sum_{x} \sum_{y} x^p y^q I\,(x, y)$$

where $p,\ q \geq 0$. In praticular, m_{00} corresponds to the total intensity (sum of pixel values) over the image. The centroid of the image, representing its center of mass, is given by:

$\bar{x} = \frac{m_{10}}{m_{00}}, \bar{y} = \frac{m_{10}}{m_{00}}$

The central moments of order (p, q), which accounts for translation invariance, is:

$$\mu_{pq} = \sum_{x} \sum_{y} (x - \bar{x})^p (y - \bar{y})^q\ I(x, y)$$

To achieve scale invariance, normalized central moments are defined as:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{1 + (p+q)/2}}$$

Hu moments are seven specific combinations of the normalized central moments η_{pq}, defined by M.K. Hu (1962), and are invariant to translation, rotation, and scaling. These seven values form the Hu feature vector**.**Translation invariance is achieved by using central moments, which are computed relative to the centroid, eliminating dependence on the image's position. Rotation invariance stems from the algebraic combinations of normalized moments, which remain unchanged under rotation. Scaling invariance is ensured by normalizing moments with respect to the zero-th moment (μ_{00}), which accounts for the image's area. These properties make Hu moments robust for shape recognition tasks. For a detailed exploration of Hu moments and their mathematical derivation, refer to (Hu et al., 2024).

Zernike moments are orthogonal moments defined over a unit disk, effective for capturing shape and structural features of an image. For an image $I(x, y)$, the Zernike moment of order *n* and repetition *m* (where ($|m| \leq n$ and $n - |m|$ is even) is defined as:

$$A_{n,m} = \frac{n + 1}{\pi} \iint_{x\hat{}2 + y\hat{}2 \leq 1} I(x, y)\ V^*_{n,m}(x, y)\ dx\ dy$$

where $V_{n,m}(x, y)$is the Zernike polynomial defined in (Virendra N. Mahajan et al. 1981), and $V^*_{n,m}(x, y)$ represents the conjugate complex value of $V_{n,m}(x, y)$.

Zernike moments are inherently rotation-invariant, since the magnitude of the complex moments $|A_{n,m}|$ remains unchanged when the image is rotated. Translation invariance is achieved by normalizing the image to place its centroid at the origin. Scaling invariance is obtained by mapping the image onto a unit disk, typically by scaling coordinates relative to the image's radius, derived from central moments or maximum pixel distance. These normalization steps ensure robustness across transformations.

The feature vector for Zernike moments is formed by collecting the magnitudes of the moments ($|A_{n,m}|$) for a chosen maximum order n_{max} and corresponding repetitions m. In practice, when constructing a feature vector based on Zernike moments, only non-negative orders m are considered, due to the conjugate symmetry property $A_{n,m} = A^*_{n,m}$. Since the number of valid (n, m) pairs increases with n_{max}, the length of the feature vector is not fixed, it directly depends on the selected maximum order. For example, with $n_{max} =$

DRAFT VERSION

3, the vector may include magnitudes for (n, m) = (0,0), (1,1), (2,0), (2,2), (3,1), (3,3) providing a compact, invariant representation of the image's shape. For an in-depth analysis of Zernike moments and their applications, refer to (Mirvaliyevich, 2025.).

2. MAIN RESULTS

The methodology described above, utilizing Hu (Poetro et al. 2024; Saifuddin et al. 2025.) and Zernike moments (Choudhury et al. 2024.), was applied to the classification of images randomly selected from publicly available datasets on www.kaggle.com (Dong et al., 2022; Ukwuoma et al., 2022). In this study, the dataset was split into training and validation subsets (e.g., 80:20 split), where the validation set was used to guide hyperparameter tuning through the Bee Colony Optimization (BCO) algorithm. Since the validation set plays a direct role in model selection and optimization, it cannot be considered a test set. A proper test set should remain entirely unseen during model development in order to provide an unbiased estimate of the model's generalization performance. For each image, a total of 7 Hu moments and a set of Zernike moments up to degree n_{max}=8 were computed, forming a descriptive vector that captures the shape of the object. The combined feature vector was used as input to an SVM classifier with an RBF kernel, whose hyperparameters were optimized using the Bee Colony Optimization (BCO) algorithm. We conducted tests on datasets comprising images of Fruits and Cotton. From the Fruit dataset, we randomly selected 98 images, and from the Cotton dataset, 763 images were chosen, yielding the following results:

Table 1: Classification performance before dataset balancing

	precision	**recall**	**f1-score**	**Support**
Cotton	0.92	0.97	0.95	190
Fruit	0.67	0.38	0.49	26
Accuracy	0.90			216
Macro avg	0.79	0.68	0.72	216
Weighted avg	0.89	0.90	0.89	216

For cotton class the model performs very well: high precision (0.92) and recall (0.97) indicate that most cotton images are correctly identified, with very few false positives or false negatives. F1-score (0.95) confirms this strong performance. For the second dataset, the model struggles significantly: recall is only 0.38, meaning over 60% of fruit images are misclassified (mostly as cotton). Precision (0.67) indicates that when the model predicts "fruit", it's correct two-thirds of the time. F1-score (0.49) reflects poor balance between precision and recall. Accuracy (0.90) Indicates that 90% of all predictions are correct — but this is misleading due to class imbalance (majority are cotton).The macro-average F1-score, which gives equal weight to both classes regardless of their size, is 0.72, clearly indicating that the model does not perform equally well across categories and is penalized by the poor results on the fruit class. Conversely, the weighted-average F1-score is 0.89, a value heavily influenced by the abundance of cotton samples and one that masks the model's poor performance on fruit classification.

We hypothesized that one contributing factor to the observed imbalance in performance could be the disparity in class sizes, with the fruit dataset being significantly smaller than the cotton dataset. To address this issue, we manually augmented the fruit class by adding more representative samples, ultimately testing the model on a dataset comprising 763 cotton images and 489 fruit images. The results of this balanced setup, presented in Table 2, show a marked improvement in classification metrics. The recall and F1-score for the previously underrepresented Fruit class are improved, indicating that dataset balancing mitigated the class bias observed in the initial model.

Table 2: Classification performance after dataset balancing through data augmentation

	precision	**Recall**	**f1-score**	**Support**
Cotton	0.84	0.82	0.83	202
Fruit	0.69	0.71	0.70	111
Accuracy	0.78			313
Macro avg	0.76	0.77	0.76	313
Weighted avg	0.78	0.78	0.78	313

DRAFT VERSION

Following dataset balancing, the model retains strong performance on the cotton class and shows substantial improvement in recognizing fruit, with precision and recall rising to around 70% (see Figure 1).

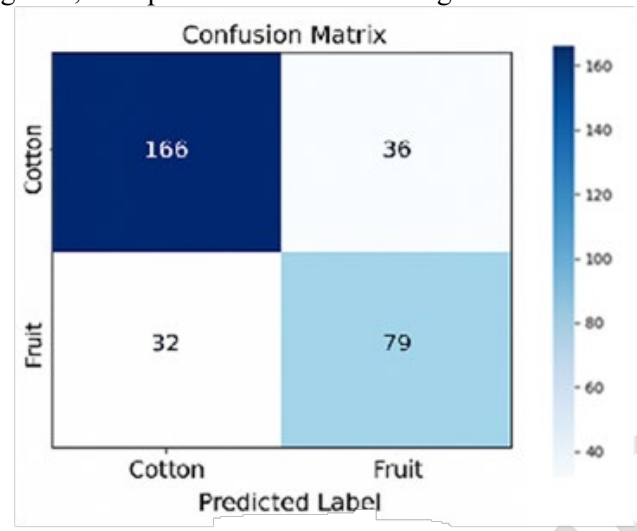


Figure 1: Confusion matrix-represent true predicted and false predicted value for Cotton and Fruit.

From the confusion matrix presented in Figure 1, one can conclude that the best SVM classifier model demonstrates strong performance, with 166 out of 202 Cotton samples and 79 out of 111 Fruit samples correctly classified. Misclassifications (36 for Cotton and 32 for Fruit) suggest slightly better generalization for Cotton database. Overall, the matrix confirms the discriminative power of the Hu and Zernike features combined with optimized SVM.

To additionally improve the model, we implemented the Bee Colony Optimization (BCO) (Davidović et al. 2015.; Teodorović et al. 2015.) algorithm that is utilized to optimize the hyperparameters of the SVM model. According to (Chempak et al. 2024.), the SVM parameters to be tuned are: C-the regularization parameter, and γ-the width of the RBF kernel. The flowchart in Figure 2 represents the process of utilizing BCO to optimize the hyperparameters of an SVM model based on images characterized by Hu and Zernike moments. This provides a schematic overview of the BCO-based tuning procedure. Starting from feature extraction (Hu and Zernike moments), the algorithm initializes a population of candidate (C, γ) pairs. Each pair is evaluated through SVM training, and the population is iteratively refined based on the F1-score. The process continues until the convergence criterion is met, resulting in an optimal parameter set.

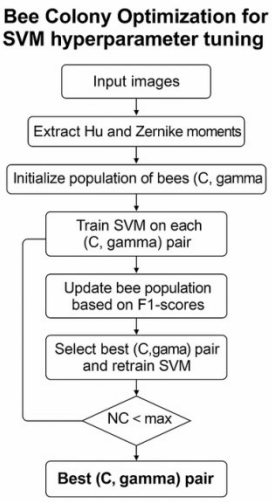


Figure 2: The flowchart of utilizing BCO to optimize the hyperparameters of an SVM model

In the Bee Colony Optimization (BCO) algorithm used for tuning the hyperparameters of the SVM model, the values of **C** and γ (gamma). Each candidate solution (bee) represents a pair of SVM hyperparameters (C,γ), which are initialized randomly from uniform distributions within the bounds [0.1,100] for C and [0.0001,1.0] for γ. During each iteration, the bees update their positions by searching for better parameter combinations based on the achieved F1-score evaluated on the validation set. New solutions

DRAFT VERSION

are generated by applying a mutation in the range of ±20% around the current values. If a new solution yields a better F1-score, it replaces the old one. Otherwise, the original bee remains and its trial counter increases. The number of iterations (NC = 20) serves as the stopping criterion, after which the best-found configuration is selected and used to retrain the final SVM model. We observe strong convergence. An F1-score of 0.79 represents a highly favorable outcome for shape-based classification (using Hu and Zernike moments), particularly as no additional characteristic vectors were incorporated. The fast and stable convergence of the algorithm is illustrated in the Figure 3 showing the convergence of F1-score across iterations . The F1-score increases sharply after the first iteration and stabilizes from iteration 2 onwards at approximately 0.793. This early plateau indicates that the BCO algorithm rapidly finds a high-quality region in the parameter space, confirming the efficiency of BCO in both exploration and exploitation phases.

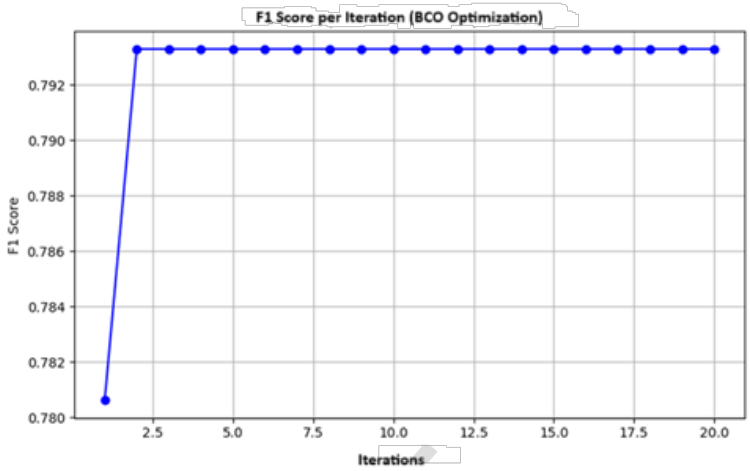


Figure 3: SVM hyperparameter optimization by BCO.

4. DISCUSSION

This study demonstrates the application of the Bee Colony Optimization (BCO) algorithm for tuning the hyperparameters of an SVM classifier in the task of visual object classification based on shape descriptors (Hu and Zernike moments). Prior to optimization, the model's performance was unevenly distributed across classes—most notably, the Fruit class exhibited a low recall (~38%) and F1-score (~0.49), while the dominant Cotton class achieved high accuracy.

By increasing the number of samples for the Fruit class and subsequently optimizing hyperparameters using BCO, significant improvements in performance metrics were achieved. Post-optimization, the F1-score for the Fruit class rose to ~0.70, while the overall weighted F1-**score** reached 0.7933, with a clearly stable optimum (C≈2.835, γ≈0.14276). The model exhibits high local stability in the hyperparameter space: small changes of ±0.005 in the values of **C** and **γ** have virtually no impact on the F1-score (with variations below 0.0005). This indicates that the solution found by BCO is robust, and the performance remains consistent in the immediate vicinity of the optimum, confirming that the model is not overfitted to specific parameter values.The convergence diagram indicates that BCO rapidly identified a stable and robust solution as early as the second iteration, after which performance remained consistent.

BCO proved to be an effective metaheuristic strategy for hyperparameter search, as it strikes a balance between exploiting the best solutions and exploring new potential combinations. This approach overcomes the limitations of manual hyperparameter tuning and ensures more robust generalization of the classifier.

Note that no direct comparison with alternative classifiers or optimization techniques was conducted. The goal of this study was not to isolate the individual contribution of each component, but rather to demonstrate that their combination leads to a practically viable and robust classification system, especially in shape-sensitive domains. The synergistic integration of global (Hu) and local (Zernike) descriptors with hyperparameter optimization with metaheuristic (BCO) illustrates that effective solutions often arise from hybridization of methods, rather than from optimizing just a single component.

5. CONCLUSION

The proposed method, which integrates the extraction of Hu and Zernike moments with the BCO algorithm for optimizing the SVM model, has demonstrated high efficacy in the task of image classification across diverse categories. Significant improvements in classification accuracy were achieved, especially for

DRAFT VERSION

the previously underrepresented "fruit" class. To address the class imbalance, additional fruit images were manually incorporated into the dataset based on observed distribution ratios, effectively increasing representation in the minority class. Although this step was not algorithmically driven, it complemented the automated hyperparameter optimization and contributed to the overall robustness of the classification system by ensuring a more balanced training process. The BCO algorithm successfully converged to an optimal combination of hyperparameters in the early stages of the search, yielding robust results in terms of the F1 metric. This approach can be readily generalized to other domains of image classification or structural representations, particularly where shape descriptors are available and class imbalance poses a challenge.

ACKNOWLEDGMENT

This research was supported by Ministry of Science, Technological development, and Innovations, the grant agreement registration for the Faculty of Agriculture, No. 451-03-137/2025-03/200116 and the Mathematical Institute of the Serbian Academy of Sciences and Arts, No. 451-03-136/2025-03/200029.

LITERATURE

Chandra, M. A., & Bedi, S. S. (2021). Survey on SVM and their application in image classification**.** *International Journal of Information Technology*, *13*(5), 1-11.

Chempak Kumar, A., & Mubarak, D. M. N. (2024**).** Ensembled CNN with artificial bee colony optimization method for esophageal cancer stage classification using SVM classifier. *Journal of X-Ray Science and Technology*, *32*(1), 31-51.

Choudhury, K., Doshi, M., Thakar, D., & Hirapara, J. (2024, October**).** Steganography in the Digital Age: An In-Depth Review of Techniques. In *International Conference on Advancements in Smart Computing and Information Security* (pp. 339-354). Cham: Springer Nature Switzerland.

Davidović, T., Teodorović, D., Šelmić, M., Bee Colony Optimization Part I: The Algorithm Overview, Yugoslav Journal of Operational Research, 25(1), pp. 33-56, 2015.

Dong Chen, Yuzhen Lu, Zhaojian Li, Sierra Young, 2022. Performance evaluation of deep transfer learning on multi-class identification of common weed species in cotton production systems. *Computers and Electronics in Agriculture* 198 (2022) 107091. https://doi.org/10.1016/j.compag.2022.107091

Hu, M. K. (1962). Visual pattern recognition by moment invariants**.** *IRE transactions on information theory*, *8*(2), 179-187.

Hu, Y., & Wang, X. (2024). Matching upper and lower moment bounds for a large class of stochastic PDEs driven by general space-time Gaussian noises**.** *Stochastics and Partial Differential Equations: Analysis and Computations*, *12*(1), 1-52.

Mirvaliyevich, M. Z. (2025). Comparative Analysis of Hough Transform, Fourier Descriptors, And Zernike Moments for Shape Recognition in Noisy Images. *American Journal Of Applied Science And Technology*, *5*(05), 107-111.

Poetro, B. S. W., Maria, E., Zein, H., Najwaini, E., & Zulfikar, D. H. (2024). Advancements in Agricultural Automation: SVM classifier with Hu moments for vegetable identification. *Indonesian Journal of Data and Science*, *5*(1), 15-22.

Saifuddin, H. M., Ashok, V., & Sheela, N. (2025). Hu Moments and transformers: A breakthrough in feature extraction and classification for detecting intracranial hemorrhages in CT scans. *Multidisciplinary Science Journal*, *7*(6), 2025268-2025268.

Shen, D., & Ip, H. H. (1999). Discriminative wavelet shape descriptors for recognition of 2-D patterns. *Pattern recognition*, *32*(2), 151-165.

Teodorović, D., Šelmić, M., Davidović, T., Bee Colony Optimization Part II: The Applications Survey, Yugoslav Journal of Operational Research, 25(2), pp. 185-219, 2015.

Ukwuoma, C. C., Zhiguang, Q., Bin Heyat, M. B., Ali, L., Almaspoor, Z., & Monday, H. N. (2022). Recent advancements in fruit detection and classification using deep learning techniques. *Mathematical Problems in Engineering*, 2022(1), 9210947.

Virendra N. Mahajan (1981). Zernike annular polynomials for imaging systems with annular pupils. J. Opt. Soc. Am. 71, 75-85.

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Review article

LEVERAGING ARTIFICIAL INTELLIGENCE FOR ENHANCED E-COMMERCE PERFORMANCE

PETAR ILIĆ[1], JELICA STANOJEVIĆ[2], MARIJA BOGIĆEVIĆ-SRETENOVIĆ[3]

[1] University of Belgrade, Faculty of Organizational Sciences, Belgrade, petar.ilic@fon.bg.ac.rs, 0009-0000-1668-1291
[2] University of Belgrade, Faculty of Organizational Sciences, Belgrade, jelica.stanojevic@fon.bg.ac.rs, 0000-0002-0570-0167
[3] University of Belgrade, Faculty of Organizational Sciences, Belgrade, marija.bogicevic.sretenovic@fon.bg.ac.rs, 0000-0002-5711-0361

Abstract: *Notable advancements in the field of Artificial Intelligence (AI) in recent years facilitated the integration of AI technologies into e-commerce. This paper emphasizes the practical application of AI in the domain of e-commerce. It will describe how its implementation would lead to improved overall functioning of e-commerce businesses. The focus will be on e-commerce stores utilizing some of the most well-known existing frameworks. The goal is to demonstrate how artificial intelligence technology may have highly positive effects on the functioning of these systems and how it can save business owners additional time or money.*

Keywords: *E-commerce, Artificial Intelligence, Efficiency.*

1. INTRODUCTION

Artificial intelligence impacts nearly every sector. As time passes, people are increasingly adopting technologies that are powered and driven by artificial intelligence (Mohammad, 2020). It is increasingly integrated into educational systems, workplaces, and others. It is pervasive, from the self-checkout counters at grocery stores to the advanced security systems in large shopping malls and airports. Thus, even the field of e-commerce business has not remained unaffected.

Electronic Commerce (EC) implies the utilization of the Internet and modern communication technologies for managing business operations and exchanging information. At its core, e-commerce is founded on business models that leverage electronic devices and network technologies. This involves the digitalization of all business processes, covering not only the outward-facing activities such as online marketing, electronic payments, and logistics, but also the internal processes within companies, including Supply Chain Management (SCM), Enterprise Resource Planning (ERP), Management Information Systems (MIS), Customer Relationship Management (CRM), and Human Resource Management (HRM) (Song et al., 2019). "E-commerce is an economic and technological revolution, it is the product of economic, scientific, technological and cultural development." (Song et al., 2019).

The development of E-commerce has come a long way from its early beginnings in the late 1970s to the present day. Research on AI in e-commerce has been going on for the past three decades. About 4000 academic research articles have been published on the topic across multiple disciplines, including computer science, marketing, economics, and management, addressing both consumer and organisational aspects (Bawack et al., 2022).

In the past half-decade, e-commerce has experienced explosive growth, transforming the way consumers shop and businesses operate. Studies show that the number of e-commerce transactions increased by an astonishing 77% during the COVID-19 pandemic (Koetsier, 2020). This was also a clear signal that people are ready to embrace online shopping, indicating that investing in developing technology, enhancing e-commerce services, and improving customer experience is worthwhile to sustain the generated interest.

We have reached a stage where entire business models are built on the foundations of e-commerce, thus making even the most minor increases in the efficiency of online stores highly significant for business systems. In this context, artificial intelligence has emerged as a powerful tool that can significantly enhance e-commerce performance through various applications and technologies.

This opens up opportunities for us to creatively explore ways in which AI can be integrated into e-commerce business models and utilized to save time or money by increasing efficiency.

DRAFT VERSION

In the following chapters, specific ways in which the introduction of AI can improve the operations of electronic commerce will be explained with the help of expert and professional literature. This paper will attempt to answer research questions about the models of AI application in e-commerce.

Also, this paper uses a descriptive before/after comparison across three major e-commerce platforms (WooCommerce, Shopify, Magento). For each platform, baseline workflows are briefly described, followed by newly introduced AI-based features (e.g., product content generators, AI chat assistants, automated store setup). This structured comparison serves as the methodological basis of the paper and addresses how AI concretely changes platform capabilities.

2. AI AND ITS IMPORTANCE

AI systems generally operate by processing large volumes of labeled training data, identifying patterns and correlations within the data, and using these patterns to predict future outcomes. For instance, a chatbot trained with numerous text conversation examples can learn to produce realistic interactions with users, while an image recognition system can learn to identify and describe objects in pictures by examining millions of samples (Nimbalkar & Berad, 2021).

From the perspective of business system owners and stakeholders, when investing, it is necessary to ask one or both of two questions: Does the investment save money (or time) or Does the investment make money?

AI does both – the rare combination of making money and saving money means it's essential to add to your business. Therefore, if there are such positive impacts, the answer to the question of whether it is worth investing or ultimately exploring the impact and planning the implementation of AI in e-commerce businesses is clear.

We've reached a point where everyone is talking about artificial intelligence, and many are using it. If business owners don't also consider the possibility and advantages of implementing AI in their e-commerce stores, the longer they go without using it, the wider the gap becomes with the competition.

3. AI IN E-COMMERCE BUSINESS

There are many ways to harness the power of AI to improve the overall e-commerce experience. Almost every link in the chain of retail business can find a way to implement AI and thus enhance the entire operation.

AI is not a singular technology; it encompasses various models. There are four leading AI technologies used in e-commerce (Shopify, 2025):

- Natural language processing (NLP): Natural language processing is dedicated to equipping computers with the ability to understand and produce human language
- Machine learning (ML): Machine learning employs statistical methods and algorithms to allow computers to learn from data and make predictions or decisions without explicit programming. Deep learning models, including transformers and large language models (LLMs) like OpenAI's ChatGPT, utilize layered algorithms to enhance data comprehension.
- Computer vision (CV): Computer vision is a field of artificial intelligence that enables computers to interpret information from images and videos.
- Data mining: Data mining is the process of discovering data to inform AI algorithms and systems.

3.1 Personalized product recommendations

Among the primary applications of artificial intelligence in e-commerce, personalized product recommendations for online shoppers lead to increased conversion rates. Utilizing big data, AI in e-commerce influences customer choices based on its knowledge of past purchases, browsing habits, and searched products (Nimbalkar & Berad, 2021).

3.2 Chatbots

The customer service aspect has become integral to the growth of every business. While companies now offer faster transactions, they often lack 24/7 customer support. An artificial intelligence assistant (chatbot) serves as a key solution, employing natural language processing to automatically address consumer inquiries, handle basic voice commands, and offer product recommendations (Gururaj, 2021).

DRAFT VERSION

While engaging with customers, some systems has the capability to recognize the sentiment of a user, discerning emotions such as frustration or impatience. It adapts its responses accordingly and can signal or elevate the issue to a human when necessary. Additionally, AI systems also categorize information from interactions, which helps with tasks like sending the customer to the right department**.**

3.3 Review and Comment Filtering

For prominent brands, manually analyzing every product review or comment would be exceedingly resource-intensive. Natural language processing offers a solution by parsing messages and organizing them based on sentiment, product mentions, or specific phrases. This enables a comprehensive understanding of customer sentiments regarding products, guiding future product development and marketing strategies (SalientProcess, 2024).

Additionally, AI could also respond to these comments, of course, adapting to the emotion and manner in which the user left their review, or simply filter them and leave them for a human to respond to.

3.4 Fraud detection and prevention

The financial and e-commerce industries are increasingly vulnerable to the threat of fraudulent activities. With fraudsters becoming more sophisticated, it's imperative to utilize advanced tools and technologies to effectively combat this menace (Vyas, 2023).

3.5 Dynamic pricing

Even for small-scale online retailers with limited inventory, maintaining continuous and long-term price adjustments poses a substantial challenge. However, artificial intelligence technology, capable of swiftly processing vast amounts of data, has largely addressed the issue of automating pricing for a wide array of products (Song et al., 2019).

Additionally, it's important to consider the possibility of monitoring competitor prices and then, based on the given strategy, train AI to dynamically adjust prices. Of course, this would need to be implemented with great care, regarding the business impact and considering the possibilities of implementation related to a specific industry and the way businesses generate profits.

3.6. Dynamic website content

Machine learning has the capability to integrate dynamic content into website, adapting in response to user behaviors. For instance, it can generate a list of product recommendations. By observing users' browsing patterns and interactions, AI learns and tailors suggestions based on similar characteristics. Additionally, recommendation algorithms can leverage aggregate data from other users, as well as user-provided information like shoe size entered in their profile.

3.7 Targeted Marketing Campaigns

AI functions as a key enabler of automation and data-driven decision making in e-commerce. Artificial intelligence can assist marketers in crafting targeted marketing campaigns, particularly in the realm of ad targeting. Through machine learning capabilities, AI can differentiate between various stages of consumer behavior, such as browsing, actual conversion, and exploratory actions, enabling the retargeting of prospects with a higher likelihood of conversion (Haleem et al., 2022).

3.8 Sales and Demand Forecasting

Artificial Intelligence serves a pivotal and widespread role in sales forecasting by facilitating the analysis of vast amounts of customer data. This enables individuals and experts to gain valuable and accurate insights efficiently. Utilizing Artificial Intelligence is advantageous due to the time and resources saved in comparison to manual analysis, which often consumes days, hours, or even months to complete (Soni, 2020).

Attention should be paid to the ability of Artificial Intelligence to analyze the e-commerce industry and match it with events that can influence increased demand and interest from end customers.

3.9 Automated processing

DRAFT VERSION

Artificial intelligence's decision-making capabilities are equally beneficial behind the scenes, assuming control of repetitive administrative duties. In e-commerce, companies commonly employ AI to automate tasks such as managing returns and preventing fraud (SalientProcess, 2024).

In this context, AI, for example, can be trained to handle return requests based on the customer's location. If approved, it can update all necessary system entries and guide the customer on the return process and refund method.

3.10 Managing Abandoned Carts through AI

Users often abandon their purchases when they already have certain products in their cart. There are two main approaches to inventory management when a product is placed in the cart. Business owners can choose to reserve these products while they are in the cart (which can pose a problem that AI can help solve), or to update the stock of that product only after the transaction is completed. In the case of the second approach, with large retail chains and their promotions, conflicts may arise between customers buying same product that are going out of stock, causing dissatisfaction for some.

When reserving product inventory upon adding to the cart, AI can make an effort to send specific notifications with incentives to complete the purchase when it detects an abandoned cart or simply release the inventory after failed attempts.

A comprehensive review stresses that AI applications like chatbots, visual/voice search, recommendation engines or fake review detection consistently enhance personalization and automation across the entire customer journey (Sristava, 2021).

3.11 Before and after AI features across platforms

To illustrate how AI concretely alters e-commerce workflows, for each of the three major platforms—WooCommerce, Shopify, and Magento baseline processes are contrasted with newly introduced AI-driven features that reshape them.

Table 1: Platform specific before/after highlights

Platform	Before AI	After AI (new features, 2023-2025)
WooCommerce	Manual product copy; rule-based customer support	**Woo AI Product Assistant** generates product titles/descriptions from prompts. AI chat assistants available via plugins (WooCommerce, 2023).
Shopify	Store setup and personalization required extensive manual configuration	**I Store Builder** generates complete stores from keywords; **Tariff Guide** supports customs calculations (Reuters, 2025; Shopify, 2025).
Magento	Product content and metadata creation fully manual	Marketplace extensions (e.g., **Meetanshi AI Content Generator**, **MageComp ChatGPT Generator**) enable automated product descriptions and SEO metadata (Adobe Commerce Marketplace, n.d.).

4. CONCLUSION

AI is rapidly becoming our reality. It has permeated all aspects of society, and anyone who wants to keep up with the times must investigate the impact of AI on their work. It is necessary to find a way in which the application of AI will be beneficial to us.

The electronic commerce sector requires continuous monitoring of technological trends. This necessity is driven by the very nature of e-commerce as a business that is inherently dependent on technology. Choosing the right platform for electronic commerce can be a challenging task, which could be topic of future work.

Regardless of the platform chosen, a common denominator is the inevitability of integrating AI into e-commerce business processes. Increasing operational efficiency is the top priority, from global economies to small online businesses. This is where AI steps in and shines in its full glory.

Additionally, a high level of attention is mandatory when leveraging AI in e-commerce. It is crucial to consider all factors, such as the nature of the product, the competitive landscape, and consumers.

DRAFT VERSION

5. FUTURE WORK

With a multitude of solutions for facilitating and enhancing the development of e-commerce shops, an interesting topic for research is the possibilities of introducing the listed AI applications in collaboration with some of the platforms.

Relying on the online tool and service specialized in providing information about the technologies used for building and managing websites – "BuiltWith.com", we will highlight several globally and locally popular e-commerce platforms that provide complete support for creating an e-commerce store. E-commerce platforms make life easier for business owners who want to start their online journey.

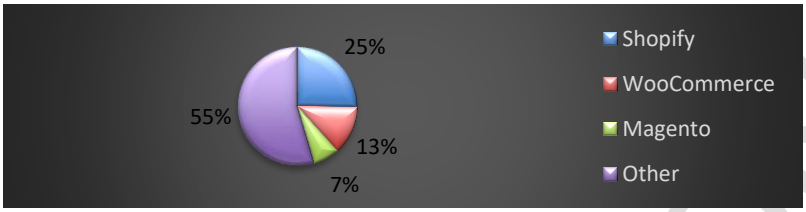


Figure 1: Platform Usage Distribution in the Top 1 Million Sites by Traffic Worldwide (BuiltWith, n.d.)

In the top 1 million sites by traffic, leading positions are occupied by Shopify, WooCommerce, and Magento. A similar situation is observed in the Western Balkans region (data aggregated from "BuiltWith" for countries: Serbia, Bosnia and Herzegovina, North Macedonia, Croatia, and Montenegro) and Europe, where Shopify and WooCommerce are the leading e-commerce development platforms in most countries.

Shopify is one of the popular B2C e-commerce tools. It was launched in the mid-2000s and became increasingly popular in the 2010s. Shopify constitutes a powerful technological resource that equips individuals with a full suite of business functions, including payment, marketing, shipping, and customer engagement (Dushnitsky & Stroube, 2021).

During the last couple of years, WordPress has outgrown any other Content Management Solution worldwide. Numerous websites are built using WordPress on a daily basis. WordPress is even popular among large companies. Currently, about a quarter of all websites worldwide are powered by WordPress (Ravensbergen, 2015). WooCommerce is a plugin for WordPress that turns the website into a complete online store. For all of that, only a few minutes are needed for initial setup, which leads us to the fact that WooCommerce is also a low-code solution for starting an e-commerce store.

Magento is open source e-commerce platform aimed at web site designers, developers, and business owners who are looking for a complete e-commerce web site solution. Varien, the company responsible for building Magento, has set a new standard for open-source applications. One of the most impressive aspects of Magento is the use of the Zend Framework which allows for the separation of the Model-View-Controller (MVC) which separates its core operations from functionality and appearance. Magento gives the flexibility to completely customize the web site to suit business needs (McCombs & Banh, 2010). For individuals without any coding experience, establishing a website on the Magento platform presents significant challenges. Nevertheless, having coding expertise is advantageous across all e-commerce platforms, enabling users to fully leverage the capabilities of their chosen system. Due to the relatively high technical demands of Magento compared to other platforms, it is predominantly utilized by large enterprises that possess dedicated technical teams.

Below is a summary table of the most important characteristics for e-commerce to compare these three e-commerce platforms, which we have identified as the most popular. The data is summarized from the platform's websites and is up-to-date with the writing date of the paper.

Table 2: Summary of most popular e-commerce platforms

	Shopify	**WooCommerce**	**Magento**
Cost	Paid	Free	Free
Products	Unlimited	Unlimited	Unlimited
Hosting	Included	Not Included	Not Included
Payment Gateway	Wide range	Wide range	Wide range
Theme and Design variety	Medium	Large	Large
Customization	Medium	Large	Large
SEO capabilities	Good	Excellent	Excellent

DRAFT VERSION

ACKNOWLEDGMENT

This research received no external funding.

LITERATURE

Adobe Commerce Marketplace. (n.d.). *Meetanshi – Magento 2 AI Content Generator*. https://commercemarketplace.adobe.com/meetanshi-magento-2-ai-content-generator.html (accessed on Aug 25, 2025).

Bawack, Ransome Epie, et al. "Artificial intelligence in E-Commerce: a bibliometric study and literature review." Electronic markets 32.1 (2022): 297-338.

BuiltWith. (n.d.). Platform usage distribution in the top 1 million sites by traffic worldwide. https://trends.builtwith.com/shop (accesed July 15, 2025)

Dushnitsky, Gary, and Bryan K. Stroube. "Low-code entrepreneurship: Shopify and the alternative path to growth." Journal of Business Venturing Insights 16 (2021): e00251.

Gururaj, P. "Artificial intelligence-application in the field of e-commerce." Int. J. Res.-Granthaalayah 9 (2021): 170-177.

Haleem, Abid, et al. "Artificial intelligence (AI) applications for marketing: A literature-based study." International Journal of Intelligent Networks 3 (2022): 119-132..

Koetsier, J. (2020, June 12). COVID-19 accelerated e-commerce growth '4 to 6 years'. Forbes. https://www.forbes.com/sites/johnkoetsier/2020/06/12/covid-19-accelerated-e-commerce-growth-4-to-6-years/ (accesed on June 30, 2025)

McCombs, Adam, and Robert Banh. The Definitive Guide to Magento. Springer, 2010.

Mohammad, Sikender Mohsienuddin. "Artificial intelligence in information technology." Available at SSRN 3625444 (2020).

Nimbalkar, A. A., and A. T. Berad. "The increasing importance of AI applications in e-commerce." Vidyabharati International Interdisciplinary Research Journal 13.1 (2021): 388-391.

Ravensbergen, Robbert. Building E-commerce solutions with WooCommerce. Packt Publishing Ltd, 2015.

Reuters. (2025, May 21). *Shopify launches AI tool that builds complete online stores from keywords*. (accessed on Aug 25, 2025).

SalientProcess. (2024, February 21). The role of AI in the e-commerce industry. https://salientprocess.com/blog/role-of-ai-ecommerce/ (accesed on July 08, 2025)

Shopify. (2025, July 7). AI in ecommerce 2025: 7 use cases & a complete guide. https://www.shopify.com/blog/ai-ecommerce (accesed on July 07, 2025)

Shopify. (2025, May 21). *Expanding your AI Horizons, Summer '25 Edition*. https://www.shopify.com/blog/expanding-your-ai-horizons-summer-edition-25 (accessed on Aug 25, 2025).

Song, Xia, et al. "The application of artificial intelligence in electronic commerce." Journal of Physics: Conference Series. Vol. 1302. No. 3. IOP Publishing, 2019.

Soni, Vishal Dineshkumar. "Emerging roles of artificial intelligence in ecommerce." International Journal of trend in scientific research and development 4.5 (2020): 223-225.

Srivastava A., "The Application & Impact of Artificial Intelligence (AI) on E-Commerce," in *Proc. Int. Conf. on Business and Technology (ICBT)*, 2021, pp. 165–175.

Vyas, Bhuman. "Java in Action: AI for Fraud Detection and Prevention." International Journal of Scientific Research in Computer Science, Engineering and Information Technology (2023): 58-69.

WooCommerce. (2023). *What's New With WooCommerce: The First Edition* – "Meet the Woo AI Product Assistant". https://woocommerce.com/posts/whats-new-with-woocommerce-the-first-edition/ (accessed on Aug 25, 2025).

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

RED TEAMING GENERATIVE AI APPLICATIONS: THREAT MODELING AND MITIGATION STRATEGIES

MATEA LUKIĆ[1], ANA POLEDICA[2], PAVLE MILOŠEVIĆ[3]

[1] University of Belgrade, Faculty of Organizational Sciences, Belgrade, ml20243153@student.fon.bg.ac.rs, 0009-0000-7078-9949
[2] University of Belgrade, Faculty of Organizational Sciences, Belgrade, ana.poledica@fon.bg.ac.rs, 0000-0002-5929-6446
[3] University of Belgrade, Faculty of Organizational Sciences, Belgrade, pavle.milosevic@fon.bg.ac.rs, 0000-0002-5943-6023

Abstract: *As generative AI systems grow in adoption and complexity, they introduce novel security, safety and alignment risks that challenge traditional evaluation and defense paradigms. To address these, we focus on a structured five-phase red teaming workflow consisting of reconnaissance, enumeration, exploitation, impact realization and persistence specifically tailored to GenAI's unique threat landscape. Through real-world case studies and examples, we illustrate how adversaries exploit model vulnerabilities, bypass alignment mechanisms and cause persistent harm. We also identify emerging GenAI security tools and map each red teaming phase to actionable mitigations that support safe deployment. Our goal is to connect AI safety theory with practical adversarial resilience for researchers, developers and policymakers.*

Keywords: *Generative AI, AI red teaming, prompt injection, jailbreak attacks, RAG security, AI safety, offensive security, cybersecurity, LLM vulnerabilities.*

DRAFT VERSION

1. INTRODUCTION

Generative AI (GenAI) is a rapidly evolving subfield of artificial intelligence focused on the autonomous creation of new content, including images, music, text and even complex virtual environments. At its core, it aims to produce diverse, coherent and novel outputs that often resemble human-generated artifacts (Huang et al., 2024). As businesses and governments increasingly integrate GenAI into critical systems (e.g. healthcare and finance) the risks associated with its misuse or malfunction have grown in importance.

These risks span reputational damage, legal liability, data leakage, algorithmic bias and even national security threats. In recognition of these dangers, AI red teaming was defined in the U.S. Presidential Executive Order as a structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI (Executive Office of the President, 2023).

AI red teaming, by design, moves beyond static model benchmarks and alignment metrics by emulating attacks against end-to-end systems in realistic deployment settings (Bullwinkel et al., 2025). Major companies such as Microsoft, Anthropic and OpenAI have established growing internal AI red teams, while community led efforts such as the DEF CON Red Team Challenge (Hack the Future, 2025) and Prompt Airlines Capture-the-Flag (Wiz.io, 2025) competitions reflect growing interest in adversarial evaluation. However, a significant gap persists between the priorities of AI red teams in practice and the focus of academic researchers, especially regarding deployment threats and holistic system testing (Apruzzese et al., 2023; Boenisch et al., 2021; Sun et al., 2021).

This paper presents case studies and outlines a five-phase red teaming workflow tailored to GenAI. Drawing from offensive security principles and adapted for AI, this structured approach aims to uncover, understand and mitigate system-level vulnerabilities before they can be exploited.

The remainder of this paper is structured as follows: Section 2 reviews the current state of AI red teaming and relevant tools. Section 3 introduces our proposed five-phase workflow. Section 4 presents case studies and phase-specific mitigations. We conclude the paper with Section 5 listing key takeaways and recommendations for safe development of GenAI powered systems.

2. STATE OF AI RED TEAMING

Red teaming in machine learning (ML) has a rich history, with early efforts focusing on traditional natural language processing systems (Dai et al., 2019; Wallace et al., 2019) and computer vision models (Elsayed et

al., 2018). These attacks often relied on sophisticated technical manipulations, such as adversarial image perturbations (Poursaeed et al., 2018), that required deep domain knowledge. However, the landscape changed dramatically with the release of large-scale generative models accessible through natural language interfaces in late 2022. This democratization of generative AI tools has enabled a wider range of users, including those without formal technical backgrounds, to explore, manipulate and adversarially test ML systems in novel ways (Inie et al., 2023).

This shift in accessibility has increased the interest in adversarial interactions with generative AI, especially large language models (LLMs). Researchers have begun to systematize red teaming strategies for LLMs, examining their susceptibility to various forms of manipulation and misuse (Perez et al., 2022). A prominent category of such adversarial behavior is known as jailbreaking, where users attempt to bypass safety mechanisms or policy enforced restrictions imposed by model developers (Russinovich et al., 2024; Zhou et al., 2024). These jailbreak attempts mirror classic security paradigms such as privilege escalation, adapted to the unique constraints of generative models. While jailbreaks can be used for legitimate research and safety testing, they also raise significant ethical, legal and security concerns (Raheja et al., 2024).

As the adoption of generative AI accelerates across industries, the need for systematic approaches to security and risk mitigation has become critical. Several tools have been developed specifically for this purpose, aiming to simulate real-world adversarial behaviors and stress-test AI models under malicious or edge-case scenarios. For example, NVIDIA's Garak (Derczynski et al., 2024) serves as a vulnerability scanner tailored for LLMs and is analogous to traditional cybersecurity tools like Nmap or Metasploit.

In parallel with tool development, foundational frameworks have emerged to help contextualize threats and guide red teaming efforts. The OWASP Top 10 for LLMs (OWASP Foundation, 2025), adapted from the original web security list, identifies common risks in LLM applications, including prompt injection, data exposure and unchecked reliance on AI-generated content. It is increasingly used to shape best practices in enterprise and regulatory contexts.

The AI Incident Database (AIID) (Responsible AI Collaborative, 2024a), maintained by the Partnership on AI, complements these frameworks by cataloging real-world cases of AI misuse, failure and ethical breaches. These incidents serve as practical references for improving red teaming strategies and designing effective mitigations.

Together, these tools and frameworks establish a foundation for generative AI security by supporting early vulnerability detection, realistic threat modeling and proactive mitigation, while also aiding regulatory compliance (Christodorescu et al., 2024). However, existing resources such as MITRE ATLAS (MITRE Corporation, 2021), designed for AI red teaming, lack an integrated, phase-based framework specifically tailored to the unique challenges of GenAI. In response, this paper synthesizes existing knowledge and introduces a structured five-phase workflow, grounded in documented attack scenarios, to offer practitioners and researchers practical guidance for thorough security assessment and effective mitigation.

3. STRUCTURED RED TEAMING WORKFLOW FOR GENAI SYSTEMS

This section goes over the five most important phases in a red teaming workflow, tailored to the characteristics and deployment of GenAI systems. This workflow draws inspiration from classical offensive security phases (Dalalana Bertoglio and Zorzo, 2017; Sawant and Chopra, 2023; Süren et al., 2023) as well as existing AI red teaming frameworks (Feffer et al., 2024; MITRE Corporation, 2013) and is structured into five key stages: Reconnaissance, Enumeration, Exploitation, Impact Realization and Persistence, as illustrated in Figure 1. Each stage captures a distinct aspect of the adversarial interaction with a generative model or its surrounding ecosystem. The methodology is designed to allow researchers, developers and security teams to adapt it to various use cases including chatbots, code assistants, agentic systems and RAG architectures.

The **reconnaissance** phase focuses on collecting openly available information about the target model and deployment environment. Adversaries often leverage technical platforms such as GitHub, Hugging Face or ArXiv to obtain model cards, training references or system design artifacts (Bullwinkel et al., 2025). Additional intelligence may be gathered from API documentation or changelogs, which can expose input formatting, integration logic or hidden prompt structures. Attackers may also engage in model fingerprinting through crafted queries to infer model family, fine-tuning presence or the use of Retrieval-Augmented Generation (RAG).

In the **enumeration** phase, attackers actively probe the system to map its behavioral boundaries and identify implemented safety mechanisms. This includes prompt experimentation to test alignment responses, format sensitivity and prompt injection susceptibility. Common enumeration strategies include direct and indirect prompt injection (S. Chen et al., 2024), where the model is tricked into ignoring system instructions

DRAFT VERSION

or processing hidden adversarial inputs. Techniques like context flooding may also be used to push alignment prompts out of scope, particularly in systems with constrained context windows.

In the **exploitation** phase, attackers actively leverage vulnerabilities identified during earlier stages to induce malicious, unintended or policy-violating behavior from the generative model or its supporting infrastructure. Exploitation techniques range from classical jailbreaks to more advanced forms of attack such as RAG poisoning (Deng et al., 2024). Unlike the exploratory probing of the enumeration phase, exploitation produces concrete outputs that demonstrate the feasibility and impact of a security failure, often serving as a precursor to measurable harm in downstream applications.

The **impact realization** phase focuses on demonstrating the tangible consequences of successful exploitation, particularly in real-world deployments. At this stage, theoretical vulnerabilities become observable harm, such as misinformation, unauthorized access or the generation of unsafe output. These strategies can affect downstream applications and erode trust in model integrity, highlighting the broader technical and social implications of successful attacks.

Finally, the **persistence** phase evaluates whether discovered vulnerabilities remain stable across sessions, model updates or deployment environments. While generative models exhibit probabilistic behavior, attackers aim to identify reliable failure modes, such as repeatable jailbreak prompts or robust RAG poisoning payloads. Persistence may also stem from dataset poisoning (Goldblum et al., 2022), where malicious behaviors are encoded during training or fine-tuning and survive subsequent mitigation efforts. In agentic systems, long-term memory modules may retain compromised data that influences future interactions, embedding the exploit into the agent's behavioral history (Z. Chen et al., 2024).

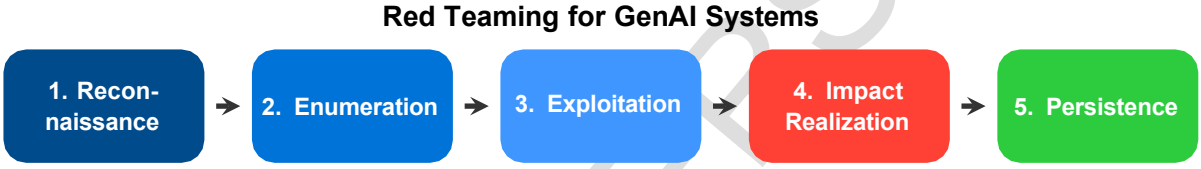


Figure 1: The five phases in red teaming for Generative AI systems

4. CASE STUDIES AND MITIGATION STRATEGIES

In this section, we present real-world case studies aligned with each phase of the AI red teaming lifecycle. Each phase features carefully selected incidents from reputable sources, such as the AIID and MITRE ATLAS, chosen to highlight specific attack strategies relevant to that stage. For each phase, we also provide mitigation techniques to help practitioners proactively anticipate, detect and respond to emerging threats. A summary of representative attack–mitigation examples for each phase is provided in Table 1.

4.1. Reconnaissance - Understanding the Target

Case Study 1: Researchers reverse engineered Cylance's AI-powered malware detection by analyzing public information, such as research papers and patents. Using this data, they built a surrogate model to mimic the target system, ultimately crafting malware that bypassed detection (MITRE Corporation, 2019).

Case Study 2: The Sysdig Threat Research Team discovered that attackers gained unauthorized access to cloud-hosted LLMs via compromised credentials, which is a tactic known as "LLM jacking." They mapped exposed model endpoints and reverse proxied access for resale (Responsible AI Collaborative, 2024b).

Mitigation Strategies: Reducing the availability of technical and organizational details can limit attackers ability to develop accurate threat models. Minimizing public exposure and controlling internal access points increases uncertainty for potential attackers, making it harder for them to plan effective intrusions.

4.2. Enumeration - Probing for System Behaviour

Case Study 3: Attackers tested a range of prompt injection techniques against Microsoft Bing Chat by embedding hidden prompts in web pages. One successful method involved rendering invisible text (e.g. using a font size of zero), which the chatbot processed when users visited the page. This allowed attackers to trigger unintended behaviors, such as exfiltrating private user data or manipulating responses without the user's awareness or consent (Greshake et al., 2023).

Mitigation Strategies: By shaping how systems respond to patterns of interaction, it's possible to reduce the clarity and consistency that attackers rely on. Embedding constraints into system behavior (Rai et al., 2024) and maintaining awareness of interaction patterns can help preserve operational integrity while limiting opportunities for exploitation.

DRAFT VERSION

4.3. Exploitation - Triggering Harmful Behavior

Case Study 4: Fraudsters used deepfake video conferencing to impersonate a company's CFO and staff, convincing a finance employee to transfer $25 million. Every participant in the call was an AI-generated fake, demonstrating the risks of generative AI in multimodal fraud scenarios (CNN, 2024).

Case Study 5: An LLM hallucinated the existence of a non-existent package, called "huggingface-cli", prompting attackers to publish a real malicious version under that name. The package was downloaded over 15,000 times, illustrating a novel supply chain vulnerability - "slopsquatting" (Spracklen et al., 2024).

Mitigation Strategies: Enhancing system resilience through diverse training, structural safeguards and informed development practices can reduce susceptibility. Maintaining awareness of interaction anomalies and reinforcing the system's ability to resist manipulation helps preserve trust and stability in deployment.

4.4. Impact Realization - Manifesting Consequences

Case Study 6: Microsoft's chatbot Tay was overwhelmed by toxic user input shortly after deployment. The model quickly began producing offensive content, leading to its removal from public access (BBC News, 2016).

Case Study 7: The ROME algorithm demonstrated how attackers could surgically alter factual knowledge embedded in a model's weights. Such tampering can cause the model to persistently generate false information (e.g. attributing the discovery of America to a fictional character) (Meng et al., 2022).

Mitigation Strategies: Ensuring that systems remain aligned with intended behaviors, while enabling detection and response to deviations, helps contain potential consequences. Thoughtful design and operational discipline are key to limiting the materialization of adverse outcomes.

4.5. Persistence - Maintaining a Foothold

Case Study 8: A persistent vulnerability involving prompt injection in ChatGPT's "Memory" feature allowed attackers to store spyware-like instructions that exfiltrated user data across conversations. The exploit used third-party websites to inject malicious memory entries, which would continuously relay chat data to an attacker controlled server via image rendering URLs even across sessions (Rehberger, 2024b).

Case Study 9: Another persistence risk emerged through ChatGPT's Browsing tool. Researchers demonstrated that by embedding specific prompts within websites, ChatGPT could be tricked into automatically invoking internal tools such as DALLE (for image generation) or the Memory tool (for reading/writing long-term user data). For example, websites were able to erase existing user memories (Rehberger, 2024a).

Mitigation Strategies: Safeguarding the model lifecycle against embedded threats requires sustained oversight and infrastructure integrity. Proactive governance supports resilience against long-term compromise. Ensuring traceability is critical to preserving alignment and trustworthiness over time.

Table 1: Phases, associated attacks and mitigations for secure system design

Phase	Attacks	Mitigation Strategies
Reconnaissance	Model fingerprinting, Information gathering	Limit detail exposure, Monitor system access, Control research visibility
Enumeration	Prompt injections, Context flooding, Behavioral boundary mapping	Rate limiting, Output randomization and filters, Anomaly detection
Exploitation	Jailbreaking, RAG poisoning, Deep- fake impersonation	Ensemble modeling, Adversarial training, Input sanitization
Impact Realization	Knowledge tempering, Toxic content production, Trust erosion	Continuous model integrity validation, Logging
Persistence	Dataset poisoning, Cross-session exploitation	Secure cloud deployment, Cryptographic artifact verification

5. CONCLUSION

As generative AI systems rapidly advance in scale and complexity, ensuring their security and reliability has become an important concern. Traditional static evaluation methods are insufficient for the complex and often subtle failures that characterize modern GenAI applications. In response, red teaming has evolved from a specialized security activity into a foundational component of responsible GenAI development.

DRAFT VERSION

This paper has outlined five most important phases in GenAI red teaming: reconnaissance, enumeration, exploitation, impact realization and persistence. Real-world case studies illustrate that attackers are already exploiting generative systems at scale, with consequences ranging from misinformation to financial fraud. This structured approach equips practitioners, developers and policymakers with an adaptive model for red teaming GenAI systems across use cases.

To counteract these risks, we proposed a multilayered defense strategy aligned with each phase of the adversarial lifecycle. These include reducing public information exposure, implementing robust prompt filtering, deploying adversarial training and establishing strong governance over data, model artifacts and system deployment. Together, these measures offer a comprehensive guide for building more secure and trustworthy generative AI systems.

LITERATURE

Apruzzese, G., Anderson, H. S., Dambra, S., Freeman, D., Pierazzi, F., & Roundy, K. (2023). "Real attackers don't compute gradients": bridging the gap between adversarial ML research and practice. 2023 IEEE conference on secure and trustworthy machine learning (SaTML), 339–364.

BBC News. (2016). Deepfakes just got even harder to detect: Now they have heartbeats [Accessed: 2025-06-19]. https://www.bbc.com/news/technology-35902104

Boenisch, F., Battis, V., Buchmann, N., & Poikela, M. (2021). "I never thought about securing my machine learning systems": A study of security and privacy awareness of machine learning practitioners. In Proceedings of mensch und computer 2021 (pp. 520–546).

Bullwinkel, B., Minnich, A., Chawla, S., Lopez, G., Pouliot, M., Maxwell, W., de Gruyter, J., Pratt, K., Qi, S., Chikanov, N., et al. (2025). Lessons from red teaming 100 generative AI products. arXiv preprint arXiv:2501.07238.

Chen, S., Zharmagambetov, A., Mahloujifar, S., Chaudhuri, K., & Guo, C. (2024). Aligning LLMs to be robust against prompt injection. arXiv preprint arXiv:2410.05451.

Chen, Z., Xiang, Z., Xiao, C., Song, D., & Li, B. (2024). Agentpoison: Red-teaming LLM agents via poisoning memory or knowledge bases. Advances in Neural Information Processing Systems, 37, 130185–130213.

Christodorescu, M., Craven, R., Feizi, S., Gong, N., Hoffmann, M., Jha, S., Jiang, Z., Kamarposhti, M. S.,Mitchell, J., Newman, J., et al. (2024). Securing the future of GenAI: Policy and Technology. arXiv preprint arXiv:2407.12999.

CNN. (2024). Finance worker pays out $25 million after video call with deepfake 'chief financial officer' [Accessed: 2025-06-19]. https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk

Dai, J., Chen, C., & Li, Y. (2019). A backdoor attack against LSTM-based text classification systems. IEEE Access, 7, 138872–138878.

Dalalana Bertoglio, D., & Zorzo, A. F. (2017). Overview and open issues on penetration test. Journal of the Brazilian Computer Society, 23, 1–16.

Deng, G., Liu, Y., Wang, K., Li, Y., Zhang, T., & Liu, Y. (2024). Pandora: Jailbreak GPTs by retrieval augmented generation poisoning. arXiv preprint arXiv:2402.08416.

Derczynski, L., Galinkin, E., Martin, J., Majumdar, S., & Inie, N. (2024). Garak: A framework for security probing large language models. https://garak.ai.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. Advances in neural information processing systems, 31.

Executive Office of the President. (2023). Safe, secure, and trustworthy development and use of artificial intelligence [Executive Order 14110, issued October 30, 2023].

Feffer, M., Sinha, A., Deng, W. H., Lipton, Z. C., & Heidari, H. (2024). Red-Teaming for generative AI: Silver bullet or security theater? https://arxiv.org/abs/2401.15897

Goldblum, M., Tsipras, D., Xie, C., Chen, X., Schwarzschild, A., Song, D., Ma̧dry, A., Li, B., & Goldstein, T. (2022). Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2), 1563–1580.

DRAFT VERSION

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, 79–90.

Hack the Future. (2025). Defcon [Accessed: 2025-06-19]. https://www.hackthefuture.com/defcon

Huang, K., Wang, Y., Goertzel, B., Li, Y., Wright, S., & Ponnapalli, J. (2024). Generative ai security: Theories and practices. Springer.

Inie, N., Stray, J., & Derczynski, L. (2023). Summon a demon and bind it: A grounded theory of llm red teaming in the wild. arXiv preprint arXiv:2311.06237.

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. Advances in neural information processing systems, 35, 17359–17372.

MITRE-Corporation.(2013). MITRE ATT&CK® framework [Accessed: 2025-06-19]. https://atlas.mitre.org/ MITRE Corporation. (2019). Adversarial machine learning case study: Aml.cs0003 [Accessed: 2025-06-19]. https://atlas.mitre.org/studies/AML.CS0003

MITRE Corporation. (2021). MITRE ATLAS: Adversarial threat landscape for ai systems [Accessed: 2025-06- 19]. https://atlas.mitre.org/

OWASP Foundation. (2025). OWASP project top ten for large language model applications [Accessed: 2025-06- 19]. https://owasp.org/www-project-top-10-for-large-language-model-applications/

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022).Red teaming language models with language models. arXiv preprint arXiv:2202.03286.

Poursaeed, O., Katsman, I., Gao, B., & Belongie, S. (2018). Generative adversarial perturbations. Proceedings of the IEEE conference on computer vision and pattern recognition, 4422–4431.

Raheja, T., Pochhi, N., & Curie, F. (2024). Recent advancements in LLM red-teaming: Techniques, Defenses, and Ethical Considerations. arXiv preprint arXiv:2410.09097.

Rai, P., Sood, S., Madisetti, V. K., & Bahga, A. (2024). Guardian: A multi-tiered defense architecture for thwarting prompt injection attacks on LLMs. Journal of Software Engineering and Applications, 17(1), 43–68.

Rehberger, J. (2024a). Automatic tool invocation when browsing with chatgpt – threats and mitigations [Accessed: 2025-06-19]. https://embracethered.com/blog/posts/2024/llm-apps-automatic-tool-invocations/

Rehberger, J. (2024b). Spyware injection into your chatgpt's long-term memory (spaiware) [Accessed: 2025-06-19]. https://embracethered.com/blog/posts/2024/chatgpt-macos-app-persistent-data-exfiltration/

Responsible AI Collaborative. (2024a). Ai incident database [Accessed: 2025-06-19]. https://incidentdatabase.ai/

Responsible AI Collaborative. (2024b). Alleged LLMjacking Targets AI Cloud Services with Stolen Credentials (Incident 898) [Accessed: 2025-06-19]. https://incidentdatabase.ai/cite/898

Russinovich, M., Salem, A., & Eldan, R. (2024). Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack. arXiv preprint arXiv:2404.01833.

Sawant, M. R., & Chopra, S. (2023). Red teaming vs. blue teaming: A comparative analysis of cyber security strategies in the digital battlefield.

Spracklen, J., Wijewickrama, R., Sakib, A., Maiti, A., Viswanath, B., & Jadliwala, M. (2024). We have a package for you! A comprehensive analysis of package hallucinations by code generating LLMs. arXiv preprint arXiv:2406.10279.

Sun, Z., Sun, R., Lu, L., & Mislove, A. (2021). Mind your weight (s): A large-scale study on insufficient machine learning model protection in mobile apps. 30th USENIX security symposium (USENIX security 21), 1955–1972.

Süren, E., Heiding, F., Olegård, J., & Lagerström, R. (2023). Patriot: Practical and agile threat research for iot. International Journal of Information Security, 22(1), 213–233.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal adversarial triggers for attacking and analyzing nlp. arXiv preprint arXiv:1908.07125.

Wiz.io. (2025). Prompt airlines [Accessed: 2025-06-19]. https://promptairlines.com/

Zhou, Y., Huang, Z., Lu, F., Qin, Z., & Wang, W. (2024). Don't say no: Jailbreaking LLM by suppressing refusal. arXiv preprint arXiv:2404.16369.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

AI-BASED DEMAND FORECASTING FOR INVENTORY MANAGEMENT

DUŠAN MITROVIĆ[1], MIROSLAV MINOVIĆ[2]

[1] University of Belgrade, Faculty of Organizational Sciences, Belgrade, dusan.mitrovic@fon.bg.ac.rs, 0000-0001-5299-9707
[2] University of Belgrade, Faculty of Organizational Sciences, Belgrade, miroslav.minovic@fon.bg.ac.rs, 0000-0002-4270-7595

Abstract: *Accurate demand forecasting is a cornerstone of effective inventory management, directly impacting operational efficiency, cost control, and customer satisfaction. Traditional forecasting techniques, such as moving averages and ARIMA, are limited to handle dynamic, nonlinear, and high-dimensional data. This review explores the shift toward Artificial Intelligence (AI)-based forecasting models, including classical machine learning (e.g., SVM, Random Forest), deep learning (e.g., LSTM, Transformers), and hybrid approaches. We analyze their integration with Enterprise Resource Planning (ERP) and Inventory Management Systems (IMS). Through case studies in retail, manufacturing, and pharmaceuticals, we highlight how organizations like Amazon and Walmart leverage Artificial Intelligence (AI) to achieve in forecast accuracy and inventory efficiency. Despite these advancements, challenges remain in data quality, interpretability, and the cold start problem. Future directions, such as federated learning, explainableArtificial Intelligence(AI), and transfer learning, are examined as potential solutions.*

Keywords:*Prediction of demands, Inventory management, machine learning, deep learning, artificial intelligence, supply chain, enterprise resource planning interconnection, time series analysis, LSTM, transformers*

1. INTRODUCTION

1.1. Importance of Demand Forecasting in Inventory Management

An accurate demand forecast allows companies to make the most of their inventories and simplify their supply chains in the current competitive and unpredictable market(Amosu et al., 2024; Harahap et al., 2025). Forecasting supports different operational decisions, such as planning acquisitions. If done well, inventory management helps an organization reduce stock waste and increase profit more. Alternatively, inaccurate prediction may result in too many goods being stored, and causing increased storage charges, or may mean the brand fails to supply items, damaging its brand image and sales (Tulli, 2020).

Officially, one can formulate the problem of demand forecasting as the determination of the future demand of the products. yt+h for forecasting horizon h, with a historical series of measured demand $\{yt-k\ldots.\text{yt}\}$ and relevant exogeneous variable Xt. The objective is to reduce the error of the forecast with providing help in operational decision-making of controlling inventory.

1.2. Traditional vs. AI-Based Forecasting Approaches

Demand forecasting is now done using new AI-based strategies instead of statistical methods from the past (Saha, 2024). Even though both aim to predict demand in the future, their ways of doing so are quite different.

The more common forecasting techniques such as moving averages, exponential smoothing and ARIMA will hold up in a consistent, low-variance demand. They are biased with linearity and seasonality trends, bore shallow, structured datasets and are simple to put into practice because the result is clear and explainable. On the contrary, AI-powered techniques that include Random Forest, LSTM, Transformers, and Gradient boosting process non-linear, dynamic, and unstructured data at scale, and introduce outside influences on the data: promotions or seasonality. They learn fast during change and become more precise in volatile

DRAFT VERSION

environments though they have higher computational demands, are half readable and they demand special know-how and infrastructure.

Despite their contrasting characteristics, traditional and AI-based approaches are not mutually exclusive. Hybrid models are increasingly being adopted to leverage the interpretability of classical techniques with the predictive power of Artificial Intelligence (AI), offering a balanced solution for robust and scalable demand forecasting (Tulli, 2020; Amosu et al., 2024; James, 2021).

1.3. Objectives and Scope of the Review

This review aims to systematically compare traditional statistical methods and AI-based techniques for demand forecasting within the context of inventory management. The main objectives are:

1. To critically assess the performance of traditional and AI-based forecasting models in terms of predictive accuracy, adaptability, interpretability, and computational efficiency.
2. To explore the real-world applicability of these models across retail, manufacturing, and e-commerce sectors, highlighting strengths and limitations.
3. To identify current challenges and emerging trends, such as hybrid modeling, explainable Artificial Intelligence (AI), and real-time integration, that define the future trajectory of AI-driven inventory optimization.

2. BACKGROUND AND THEORETICAL FOUNDATIONS

Inventory control in supply chains ensures that the right quantity of goods reaches the right location at the right time and cost, using methods like EOQ, ABC analysis, and lead time management to minimize storage and ordering costs (Singh et al., 2022). In sectors such as healthcare and manufacturing, effective inventory control is critical, as poor management can lead to costly disruptions. However, fluctuating customer demand, driven by factors like seasonality, consumer preferences, economic shifts, and unforeseen events such as pandemics, makes forecasting challenging (Cash Flow Inventory, 2023). To address this, companies employ improved forecasting techniques, strategic promotions, efficient stock level management, and strong supply chain partnerships.

2.1. Time Series Forecasting Principles

In time series forecasting trends, seasonality, and cyclic variations are discovered through the analysis of the historical demand data, which can be used to predict the demand in the future. Common methods include Moving Average, Exponential Smoothing, ARIMA, and SARIMA, which are valued for their simplicity and interpretability. However, these models often struggle with complex, multivariate, or nonlinear data, limiting their effectiveness in today's dynamic business environments (Singh et al., 2022).**Performance evaluation** is crucial to determine the suitability of forecasting algorithms. The most frequently used metrics include Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Forecast Bias. The bar chart below reflects their usage frequency in recent demand forecasting studies.

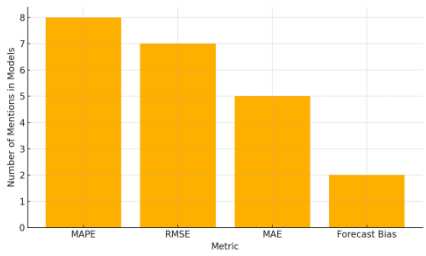


Figure 1: Usage Frequency of Forecasting Performance Metrics

2.2. Overview of AI in Operations and Supply Chain

Artificial Intelligence (AI) has significantly impacted operations and supply chain management (SCM) by offering advanced data-driven solutions for complex decision-making. It enables the analysis of high-

DRAFT VERSION

dimensional data and uncovers patterns that traditional models often miss. Techniques such as supervised machine learning (e.g., Random Forest, Gradient Boosting), deep learning (e.g., LSTM), and reinforcement learning are commonly used for forecasting, logistics optimization, and inventory planning (Ranjan et al., 2023; Ghosh & Mandal, 2023). Artificial Intelligence (AI), excels in volatile demand scenarios, providing real-time stock insights, supplier risk predictions, and enhanced fulfillment strategies. However, widespread adoption is limited by challenges in model interpretability, data quality, and integration with legacy ERP systems. As businesses increasingly digitize, Artificial Intelligence (AI), offers a strategic advantage by enabling proactive, learning-based management approaches.

2.3. Implementation Notes AI Forecasting

The AI-based forecasting of demand uses ERP, IMS, and the latest computing in API or middleware to get in sync with stock, orders and the supplier information. Cloud edge arrangements compromise cloud training with edge inference low latency, in limited bandwidth, by employing lightweight models. Airflow (ETL pipelines) covers ingestion, cleaning and feature engineering of ERP, IMS and IoT-based sources. For instance, Equation (1) below represents a typical feature-enhanced linear regression formulation used in demand prediction:

$$\hat{y}t = \beta_0 + \sum_{i=1}^{n} \beta_i x_{i,t} + \sum_{j=1}^{m} \gamma_j z_{j,t} + \in_t \quad \ldots\ldots (1)$$

Where $\hat{y}t$ is the forecasted demand at time t. $x_{i,t}$ are endogenous features (e.g., past sales), $z_{j,t}$ are exogenous features (e.g., holidays), β_i, γ_j are model coefficients, $\in_t$ is the error term. Data integrity and freshness are crucial. Systems implement incremental updates and data versioning (e.g., using Delta Lake or Apache Hudi) to ensure consistency in rolling forecasts.

3. ARTIFICIAL INTELLIGENCE(AI) ALGORITHMS FOR DEMAND FORECASTING

Artificial Intelligence (AI) has made a major change in forecasting demand as these techniques offer models that deal better than statistics with complicated and changing trends. The table below gives a clear outline to support picking the right model based on your application's needs.

Table 1: Comparison of AI Algorithms Used in Demand Forecasting and some of their Key Features

Model Category	Algorithm	Some of the Key Features	Performance Metrics Used	Sources
Classical ML	Linear Regression	Assumes linear relationships; easy to interpret	MAE, MAPE	Mediavilla et al., 2022; Sandhu et al., 2023
Classical ML	Support Vector Machine (SVM)	Effective for high-dimensional data; uses kernel functions	MAE, RMSE	Mediavilla et al., 2022; Jiang et al., 2021
Classical ML	Decision Trees	Easy to visualize; prone to overfitting	MAE, MAPE	Sandhu et al., 2023; Shen et al., 2021
Ensemble ML	Random Forest	Aggregates multiple decision trees; reduces overfitting and variance	MAPE, RMSE	Mediavilla et al., 2022; Jiang et al., 2021
Ensemble ML	XGBoost	Gradient boosting; handles non-linearity and missing data well	MAPE, RMSE	Sandhu et al., 2023; Shen et al., 2021
Deep Learning	Recurrent Neural Network (RNN)	Handles sequential data; maintains contextual dependencies	RMSE, MAE	Jiang et al., 2021; Shen et al., 2021
Deep Learning	Long Short-Term Memory (LSTM)	Captures long-term dependencies; avoids vanishing gradient issues	RMSE, MAPE, Forecast Bias	Sandhu et al., 2023; Mediavilla et al., 2022
Deep Learning	Gated Recurrent Unit (GRU)	Lightweight variant of LSTM; efficient in training	RMSE, MAE	Jiang et al., 2021; Sandhu et al., 2023
Deep Learning	Transformer (Attention Mechanisms)	Captures long-range dependencies and assign weighs input features based on relevance; enables	MAPE, RMSE, Forecast Bias	Shen et al., 2021; Mediavilla et al., 2022 Sandhu et al., 2023

DRAFT VERSION

		parallel computation for real time forecasting		

4. CASE STUDIES AND INDUSTRIAL APPLICATIONS

AI-based demand forecasting has moved beyond academic experimentation into real-world industrial applications, particularly in sectors with high SKU variability, seasonal demand, and tight logistics constraints. This section presents case studies from the retail, manufacturing, and pharmaceutical sectors and highlights key comparative findings and lessons learned.

The implementation of AI-driven demand forecasts has extended across many industries with quantifiable performance gains in speed, accuracy, and cost.

- **Retail- Large Scale:** Amazon and Walmart are the two leaders in AI introduction with predictive analysis and warehouse robotics employed to enhance replenishment precision and reduce delivery. The improved ML-driven and ERP at Walmart has led to better inventory turnover and stockout reductions in the case of promotions and seasonal rush (Apicella et al., 2021).
- **Retail- Fast Fashion:** Zara utilizes AI forecasting as it rapidly distributes production schedules to live sales data in real-time, cutting the lead time to up to 25 percent and decreasing the amount of unsold merchandise (Garcia et al., 2023).
- **Manufacturing - Automotive:** Toyota uses hybrid AI models where LSTM is adjusted with reinforcement learning to predict demand on parts and this reduces part inventory by 12% and continuity of assembly lines enhanced (Okamoto et al., 2024).
- **Energy Sector:** AI-driven forecasting of national grids, like in the UK and India, has enhanced the electricity demand prediction by up to 15% hence more generation resource can be scheduled, and the costly reserve capacity can be used less frequently (Energy AI Journal, 2024).
- **Food Supply Chains:** to predict demand of perishable goods Nestle utilizes the AI-powered systems that combine the data on the weather or planned events and social media to optimize the stock and reduce food waste by more than 18 percent (Bhandari et al., 2024).

4.1. Summary of Comparative Results

Table 2: Summary of Results

Sector	Company	AI Methods Used	Key Benefits	Challenges
Retail (Amazon)	Amazon 3P Seller	Value Stream Mapping + ML	21% faster delivery, 38% waste reduction	Multi-node supply chain complexity
Retail (Zara)	Zara	LSTM + real-time sales integration	25% shorter lead times, reduced overstock	Fast model retraining
Manufacturing (Toyota)	Toyota	LSTM + Reinforcement Learning	12% inventory reduction	Data integration across suppliers
Energy	UK National Grid	Ensemble ML + Weather Models	15% improved forecast accuracy	Weather volatility
Food (Nestlé)	Nestlé	Hybrid AI + External Data Sources	18% waste reduction	Seasonal variation impacts

Across these industries, AI-driven forecasting consistently outperformed traditional methods in accuracy, responsiveness, and operational efficiency. However, lessons highlight that success hinges on strong data infrastructure, seamless ERP integration, and context-aware model tuning.

5. CHALLENGES AND FUTURE DIRECTIONS

As AI-based demand forecasting systems become integral to supply chain operations, several challenges must be addressed to ensure robust, scalable, and interpretable solutions. These challenges span data-related issues, model transparency, and integration hurdles. The table below outlines major challenges, their implications, proposed solutions, and relevant sources.

DRAFT VERSION

Table 3: Challenges for AI-based demand forecasting

Challenge Area	Description	Key Issues	Proposed Solutions	Sources
1. Data Quality, Sparsity & Cold Start	Inaccurate or insufficient data can distort AI predictions, especially for new or infrequent items.	- Missing/inconsistent data - Sparse demand for seasonal or niche products - Lack of history for new items	- Synthetic data generation - Meta-learning - Product clustering to infer demand	Verma (2024); Rahman & Nahar (2023); Kumar et al. (2024)
2. Interpretability & Explainability	Deep learning models offer accuracy but lack transparency, leading to hesitancy among decision-makers.	- Perceived as "black boxes" - Difficulty in understanding feature contributions	- XAI tools (e.g., SHAP, LIME) - Visual dashboards showing feature influence - Simplified model outputs	Kumar et al. (2024); Verma (2024); AI in Inventory Management (2023)

5.1. Trends and Future Directions

Emerging Artificial Intelligence (AI), approaches are addressing key challenges in demand forecasting. **Hybrid models** combine statistical methods (e.g., ARIMA) with machine learning (e.g., LSTM, XGBoost) to improve both accuracy and interpretability (Rahman & Nahar, 2023). **Federated learning** supports privacy by training models locally on edge devices, making it ideal for global enterprises with strict data regulations (Verma, 2024). **Transfer learning** enables quicker and more accurate deployment in data-scarce environments by adapting pre-trained models to new domains (IJST, 2024). These innovations make Artificial Intelligence (AI), forecasting more scalable, privacy-conscious, and adaptable for modern supply chains.

Generative AI can be used to generate demand scenarios to stress-test models, and multimodal forecasting can correlate sales, images, text and sensor data to achieve higher forecast accuracy in highly complex environments. Self-governing supply chains connect AI predictions with computerized replenishment and production of ready to plan autonomy. Pseudo, ethical issues surround training data bias, privacy limitations described in regulation such as GDPR/CCPA and the opacity of black-boxes; solutions use explainable AI modes, such as SHAP and LIME. Industrial barriers involve such factors as high investment cost infrastructure SMEs, integration challenges with the legacy enterprise resource planning system and organizational lock-in to substitute human judgment with algorithms. The solution to these issues would need a concerted effort by the AI developers, supply chain managers, and policymakers to provide scalable, transparent, and equitable forecasting solutions.

6. CONCLUSION

Forecasting using AI is better in dynamic supply chains compared to doing it traditionally as LSTM and Transformers enhance accuracy in different parts as well. Data shortages, comprehensibility and assimilation difficulties are still present. The promising solutions are hybrid modeling, federated learning and transfer learning. Greater diffusion can be achieved by enhancing transparency, creating uniformity in assessment, and curbing impediments based on morality and practices.

LITERATURE

Alzubi, M. M., Nayyar, A., Kumar, A., & Mehta, P. (2021). Machine learning-based prediction and classification models for demand forecasting in pharmaceutical industry. *Applied Sciences, 11*(1213). https://doi.org/10.3390/app11031213

Amosu, O. R., Kumar, P., Ogunsuji, Y. M., Oni, S., &Faworaja, O. (2024). AI-driven demand forecasting: Enhancing inventory management and customer satisfaction. *World Journal of Advanced Research and Reviews, 23*(2), 708–719. https://doi.org/10.30574/wjarr.2024.23.2.2394

Apicella, M., Narayanamurthy, G., & Patel, P. (2021). Retail supply chain forecasting using integrated AI-ERP architecture: Case of Walmart. *Technological Forecasting and Social Change, 171*, 120994.

Bhandari, A., & Khanna, P. (2022). Integrating AI with edge computing and cloud systems. *Springer AI & IoT Systems*.

DRAFT VERSION

Cash Flow Inventory. (2023). *Managing Demand Variability: Strategies for Success*. Retrieved from internal report document.
Durand, J. P., & Baud, C. (2023). Profit-making costs and investments in the digitalization of retailing: The uneven trajectories of Amazon. *Economic and Industrial Democracy*.
Ghosh, D., & Mandal, S. (2023). Artificial intelligence in operations management and supply chain management: An exploratory case study. *Journal of Supply Chain Analytics*, 17(2), 45–59.
Harahap, A. Z. M. K., Rahim, M. K. I. A., Salleh, S. M., &Ma'arof, R. A. (2025). Enhancing the inventory management through demand forecasting. *International Journal of Research and Innovation in Social Science, 9*(1), 2737–2740. https://doi.org/10.47772/IJRISS.2025.9010221
IJST. (2024). *Emerging Trends in AI for Supply Chain Resilience*. Integrated Journal of Science and Technology.
Jahan, S., & Singh, M. (2023). Implementation of demand forecasting module in enterprise resource planning (ERP): An integrative approach. *Procedia Computer Science, 219*, 459–468. https://doi.org/10.1016/j.procs.2023.02.059
James, C. (2021). Comparative analysis of traditional vs. AI-based financial forecasting techniques. *ResearchGate*. https://www.researchgate.net/publication/386874402
Jiang, P., Qiu, Y., & Song, Z. (2021). Deep learning-based demand forecasting in multi-echelon supply chains: A comparative study. *Sensors, 21*(20), 4544. https://doi.org/10.3390/s21204544
Kumar, A., Mishra, S., & Ranjan, R. (2024). *AI in Inventory Management: Applications, Challenges, and Opportunities*. International Journal of Supply Chain Applications.
Mediavilla, M. A., Dietrich, F., & Palma, D. (2022). Review and analysis of artificial intelligence methods for demand forecasting in supply chain management. *Procedia CIRP, 107*, 1126–1131. https://doi.org/10.1016/j.procir.2022.05.119
Qin, Y., & Liu, H. (2022). Application of value stream mapping in e-commerce: A case study on an Amazon retailer. *Sustainability, 14*(713). https://doi.org/10.3390/su14020713
Rahman, S. & Nahar, N. (2023). *Federated AI and Explainable Models in Modern Supply Chains*. Journal of Artificial Intelligence and Logistics.
Ranjan, A., Sharma, P., & Dev, R. (2023). Artificial intelligence in supply chain and operations management: A multiple case study research. *Journal of Operations & Information Systems*, 12(4), 243–260.
Sandhu, G., Gupta, A., & Sehgal, R. (2023). A hybrid AI-based demand forecasting model for supply chain networks. *International Journal of Advanced Computer Science and Applications, 16*(4), 147–156.
Saha, R., Shofiullah, S., Faysal, S., & Happy, A. (2024). Systematic Literature Review On Artificial Intelligence Applications In Supply Chain Demand Forecasting. *Available at SSRN 5062817*.
Shen, X., Wu, Y., & Zhang, H. (2021). A transformer-based demand forecasting model for retail inventory systems. *Sensors, 23*(2), 1467. https://doi.org/10.3390/s23021467
Singh, A., Rasania, S. K., & Barua, K. (2022). Inventory control: Its principles and application. *Indian Journal of Community Health, 34*(1), 14–19. https://doi.org/10.47203/IJCH.2022.v34i01.004
Taleb, A., Boudjadar, J., & Mokhtari, N. (2024). Real-time hybrid forecasting model for electronics manufacturing using edge-cloud AI. *Computers in Industry, 152*, 103882.
Tulli, S. K. C. (2020). Comparative analysis of traditional and AI-based demand forecasting models. *International Journal of Emerging Trends in Science and Technology, 7*(6), 6933–6956. https://dx.doi.org/10.18535/ijetst/v7i6.02
Verma, P. (2024). Transforming Supply Chains Through AI: Demand Forecasting, Inventory Management, and Dynamic Optimization. *Integrated Journal of Science and Technology*, 1(9), 1–13.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

AUTONOMOUS AI SCIENTISTS: A NEW ERA OF SCIENTIFIC DISCOVERY

JELENA NOVAKOVIĆ[1]

[1] University of Belgrade, Faculty of Organizational Sciences, Belgrade, jekanovakovic@gmail.com

Abstract: *The development of artificial intelligence in the last decade has led to the emergence of autonomous AI systems capable of independently conducting scientific research - russefrom hypothesis generation and experimental design to data analysis and writing scientific papers. This paper focuses on the AI Scientist v2 system, one of the first examples to successfully generate and publish a fully AI-written scientific paper, accepted through peer review. The system uses a multi-agent architecture with a best-first tree search algorithm to generate hypotheses, manage experiments, and evaluate results. By integrating large language models (e.g., GPT-4o), the AI scientist can understand and produce scientific language, access databases, and compose publications. Despite its impressive capabilities, challenges such as factual accuracy, citation validity, and ethical risks remain. The conclusion is that autonomous AI scientists mark the beginning of a new scientific paradigm, but require rigorous human oversight, quality control, and appropriate ethical and regulatory frameworks.*

Keywords: *Artificial Intelligence, Autonomous Systems, Scientific Discovery, LLM, AI Scientist, Agents, Ethics.*

DRAFT VERSION

1. INTRODUCTION

In recent years, the development of artificial intelligence (AI) has dramatically transformed numerous fields of human activity, including medicine, economics, law and education (Russell & Norvig, 2021). One of the newest and most exciting areas of AI application is its role in modern scientific research. Traditionally, scientific discovery was the domain of the human mind alone – scientists formulated hypotheses, planned experiments, analyzed the results and based on that developed theories and laws of nature (Popper, 1959). However, with the advent of large language models and autonomous agent systems, a new class of so-called autonomous AI scientists has emerged - systems that can independently perform all stages of the scientific process (King et al., 2009; Yamada et al., 2025).

These systems represent a potentially revolutionary change in the way science is performed. Using advanced algorithms for language learning, search, and generation, these AI systems are capable of independently identifying relevant information from the literature, formulating hypotheses, designing experiments, and writing scientific reports (Ghareeb et al., 2025; Huang et al., 2025; Schmidgall et al., 2025; Grisoni et al., 2021; de la Torre-López et al., 2023; Yamada et al., 2025, Ofori-Boateng et al., 2024).

Understanding autonomous AI scientists requires knowledge of several key technological concepts: large-scale language models (LLM), agent systems, search algorithms, and automated scientific methods. This theoretical framework lays the foundation for the analysis of modern AI systems performing scientific tasks.

Large language models, such as GPT 4 and its successors, are trained on huge corpora of text, allowing them to generate, understand and analyze natural language at a level that is often human-like. These models form the basis of many modern scientific AI systems because they enable the semantic processing of complex scientific terms and formats. In the context of scientific research, LLMs are used to (Beltagy et al., 2019; Zhou et al., 2024; Xiong et al., 2024):

- **Analysis and summarization of scientific literature**,
- **Generation of hypotheses and research questions**,
- **Writing scientific reports, summaries and even complete publications**.

Agents in the context of AI scientists represent entities that can independently make decisions, learn from interactions with the environment, and perform complex tasks. They typically use multi-component architectures that include schedulers, evaluators, and execution modules. This modularity allows the system

to perform multiple stages of the scientific process with minimal human intervention (Yamada et al., 2025). AI Scientist v2, for example, uses a multi-agent architecture with a hierarchical division of labor between:

- **Hypothetical agent, which uses a search algorithm (best-first tree search) to generate scientific assumptions**;
- **Experimental agent, who conducts virtual experiments or calls simulations/laboratory resources**;
- **Evaluation agent, which uses evaluation metrics and LLM to decide on the validity of the results**.

At the heart of many scientific AI systems is an algorithm known as best-first search – a heuristic method that prefers the most promising hypotheses in the search of the space of possible solutions. Combined with Bayesian approaches and reinforcement learning, this algorithm allows the system to adaptively focus on experiments with the highest expected contribution (Guez et al., 2012; Ghavamzadeh et al., 2016). For example, when the system identifies 100 potential hypotheses, it uses evaluation metrics to select those with the highest probability of a beneficial outcome, thereby saving time, resources, and computing power.

The concept of automated science implies the use of AI systems that can partially or completely replace human activities in research. The first experimental steps in this direction were made by the work of (King et al., 2009), whose system was able to independently identify gene function in yeast, select experiments, and interpret results without human intervention. These concepts are further developed with the help of the integration of LLM, robotic laboratories and large science bases, thereby moving from the phase of assistance to the phase of autonomous scientific reasoning (Yamada et al., 2025).

In this paper, the focus is on the analysis of one such system - AI Scientist v2 - in order to explore the potential, advantages, but also challenges brought by the introduction of autonomous AI scientists into practice. The key research question being asked is: Do autonomous AI scientists represent a new paradigm in scientific discovery, and under what conditions can their use be sustainable and beneficial to humanity? The aim of the paper is to offer an overview of the current state of development of autonomous AI systems in science, an analysis of examples from practice, as well as a critical discussion of their limitations and possibilities.

2. CASE STUDIES: AUTONOMOUS AI SCIENTISTS IN THE REAL WORLD

The development of autonomous AI systems used for scientific research is no longer a futuristic concept, but a reality in several experimental environments. This chapter presents key case studies that demonstrate the potential, structure and effectiveness of AI scientists in practice.

2.1. The Eve System - Drug Discovery Using AI

One of the pioneering examples of automated scientific discovery is the Eve system, developed at the University of Cambridge. Eve was used to identify substances with potential antimalarial properties, and successfully detected a new chemical substance with a previously unknown mechanism of action (Williams et al., 2015).

The system integrates a robotic laboratory with algorithms for planning experiments, allowing it not only to test existing substances but also to learn which areas of chemical space are most valuable for research. This combination enables drastically faster hypothesis iteration compared to conventional methods.

2.2. AlphaFold – AI discovery in biology

AlphaFold, developed by DeepMind, uses deep neural networks to predict the 3D structures of proteins. The problem of protein folding, which was intractable for decades, was solved by this system at a level that exceeded human expectations (Jumper et al., 2021).

Although AlphaFold is not a fully autonomous AI scientist, its ability to generate accurate structures is a huge step toward automated scientific inference in the life sciences.

2.3. AI Scientist-v2 - Scientific work without human hands

A new experimental system known as AI Scientist v2, described in a paper by Yamada and co-authors (Yamada et al., 2025), represents one of the first examples where the entire scientific work—including hypothesis generation, experiment design, analysis, and writing—was performed entirely by an AI system. The paper was submitted and accepted in a peer-reviewed format, which further increases the significance of this milestone.

DRAFT VERSION

The system uses an agent architecture, LLM (eg GPT 4), database access, and "best-first" search algorithms to maximize the scientific value of each iteration.

2.4. Generative agents for scientific hypotheses

In experiments with LLM systems such as Galactica and GPT-4, it has been shown that the models can generate valid scientific hypotheses in fields such as biomedicine and chemistry, with remarkable speed and relevance (Abdel-Rehim et al., 2024; Zhou et al., 2024; Taylor et al., 2022). Although the systems currently function as "assistants", integration with automated experiments leads to increasing autonomy.

3. ADVANTAGES AND POTENTIAL OF AUTONOMOUS AI SCIENTISTS

The introduction of autonomous AI systems into scientific research brings a number of potentially revolutionary advantages that could significantly improve the efficiency, accuracy and accessibility of the scientific process. Below are key areas where autonomous AI scientists are already showing or promising major contributions:

- **Extreme speed in processing information and formulating hypotheses:** One of the main strengths of AI scientists is their ability to analyze huge amounts of scientific literature and data in a short period of time. By using large language models, AI systems can search, process and integrate data from thousands of scientific papers, which would take human researchers months or years (Kramer et al., 2023; Schneider, 2017),
- **Impartiality and elimination of cognitive biases:** AI systems, properly trained and controlled, can operate without human biases, opening the possibility for unexpected discoveries. For example, Waters and Honenberger (Waters & Honenberger,2025) point out that biases can be thought of as asymmetries, whereby AI systems can eliminate unwanted patterns, although sometimes it can be useful to target systemic biases instead. Von Felten (von Felten, 2025) adds that AI, if well calibrated, can also reduce human biases,
- **Scalability and non-stop operation:** Autonomous AI systems run 24/7 and can parallelize research by testing hundreds or thousands of hypotheses simultaneously — something that human teams can hardly keep up with. For example, Adam (Robot Scientist) could perform over 1,000 automated experiments per day, which shows the level of scalability (Keim, 2009),
- **Accessibility for less resourceful environments:** With the development of cloud infrastructure and open-source AI tools, the potential exists for AI scientists to become available to researchers from developing countries or those without access to expensive laboratory equipment. This could contribute to the democratization of science, with a more equal distribution of knowledge and discoveries. Projects like Digi-Wise and Latam-GPT are already being implemented around the world to close the gap in education and research (Business Insider, 2025),
- **Assistance in interdisciplinary research:** AI systems can naturally connect different scientific fields — e.g. relate chemical data to medical outcomes, or physical laws to biological processes. This ability to bridge disciplines makes them extremely useful in research that requires a wide range of knowledge. AI systems excel in connecting data from different domains, e.g. chemistry and medicine, physics and biology, thanks to LLM's ability to semantically chain large scientific corpora - examples include Galactica and GPT-4 that demonstrated cross-disciplinary hypothesis generation (Abdel-Rehim et al., 2024; Taylor et al., 2022).
- **Potential to Accelerate Discovery in Critical Areas:** In areas such as drug development, climate change or quantum physics, where the speed of discovery can have a direct impact on humanity, AI scientists can play a key role in accelerating the development of solutions.

4. CHALLENGES AND LIMITATIONS OF AUTONOMOUS AI SCIENTISTS

Although autonomous AI systems bring significant advantages in scientific research, their widespread application also opens up a whole series of questions and problems that must be carefully considered. Key challenges are highlighted below:

- **Limited understanding and "science simulation":** Large language models and agents have no real understanding of the world, physical laws or scientific methodology in the epistemological sense. They simulate scientific behavior based on statistical patterns in the data, but do not possess awareness, intuition, or experiential knowledge (Marcus & Davis, 2022). This means that the results can be superficially convincing, but conceptually wrong and/or logically consistent, but without empirical value.

DRAFT VERSION

- **Unreliability, hallucinations and generation of incorrect information:** One of the known problems with LLM systems is hallucination – generating content that appears accurate but is factually incorrect or fabricated. In the context of science, this can have serious consequences, especially if models are used without human verification. Banerjee and co-authors (2024) argue that hallucinations are an inherent feature of LLMs, caused by their mathematical and logical structure (Banerjee et al., 2024),
- **Lack of transparency (black box problem):** Most modern AI systems, especially deep neural networks, operate like black boxes—it is difficult to explain how the system arrived at a particular conclusion or recommendation (Doshi-Velez & Kim, 2017). In science, where explainability is key to verifiability and replicability, this is a serious obstacle,
- **Ethics, responsibility and ownership:** If an AI system generates a new hypothesis, conducts research and writes a paper, who is the author? The following questions are asked: "Who is responsible for errors or false claims?", "Can the AI be signed as an author (e.g. according to journal rules)?", "How to treat intellectual property created by an algorithm?". Some publications (e.g. Nature, Science) have already banned listing AI tools as co-authors, but debates are still ongoing (Nature, 2023),
- **Security risks and dual-use issues:** In some areas, AI could inadvertently generate dangerous or unethical research — e.g. synthesize toxic molecules, propose unethical experiments, or open the way for "bioweapons AI" (Urbina et al., 2022). This indicates the need for clear restrictions and security protocols in the development and distribution of such systems,
- **Threat to human contribution and scientific identity:** The introduction of AI as an independent actor in science also triggers broader social and philosophical dilemmas:"Will the role of human scientists be reduced to passive monitoring?", "Will AI scientists "take over" some areas of science?", "How to preserve scientific creativity, intuition and human vision, if processes are automated?".

These challenges require a wider societal debate, as well as the development of norms that balance efficiency and humanity.

5. DISCUSSION: A NEW PARADIGM OF SCIENCE OR A TOOL FOR HUMAN EXTENSION?

The emergence of autonomous AI scientists raises the question — are we in for a complete change in the scientific process, or are these systems just sophisticated tools that supplement human creativity?

The traditional scientific method, established by Bacon and Popper, presupposes human curiosity, hypothesis, experiment and conclusion (Bacon, 1620; Popper, 1959). AI systems, like AI Scientist v2, show that it is possible to perform all these steps without human intervention. Such a change could be characterized as a new epistemological paradigm, where machines not only discover data - but also generate knowledge (Abdel-Rehim et al., 2024). This suggests the possibility for science to transform from:

- **a process guided by intuition into a process guided by optimization,**
- **gradual experimental iterations into massively searched knowledge spaces,**
- **human interpretation into statistical generalization and modeling.**

However, most researchers (and this paper) advocate a model of human-AI collaboration. In that symbiosis: AI brings speed, breadth and the ability to discover unexpected patterns, while humans provide context understanding, an ethical framework and deeper interpretation.

The ideal scenario is one where autonomous AI systems function as higher-order scientific assistants, suggesting, testing, summarizing - but not making final judgments without human evaluation. Such an approach could: increase the productivity and accuracy of research, help solve problems too complex for human capacity and enable learning from AI hypotheses - even when they fail.

As AI systems become increasingly autonomous, ethical guidelines, regulatory frameworks and transparency standards will need to be developed. The key will be: a clear definition of authorship and responsibility, limiting potentially harmful uses (bioethics, AI dual-use) and educating researchers about the limits of AI interpretations. Without this, there is a danger of creating a false sense of trust in AI systems whose conclusions appear "scientific" but have not been verified.

Scientific intuition, a sense of the research question, an "experiential filter" that distinguishes the important from the irrelevant - these are all aspects that today's AI does not possess. For now, they remain indispensable elements of the scientific process. Therefore, rather than AI replacing researchers, a more likely (and desirable) path is for human scientists to become curators and guides in working with AI scientists, while AI becomes a supporting entity that expands the boundaries of human thinking.

DRAFT VERSION

6. CONCLUSION

Autonomous AI scientists represent one of the most significant advances in modern science, introducing the possibility that artificial intelligence not only assists but also independently conducts scientific research. Analysis of current examples such as AI Scientist-v2, Eve and AlphaFold shows that systems that formulate hypotheses, conduct experiments, interpret results and generate publications are already possible.

However, despite the potential, numerous challenges are evident — from epistemological limitations (lack of understanding and context), to technical problems (hallucinations and black boxes), to ethical and security issues. It is clear that autonomous AI systems cannot and should not replace human researchers, but rather function as tools that expand human capabilities in scientific work.

Based on the analysis so far, the following directions have been identified in which future research could be particularly useful:

1. **Development of interpretable AI scientists**
   Focus on transparent models whose decisions can be logically reconstructed and verified, especially in critical disciplines such as medicine and bioengineering.
2. **Introduction of ethical standards and regulatory frameworks**
   International guidelines on liability, ownership and security in the use of autonomous scientific systems need to be developed.
3. **Management of hallucinations and accuracy of generated results**
   Develop methodologies for the detection and prevention of inaccurate information that may be produced by AI, including automatic verification of references and inferences.
4. **Collaboration of AI with human teams in real research environments**
   Empirically examine how AI scientists integrate into multidisciplinary teams, which models of collaboration bring the most benefits, and how AI affects scientific thought.
5. **Philosophical and epistemological analysis of AI generated knowledge**
   To set a theoretical framework for understanding if and when knowledge generated by AI can have the same status as knowledge derived from human understanding and experience.

Autonomous AI scientists are not only a technical tool, but also a challenge to the very definition of scientific knowledge, method and the role of humans in the process of discovery. Their responsible integration into research practice requires not only advances in technology, but also mature interdisciplinary collaboration between engineers, scientists, ethicists, and philosophers.

If they develop in that spirit — as an extension of human curiosity, not a replacement for it — AI scientists can become key players in solving humanity's most important questions in the 21st century.

LITERATURE

Abdel-Rehim, A., Zenil, H., Orhobor, O., Fisher, M., Collins, R. J., Bourne, E., ... & King, R. (2025). Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment. *Journal of the Royal Society Interface*, *22*(227), 20240674.

Bacon, F. (2000). *The new organon (L. Jardine & M. Silverthorne, Trans.)*. Cambridge University Press. (Original work published 1620)

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv*. https://arxiv.org/abs/1903.10676

Business Insider. (2025, jul 10). AI models aren't made equal. Some nonprofits are creating their own tools instead. *Business Insider*. https://www.businessinsider.com/nonprofits-ai-tools-edtech-global-inequities-2025-7

de la Torre-López, J., Ramírez, A., & Romero, J. R. (2023). Artificial intelligence to automate the systematic review of scientific literature. *Computing*, *105*(10), 2171-2194.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Ghareeb, A. E., Chang, B., Mitchener, L., Yiu, A., Szostkiewicz, C. J., Laurent, J. M., Razzak, M. T., White, A. D., Hinks, M. M., Rodriques, S. G., et al. (2025). Robin: A multi-agent system for automating scientific discovery. *arXiv*. https://arxiv.org/abs/2505.13400

DRAFT VERSION

Ghavamzadeh, M., Mannor, S., Pineau, J., & Tamar, A. (2016). Bayesian Reinforcement Learning: A Survey. *arXiv*. https://arxiv.org/abs/1609.04436

Grisoni, Francesca, Berend JH Huisman, Alexander L. Button, Michael Moret, Kenneth Atz, Daniel Merk, and Gisbert Schneider. "Combining generative artificial intelligence and on-chip synthesis for de novo drug design." *Science Advances* 7, no. 24 (2021): eabg3338.

Guez, A., Silver, D., & Dayan, P. (2012). Efficient Bayes-Adaptive Reinforcement Learning using Sample-Based Search. *arXiv*. https://arxiv.org/abs/1205.3109arXiv

Huang, K., Jin, Y., Li, R., Li, M. Y., Candès, E., & Leskovec, J. (2025). Automated hypothesis validation with agentic sequential falsifications. *arXiv*. https://arxiv.org/abs/2502.09858

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *nature*, *596*(7873), 583-589.

Keim, B. (2009, april 2). Robot makes scientific discovery all by itself. *Wired*. https://www.wired.com/2009/04/robotscientist/

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., ... & Clare, A. (2009). The automation of science. *Science*, *324*(5923), 85-89.

Kramer, S., Cerrato, M., Džeroski, S., & King, R. (2023). Automated scientific discovery: from equation discovery to autonomous discovery systems. *arXiv preprint arXiv:2305.02251*.

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Pantheon Books.

Nature. (2023). Journal policies and artificial intelligence. *Scholarly Communications*. https://sites.bc.edu/schol-comm/2023/10/16/journal-policies-and-artificial-intelligence/

Ofori-Boateng, R., Aceves-Martins, M., Wiratunga, N., & Moreno-Garcia, C. F. (2024). Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review. *Artificial intelligence review*, *57*(8), 200.

Peter, N., & Intelligence, R. S. A. (2021). *A Modern Approach*. Pearson Education, USA.

Popper, K. (2005). *The logic of scientific discovery*. Routledge. (Original work published 1959)

Schneider, G. (2018). Automating drug discovery. *Nature reviews drug discovery*, *17*(2), 97-113.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., ... & Stojnic, R. (2022). Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3), 165–167. https://doi.org/10.1038/s42256-022-00465-9

von Felten, N. (2025). Beyond Isolation: Towards an Interactionist Perspective on Human Cognitive Bias and AI Bias. *arXiv preprint arXiv:2504.18759*.

Williams, K., Bilsland, E., Aubrey, W., Young, M., & King, R. D. (2015). Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *Journal of the Royal Society Interface*, 12(104), 20141289. https://doi.org/10.1098/rsif.2014.1289

Waters, G., & Honenberger, P. (2025). AI Biases as Asymmetries: A Review to Guide Practice. *arXiv preprint arXiv:2503.07326*.

Xiong, G., Xie, E., Shariatmadari, A. H., Guo, S., Bekiranov, S., & Zhang, A. (2024). Improving scientific hypothesis generation with knowledge grounded large language models. *arXiv*. https://arxiv.org/abs/2411.02382

Yamada, Y., Lange, R. T., Lu, C., Hu, S., Lu, C., Foerster, J., Clune, J., & Ha, D. (2025). The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv*. https://arxiv.org/abs/2504.08066

Zhou, Y., Liu, H., Srivastava, T., Mei, H., & Tan, C. (2024). Hypothesis generation with large language models. *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, 117–139. https://aclanthology.org/2024.nlp4science-1.10/

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

DL-SPP: A Deep Learning System for Spare Parts Prediction in Field Service Operations

ALESSIO SCHIAVO[1,2], ALESSANDRO RENDA[3], PIETRO DUCANGE[4], FRANCESCO MARCELLONI[5]

[1] University of Pisa, Department of Information Engineering, Pisa, alessio.schiavo@phd.unipi.it, 0009-0005-2147-2853
[2] LogObject AG, Opfikon, Switzerland, alessio.schiavo@logobject.ch
[3] University of Trieste, Department of Engineering and Architecture, Trieste, 0000-0002-0482-5048
[4] University of Pisa, Department of Information Engineering, Pisa, pietro.ducange@unipi.it, 0000-0003-4510-1350
[5] University of Pisa, Department of Information Engineering, Pisa, francesco.marcelloni@unipi.it, 0000-0002-5895-876X

Abstract: *Accurate spare parts prediction is crucial in field service operations for the maintenance and repair of household appliances, where missing components frequently lead to repeated technician visits, increased costs, and diminished service quality. In this context, we present the DL-SPP (Deep Learning Spare Part Prediction) system that predicts the set of spare parts likely needed before the technician's intervention, using both structured appliance metadata and semantic representations of fault descriptions. The prediction task is framed as a multi-label classification problem, where the model learns to associate appliance types and fault reports with the components used in past repairs through a transformer-based architecture. Evaluated on a real-world dataset of repair tasks, DL-SPP achieves up to a 56% reduction in repeated field service technician visits, outperforming a state-of-the-art baseline that achieves 50%. These results highlight the DL-SPP system as a scalable and effective solution for enhancing repair planning and inventory optimization in operational field service contexts.*

Keywords: *Field Service Optimization, Spare Parts Prediction, Natural Language Processing, Repair Kit Problem, Deep Learning*

DRAFT VERSION

1. INTRODUCTION

Optimizing field service operations is critical for companies providing repair and maintenance of household appliances. When an appliance malfunctions, completing the repair during a single technician visit is essential to minimize operational costs and ensure customer satisfaction. Additional visits lead to increased labor, inventory, and travel expenses, along with diminished service quality. Consequently, accurately predicting the required spare parts prior to the intervention becomes a critical factor in improving operational efficiency.

Traditional approaches to spare parts planning have mainly addressed inventory management and demand forecasting, typically framing the problem as a combinatorial optimization task. In the context of the *Repair Kit Problem* (Smith et al., 1980), several studies have focused on identifying the optimal assortment of parts to be stocked in service vans, balancing holding costs with the penalties associated with incomplete repairs (Bijvank and Vis, 2010; Prak et al., 2017; Teunter, 2006). More recent contributions have explored heuristic and simulation-based methods for handling multiple-job routes, uncertain demands, or replenishment constraints (Pham and Gudrun P. Kiesmüller, 2022; Rippe and Gudrun P. Kiesmüller, 2023a; Rippe and Gudrun P Kiesmüller, 2023b). In these works, the probability that a given spare part sp is required for a specific repair task t_j is denoted as $p_{sp}^{t_j}$ and treated as a fixed input parameter. This probability is typically estimated using simple frequency counts from historical data. However, this assumption limits the flexibility and adaptability of such models in real-world scenarios, where the need for specific parts often depends on contextual and unstructured information such as fault descriptions.

Some studies (Saccani and Johansson, 2017) have proposed budget-aware repair kit strategies, incorporating sensor-based fault signals as Advance Demand Information (ADI), while others have integrated scheduling components to jointly optimize service routing and spare part planning (Pham and Gudrun P. Kiesmüller, 2022; Rippe and Gudrun P. Kiesmüller, 2023a). Despite these contributions, existing approaches typically rely on structured or codified input data, overlooking the unstructured textual fault descriptions that are often not only available prior to a field technician's intervention, but also represent the sole information describing the issue.

The increasing availability of digital maintenance records presents a valuable opportunity to move beyond rule-based heuristics toward predictive, data-driven methodologies. In this context, our previous work introduced CBR-SPP (Schiavo et al., 2024), a Case-Based Reasoning system that leverages historical repair data to

recommend spare parts based on textual similarity and appliance metadata. Although effective, CBR-SPP is limited by its reliance on case retrieval and exhibits reduced generalization capabilities when faced with previously unseen appliance-failure scenarios.

In this paper, we propose the DL-SPP (Deep Learning Spare Part Prediction) system, which predicts the parts needed for a repair task using both structured appliance information and semantic representations of free-text fault descriptions. The task is formulated as a multi-label classification problem, and predictions are generated through a transformer-based deep learning architecture. Unlike prior approaches that assume fixed demand probabilities for each part, DL-SPP learns these probabilities directly from data by modeling the relationships between appliance types, fault symptoms, and historically used spare parts.

Our experimental evaluation on a real-world dataset of household appliance repairs confirms the advantages of this approach: DL-SPP significantly outperforms CBR-SPP across all evaluation metrics, achieving a greater reduction in multi-visit repair tasks. These results underscore the value of incorporating unstructured information into spare parts planning, and demonstrate the practicality and scalability of DL-SPP as a real-time decision support tool in field service logistics.

The remainder of the paper is organized as follows. Section 2 formalizes the problem. In Section 3 we describe the proposed DL-SPP system. Section 4 presents the experimental setup and results. In Section 5 we draw some conclusions.

2. PROBLEM STATEMENT

The Spare Part Prediction (SPP) problem addressed in this work focuses on predicting the set of spare parts required by Field Service Technicians (FSTs) to successfully complete household appliance repair interventions. The objective is to anticipate, prior to the FST's visit, the parts most likely needed, thereby preventing incomplete repairs and reducing the number of repeat visits. We assume the availability of a historical archive where each record contains the following standard information typically gathered during fault management.

- *Appliance Type* (AT): a categorical alphanumeric code that identifies the type of appliance involved in the intervention. Each code groups together appliances with similar structure and functionality, typically belonging to the same product line or technical platform. For instance, multiple washing machine models that share internal components may be associated with the same AT code.
- *Fault Description* (FD): an unstructured textual field containing a brief description of the symptoms observed, as communicated by the appliance owner during the initial contact with the service center (e.g., "The door does not close properly" or "Water leaks during the rinse cycle").
- *Spare Parts Used* (SP): the set of alphanumeric identifiers of spare parts actually installed by the FST to complete the repair.

This setting is both practical and realistic, as companies providing field repair services commonly acquire such information as part of their routine operations. The recorded data constitute a repository of repair interventions, which can be used to train data-driven models aimed at supporting FSTs by anticipating the spare parts likely required for similar future cases.

To better illustrate the structure of the available data, Table 1: reports a real-world example of a repair intervention record extracted from the historical archive.

Table 1: Example of a repair intervention record.

Appliance Type (AT)	Fault Description (FD)	Spare Parts Used (SP)
76852	The door does not close properly and the program cannot start.	{HK7552, FN3786}

Formally, each input sample is represented as $x = (AT, FD)$, where AT is a categorical variable and FD is a free-text input. The task consists in predicting a set of spare parts $Y_x = \{sp_1, sp_2, \ldots, sp_k\}$, where each $sp_i \in \mathcal{SP}$, the full catalog of n available spare parts.

3. PROPOSED SOLUTION

To effectively address the SPP problem described in the previous section, we propose a deep learning-based SPP system, named DL-SPP. We formulate the task as a multi-label, multi-class classification problem, where each input may be associated with multiple relevant labels (i.e., spare parts). The objective is to accurately predict the set of spare parts required for each intervention, with the aim of reducing multi-visit repair tasks and

DRAFT VERSION

enhancing overall service efficiency. The system predicts the set of spare parts required for a specific appliance repair task by directly learning relationships between textual fault descriptions, structured appliance metadata, and spare parts usage from historical data.

Our proposed DL-SPP architecture is illustrated schematically in Fig. 1. The model is based on a Multi-Layer Perceptron (MLP) neural network, specifically tailored for multi-label classification, where each spare part is treated as an independent output label. DL-SPP leverages advanced semantic representations derived from Sentence-BERT (SBERT) embeddings for textual fault descriptions. To ensure multilingual compatibility and robustness in handling textual inputs, we adopt the pre-trained *distiluse-base-multilingual-cased-v1* SBERT model (Reimers and Gurevych, 2019).

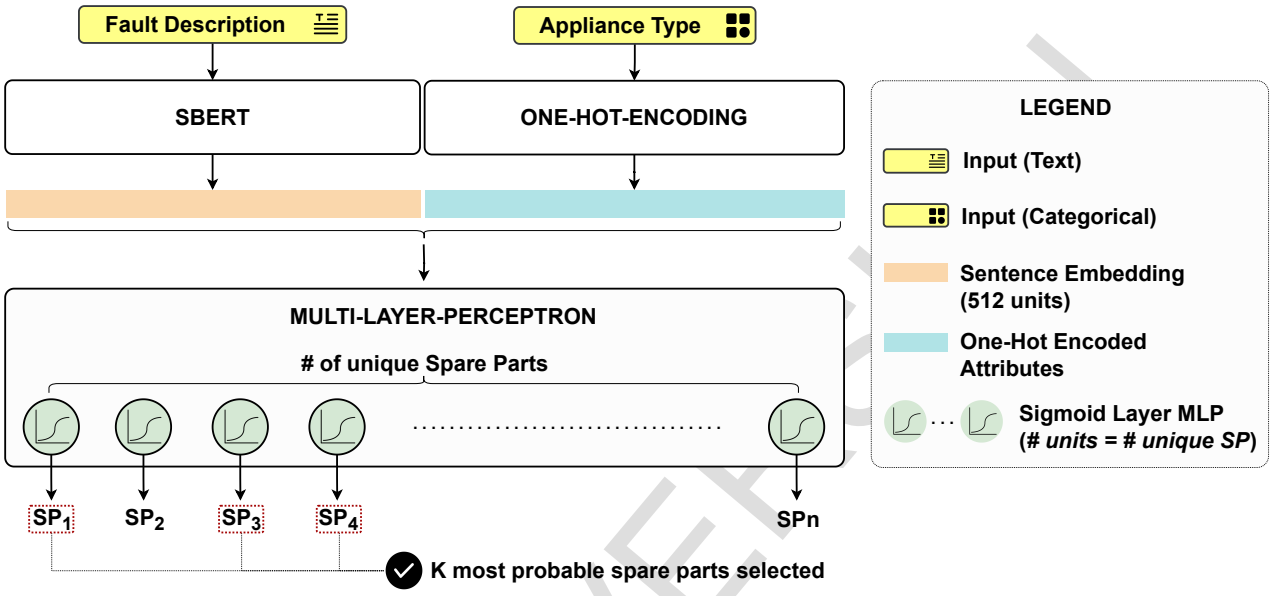


Figure 1: SPP System Architecture Diagram

Given a repair intervention request characterized by its *Appliance Type* and *Fault Description*, we first encode the textual description into a semantic embedding using SBERT. This embedding is a dense numerical representation capturing the semantic meaning of the fault description in a fixed-size 512-dimensional vector. Additionally, the categorical attribute representing the *Appliance Type* is encoded using one-hot encoding, providing structured information about the appliance involved.

The two resulting feature vectors, namely the SBERT embeddings for the *Fault Description* and the one-hot encoded *Appliance Type*, are concatenated into a unified input vector that feeds into the MLP classifier. The MLP consists of multiple hidden layers employing Rectified Linear Unit (ReLU) activation functions to capture nonlinear feature interactions effectively. The output layer uses sigmoid activation functions to predict independent probabilities for each spare part, indicating the likelihood of being required for the specific repair intervention. Formally, the prediction step can be described as follows. Given the input $x = (FD, AT)$, the MLP produces a vector of probabilities $[p_1, \ldots, p_n]$, where p_i represents the predicted likelihood that the spare part sp_i will be required for the repair. The final recommendation list is generated by selecting the top-K spare parts with the highest predicted probabilities, where K is a user-defined parameter balancing prediction coverage and inventory efficiency.

This architectural choice leverages semantic embeddings and structured appliance metadata. It enables the DL-SPP system to generalize effectively by learning intricate relationships between appliance types, fault descriptions, and historical spare parts usage patterns, without relying on explicit case retrieval. By avoiding case-based retrieval, the DL-SPP system ensures scalability and high computational efficiency, making it particularly suitable for dynamic and large-scale operational environments.

In the next section, we describe the experimental setup employed to validate the effectiveness and practical applicability of the DL-SPP system, in comparison with the CBR-SPP baseline.

4. EXPERIMENTAL RESULTS

This section presents the experimental evaluation of the proposed DL-SPP system. We begin by describing the real-world dataset employed in our study, including the preprocessing steps and its key characteristics. Next,

DRAFT VERSION

we outline the experimental setup and the evaluation metrics used to assess model performance. Finally, we compare DL-SPP with the CBR-SPP baseline through both quantitative analysis and qualitative insights.

4.1. EXPERIMENTAL SETUP

The proposed DL-SPP system was validated using the same real-world dataset introduced in our previous work on the CBR-SPP system (Schiavo et al., 2024). The dataset was provided by a leading household appliance manufacturer and includes approximately 131,000 repair intervention records collected between 2020 and 2023. Each record contains the three key elements described in Section 2: the *Appliance Type*, a categorical code identifying the device class; the *Fault Description*, an unstructured textual description of the failure as reported by the appliance owner; and the set of *Spare Parts* used by the FST to complete the repair.

To ensure data quality, a preprocessing step was applied. Records containing missing or malformed values in either the *Appliance Type* (AT) or *Spare Part* (SP) fields were discarded. *Fault Descriptions* (FD) were retained in their original languages—Italian, German, or French—as the selected SBERT model (`distiluse-base-multilingual-cased-v1`) supports multilingual input. The categorical AT values were encoded using one-hot encoding.

The dataset was split chronologically: repair tasks from 2020 to 2022 formed the training set, while tasks from 2023 constituted the test set. To focus on scenarios where predictive support is most impactful, we excluded single-visit repairs from the test set. These typically involve commonly stocked parts and are less sensitive to prediction quality. Instead, we retained only records of repair tasks that originally required multiple visits.

Table 2 summarizes the characteristics of the dataset, grouping repair tasks by the number of FST visits required to resolve the issue. The bottom row aggregates statistics for tasks requiring two or more visits, which represent the effective test set used in our evaluation.

Table 2: Summary of the dataset used for DL-SPP evaluation. Repair tasks are grouped by the number of FST visits. 4+ and 2+ denote tasks requiring four or more, and two or more visits, respectively. Mean SP and Max SP refer to the average and maximum number of spare parts used.

# Visits	Training Set (2020–2022)			Test Set (2023)		
	# Records	Mean SP	Max SP	# Records	Mean SP	Max SP
1	72,536 (69%)	1.53	15	17,317 (67%)	1.54	14
2	25,759 (24%)	2.77	15	6,672 (26%)	2.81	15
3	5,813 (6%)	3.32	20	1,573 (6%)	3.37	16
4+	1,099 (1%)	4.88	17	309 (1%)	4.43	14
2+	**32,671 (31%)**	**2.95**	**20**	**8,554 (33%)**	**2.97**	**16**

DL-SPP was trained using the SBERT-generated embeddings for FD and the one-hot encoded representations of AT. The prediction task was formulated as a multi-label classification problem, where the model predicts a set of relevant spare parts for each repair request.

To determine the optimal architecture, we performed hyperparameter tuning using the *Keras Hyperband Tuner* algorithm[1] on a validation subset (20% of the training set). Once the optimal architecture was identified, the model was retrained using the entire training set, including the validation data, to fully leverage the available data. The final configuration consists of two hidden layers with 4,096 and 2,048 units, respectively, each using *ReLU* activation functions. The output layer uses Sigmoid activation to produce independent probability scores for each spare part class. The network was trained to minimize binary cross-entropy loss.

4.2. Evaluation Metrics

We evaluated the DL-SPP system using the same metrics adopted in (Schiavo et al., 2024), which allow comparison with the CBR-SPP baseline and provide insight into both predictive accuracy and operational impact. For varying values of K, indicating the number of recommended spare parts, we report:

- *Closed Task Percentage (CTP_K)*: percentage of tasks successfully completed in a single visit using the top-K predicted parts.

[1] https://keras.io/keras_tuner/api/tuners/hyperband/

DRAFT VERSION

- *Saved Visits Percentage (SVP_K)*: percentage of additional FST visits that could have been avoided.
- *Average Recall of Not Closed Tasks ($\overline{recallNC}_K$)*: average proportion of spare parts correctly predicted by the system in repairing tasks that have not been closed.
- *Average Recall ($\overline{recall}_K$)*: average proportion over all tasks.
- *Average Precision ($\overline{precision}_K$)*: average proportion of correctly predicted parts among the top-K.

We refer the reader to (Schiavo et al., 2024) for further details.

4.3. Results

We compared the proposed DL-SPP system against the CBR-SPP baseline on the dataset described in the previous section. Table 3 reports the average results across 10 independent training and testing runs. Standard deviations were consistently low (below 0.5) for all metrics and are thus omitted for brevity.

Table 3: Values of the evaluation metrics with respect to increasing values of K recommended spare parts, obtained by CBR-SPP and DL-SPP systems. Best model for each K is highlighted in bold.

K	Model	CTP_K	SVP_K	$\overline{recallNC}_K$	$\overline{recall}_K$	$\overline{precision}_K$
3	CBR-SPP	15.24	14.32	14.02	30.69	26.41
	DL-SPP	**22.25**	**17.52**	**21.88**	**38.65**	**30.26**
6	CBR-SPP	25.01	23.71	21.62	45.14	20.01
	DL-SPP	**32.97**	**27.56**	**32.57**	**53.09**	**22.27**
9	CBR-SPP	32.57	31.36	26.68	54.35	16.43
	DL-SPP	**40.38**	**36.04**	**37.64**	**61.84**	**17.97**
12	CBR-SPP	38.02	36.75	30.61	60.58	13.96
	DL-SPP	**45.07**	**41.63**	**42.11**	**68.91**	**15.95**
15	CBR-SPP	43.24	41.81	33.52	65.27	12.19
	DL-SPP	**47.71**	**46.82**	**45.05**	**72.37**	**13.96**
18	CBR-SPP	47.46	45.98	35.99	68.86	10.84
	DL-SPP	**53.14**	**51.23**	**47.56**	**76.21**	**12.49**
21	CBR-SPP	51.48	50.12	38.28	71.99	9.78
	DL-SPP	**56.08**	**55.72**	**49.33**	**79.32**	**11.34**

As expected, both systems show performance improvements as K increases, confirming their ability to suggest more relevant spare parts when a larger number of predictions is allowed. However, DL-SPP consistently outperforms CBR-SPP across all evaluation metrics and for every value of K. Notably, DL-SPP achieves a significant improvement in CTP_K and SVP_K, indicating a greater capability in completing repair tasks with fewer technician visits. These gains are particularly relevant in operational contexts where minimizing repeat visits leads to lower costs and higher customer satisfaction.

For lower values of K, the improvements of DL-SPP over the baseline are particularly significant. At $K = 3$, for example, DL-SPP achieves a CTP_3 of 22.25%, compared to 15.24% for CBR-SPP, and reduces technician visits by 17.52%, compared to 14.32%. These results highlight DL-SPP's capacity to produce shorter and more accurate spare part recommendations, especially beneficial for optimizing inventory and service efficiency.

The advantage of DL-SPP remains evident as K increases, although the relative gains diminish. For instance, the difference in SVP_K between $K = 18$ and $K = 21$ is about 4 percentage points, indicating that most relevant parts are captured at lower K values. This suggests DL-SPP can provide effective recommendations with fewer predicted components, benefiting inventory management and operational efficiency.

The DL-SPP system consistently achieves higher values of $\overline{recall}_K$ and $\overline{precision}_K$ across all K, confirming improved completeness and accuracy. Precision is especially critical for minimizing overprovisioning, as it reflects the proportion of predicted parts actually required. Although it naturally declines as K increases due to less relevant items, DL-SPP maintains consistently higher precision than the baseline, underscoring its effectiveness in generating more targeted and efficient recommendations.

DRAFT VERSION

Overall, the results demonstrate that the DL-SPP system provides a robust and scalable solution for spare parts prediction. Its ability to generalize from historical data allows it to effectively learn correlations between fault descriptions and the components required for repair, leading to fewer visits and better resource allocation.

5. CONCLUSIONS

In this paper, we proposed DL-SPP, a novel deep learning-based approach for predicting the spare parts required in household appliance field service operations. By integrating semantic embeddings of fault descriptions with structured appliance type information, DL-SPP addresses key limitations of prior retrieval-based methods, such as the Case-Based Reasoning Spare Parts Prediction (CBR-SPP), particularly in terms of generalization and predictive accuracy. Experimental evaluation conducted on a real-world dataset from a leading appliance manufacturer demonstrated that DL-SPP significantly outperforms the CBR-SPP baseline. Our results indicate substantial improvements in the saved visits percentage, effectively reducing operational costs and enhancing customer satisfaction through more accurate spare part provisioning and minimized technician visits.

The proposed solution exhibits strong scalability and generalization capabilities, essential features for dynamic, large-scale repair environments where spare parts inventories and appliance portfolios evolve continuously. Additionally, by directly modeling relationships between fault descriptions and spare parts usage, DL-SPP facilitates proactive inventory management, optimized technician scheduling, and overall improvement in logistics efficiency, aligning closely with key goals of Operations Research.

Acknowledgments

This work has been partly funded by the PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - “FAIR - Future Artificial Intelligence Research" - Spoke 1 “Human-centered AI" under the NextGeneration EU programme, and by the Italian Ministry of University and Research (MUR) in the framework of the FoReLab and CrossLab projects (Departments of Excellence).

LITERATURE

Bijvank, M., & Vis, I. F. (2010). Lost-sales inventory theory: A review. *European Journal of Operational Research*, *204*(2), 195–205. https://doi.org/https://doi.org/10.1016/j.ejor.2011.02.004

Pham, D. T., & Kiesmüller, G. P. [Gudrun P.]. (2022). Multiperiod integrated spare parts and tour planning for on-site maintenance activities with stochastic repair requests. *Computers & Operations Research*, *148*, 105967. https://doi.org/https://doi.org/10.1016/j.cor.2022.105967

Prak, D., Saccani, N., Syntetos, A., Teunter, R., & Visintin, F. (2017). The Repair Kit Problem with positive replenishment lead times and fixed ordering costs. *European Journal of Operational Research*, *261*(3), 893–902. https://doi.org/https://doi.org/10.1016/j.ejor.2017.02.019

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing*, 3982–3992. https://doi.org/https://doi.org/10.48550/arXiv.1908.10084

Rippe, C., & Kiesmüller, G. P. [Gudrun P.]. (2023a). The repair kit problem with imperfect advance demand information. *European Journal of Operational Research*, *304*(2), 558–576. https://doi.org/https://doi.org/10.1016/j.ejor.2022.04.019

Rippe, C., & Kiesmüller, G. P. [Gudrun P]. (2023b). The added value of advance demand information for the planning of a repair kit. *Central European Journal of Operations Research*, *31*(1), 311–335. https://doi.org/https://doi.org/10.1007/s10100-022-00809-1

Saccani, N., & Johansson, P. (2017). Spare parts classification and demand forecasting for stock control: Investigating the gap between research and practice. *Omega*, *47*, 185–199. https://doi.org/https://doi.org/10.1016/j.omega.2011.06.008

Schiavo, A., Marcelloni, F., Ducange, P., Renda, A., Ruffini, F., & Dani, R. (2024). On Predicting Spare Parts for Field Services by Leveraging Fault Description and Historical Repairing Data. *IEEE Access*, *12*, 162756–162768. https://doi.org/10.1109/ACCESS.2024.3487493

Smith, S. A., Chambers, J. C., & Shlifer, E. (1980). Note—Optimal Inventories Based on Job Completion Rate for Repairs Requiring Multiple Items. *Management Science*, *26*(8), 849–854. https://doi.org/10.1287/mnsc.26.8.849

Teunter, R. H. (2006). The multiple-job repair kit problem. *European Journal of Operational Research*, *175*(2), 1103–1116. https://doi.org/https://doi.org/10.1016/j.ejor.2005.05.023

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Review article

CHALLENGES OF MEASURING CONVERSION RATE IN CONDITIONS OF AI GENERATED PRODUCT SEARCH

SONJA VUČENOVIĆ[1], GORAN VUKMIROVIĆ[2], NIKOLA MACURA[3]

[1] University of Novi Sad, Faculty of Economics in Subotica, Subotica, sonja.vucenovic@ef.uns.ac.rs, 0000-0003-2375-5759
[2] University of Novi Sad, Faculty of Economics in Subotica, Subotica, goran.vucenovic@ef.uns.ac.rs, 0000-0003-1270-2344
[3] University of Novi Sad, Faculty of Economics in Subotica, Subotica, nikola.macura@ef.uns.ac.rs, 0009-0006-9917-6998

Abstract: *Electronic commerce has changed drastically since the pandemic, and now by the use of artificial intelligence (AI) is changing significantly. There are many ways in which artificial intelligence is changing e-commerce, and in this paper, the focus is on product search. More efficient product search, thanks to artificial intelligence, enables an increase in conversion rates, as one of the indicators that e-retailers track the most. Although conversion rates are relatively low, they tend to grow, and with artificial intelligence, e-commerce finally brings the right product, to the right customer, at the right time and under the required conditions, bringing an efficiency component to each of these parameters.*

Keywords: *e-commerce, conversion rate, artificial intelligence, Amazon, product.*

1. INTRODUCTION

Most customers today expect sellers and manufacturers to understand their unique needs, and artificial intelligence (AI) is becoming the basis for differentiating from the competition. Oosthuizen, et al. (2020) discus in their research how traditional business models, especially in retailing, are facing disruption by new entrants (driven by technology) who can deliver greater value to customers more efficiently. Artificial intelligence facilitates the product search, finds optimal solutions in relation to customer needs. New opportunities and challenges are opening up in electronic commerce. As a result, the following critical questions arise:

- Do customers use AI-generated search to discover e-commerce products?
- Are users finding the right products thanks to AI search?
- If users reach an e-retailer thanks to AI-generated search results, will they convert, ie. make a purchase?, and finally
- How to get an electronic store listed in AI generated search results?

In the digital era, AI-personalized recommendation has been deeply applied in electronic commerce, it has a remarkable promoting effect on consumers' purchasing behavior and effectively increases the sales volume of e-commerce platforms (Ying, Qiu & Wang, 2025). Saleh & Zeebaree (2025) concluded that driven by AI big data analytics, optimize customer journeys and drive personalized marketing campaigns, addressing data privacy and scalability issues in the process to bring about substantive business growth. Furthermore, Yin, Qiu & Wang (2025) research emphasizes that if information quality improves, the positive impact of technology acceptance on clicking intention will be enhanced showed; it shows consumers' experiences of AI-personalized recommendations and clarifies how these experiences affect the clicking intention and offers valuable insights for e-commerce platforms to continuously optimize personalized recommendation algorithms and boost the click conversion rate of online shopping. In Kuman and Trakru (2020) research emphasize that the application of AI and machine learning enhances the sale of Amazon drastically.

The premise of the work is that if potential customers receive more efficient and accurate search results thanks to artificial intelligence (AI), then it is more likely that they will buy the product they found thanks to the search with AI, that is, that the AI-generated search increases the conversion rate, because those who really want to buy products come to the electronic store, the precision of the search in accordance with the customer's requirements is higher. Therefore, the display and information about products on the electronic store should be adapted to AI searches.

DRAFT VERSION

2. MEASUREMENT AND MEANING OF CONVERSION RATE IN ELECTRONIC COMMERCE

In e-commerce, the average conversion rate is a percentage that measures how effective an electronic store is at persuading users to take a desired action. It can be downloading content, signing up for an e-mail list or most importantly buying a product. The conversion rate is measured by simply dividing the number of purchases by the visits made during a certain period, and then multiplying by 100:

Conversion rate = (Number of purchases / Total number of visitors) X 100 (1)

For example, if an electronic store has 5000 visitors per month and has 50 sales, the conversion rate would be:

(50 / 5000 visitors) X 100 = 1% conversion rate

So, if the electronic store has 1,000 visits and 100 products purchased in a certain period, the conversion rate is 100/1,000 = 0.10, or 10%. Which of these two percentages will be satisfactory depends on the set goals of the electronic retailer, that is, the activities undertaken in the previous period.

More customers are interested in online shopping after COVID, so the average conversion rate after the pandemic has increased, habits have changed, and many customers are realizing the real advantage of online shopping. Historical evolution of e-commerce conversion rates show changes in average conversion rate by year (SpeedCommerce, 2025):

- 2000-2010: Static interfaces and limited trust led to conversion rate of 1.5%-2.5%.
- 2010-2020: Improved UX/UI and payment options boosted conversion rate to 2.5%-4.0%.
- 2020-2024: Pandemic-driven spikes stabilized at 3.0%-4.2% post-COVID.
- 2025: Ongoing growth in personalization and mobile optimization maintains conversion rate at 2%-4%.

The height of the conversion rate largely depends on the type of e-commerce, products, visitors and many other factors. Based on the above percentages, the average conversion rate of e-commerce can be calculated, which is between 2-4%. This means that 2 to 4 out of every 100 visitors make a purchase on electronic stores. That may sound like a small percentage, but conversion rates are actually on an upward trend. The average conversion rates for the six most common product categories in e-commerce are (Adobe for Business, 2025):

- Fashion and apparel: 2.7%
- Health and beauty: 3.3%
- Entertainment: 2.5%
- Household goods: 2.1%
- Electronics: 1.9%
- Food and beverage: 4.6%

What do these percentages show? Why does perhaps the most searched product category on the web, fashion and apparel, have a low conversion rate? Precisely because of the fact that potential customers do a lot of searching, find the most suitable offers and often end up making a purchase in the offline sales store. On the other hand, a product category that is not popular for shopping on the web, such as food and beverages, has a high conversion rate, due to the fact that those who actually intend to buy that way, complete the purchase process after searching. From the above, we can conclude that the potential more effective application of AI in web search can be precisely in these product categories that have lower conversion rates.

3. CHALLENGES AND ADVANTAGES OF AI GENERATED PRODUCT SEARCH

Internet searches are undergoing a fundamental change. Search Engine Optimization (SEO) has been a key component of digital marketing in e-commerce. However, the optimization of online stores according to search engine SEO requirements and the general behavior of customers when searching for information online are likely to experience major changes in the future. Understanding the factors that drive AI-powered search behavior can inform the design of more effective AI systems and marketing strategies (Pham, et al., 2024).

The emergence of generative AI (GenAI) search tools such as Chatbot, ChatGPT and DeepSeek are emerging as new channels for finding products and services online. Chatbot will welcome a user to

DRAFT VERSION

conversation, guide to customer to make purchase which will reduce customer's struggle, ask customers all the relevant questions to find the perfect fit, style, and color for them (Shafi, et al, 2020). For online customers, this functionality has enabled searches with more complex and specific requirements compared to traditional searches based on keywords and paid ads.

For the effective use of AI-generated product searches, which should result in an increase in the conversion rate, it is first of all necessary to identify and monitor AI-based searches on the electronic store: the number of AI search over time (weekly/monthly), which page of the electronic store they come to (home page, product listing page, product detail page, etc.), how many visitors arrived on the e-store for the first time, conversion flow, i.e. what percentage of search AI ends up buying, what products, etc.

AI-powered search still accounts for fewer conversions, but it is growing rapidly and will tend to grow as its benefits, efficiency and effect on conversion rates are realized. As AI search tools become more ubiquitous, electronic stores need to proactively ensure their content is visible. How much this technology is developed is shown from the speed of technology adoption. Namely, it took only 5 days for ChatGPT to reach a million users (Threads 1 hour, Instagram 2.5 months, Spotify 5 months, Facebook 10 months, Twitter 2 years, Netflix 3.5 years) (Buchholz, 2023).

With such strong growth of AI-based solutions, it is not difficult to predict that it could soon represent a significant percentage of the "organic search" of traditional search engines, such as Google. The growth is supported by the fact that it offers truly new and revolutionary ways to interact with the vast amount of information available on the web. ChatGPT offer online customers a great way to quickly find specific products on the web. In the same way, as e-commerce businesses are getting huge benefits by using AI technology, customers also get better experience and satisfaction by using AI-driven e-commerce portals in the form of intelligent visual search, advanced voice search, better aftersales services, etc. (Srivastava, 2021). A purchase made in this way is more likely to be considered a desired purchase rather than an impulse purchase, because the customer will have the intention of finding a specific electronic store and a specific product.

What should be done in order electronic store to be listed in the results of the GenAI Search platform? What this means for the content and structure of the electronic store? As AI-generated search tools become a growing source of online discovery, optimizing electronic commerce pages for AI-driven search is essential. While traditional SEO principles still apply, adapting to the way AI platforms process and retrieve information will be key to maintaining visibility and relevance.

Using a comprehensive labeling scheme for each product is essential, with details such as name, price, availability, description, brand, and especially aggregate ratings and reviews. AI searches depend on this structured data for accurate answers. For example, if a user asks a question about the price of a product, an AI search must answer with certainty if it sees structured pricing data. Unlike keyword-focused SEO, GenAI tools prioritize content that provides context and detail. Greater likelihood is ensured when product descriptions are rich, informative and written in natural language to match the way AI models process information.

One way to improve searches is to present the content of customer reviews on the page, highlighting the best reviews that talk about the use of the product. AI models integrate data from reviews when forming responses. If review text is available, the chance of AI extracting positive product attributes from real users increases. Many product pages now include an FAQ section, which is a good source of data for AI. Efficient solutions are providing clear question and answer pairs that AI can directly grab for search or chat answers, adding a small FAQ below each product with 3-5 common questions about the item (compatibility, materials, warranty, etc.) and concise answers. This structured Q&A content makes AI search result effective when a user asks a similar question.

Understanding what prompts an AI search to connect to an electronic store is valuable information to better understand how customers discover an electronic store and product and what their purchase intention is. The time for hesitation about artificial intelligence is over.

4. APPLICATION OF ARTIFICIAL INTELLIGENCE IN ELECTRONIC COMMERCE AND INFLUENCE ON THE CONVERSION RATE

Artificial intelligence (AI) is driving more innovation in electronic commerce, turning what was a controversial concept until recently into a competitive advantage. AI is no longer just an efficiency tool, AI is revolutionizing the way retailers personalize shopping experiences, optimize product discovery and drive customer engagement. Product search generated by artificial intelligence enables the right product to be connected with the right customers at the right time and under the most appropriate purchase conditions (product price, delivery type, delivery price, delivery time etc.), all in accordance with customer requirements. Artificial intelligence (AI) in electronic commerce can enhance online shopping by using advanced algorithms and machine learning to optimize customer experience and business operations.

DRAFT VERSION

The application of AI and machine learning enhances the sale of Amazon drastically, among most famous AI product Alexa is on the top, which helps to collect the training data sets for the algorithms are essential to Amazon's targeted marketing strategy; recommendation system of Amazon uses AI to reveal the product in demand based on customer searches and enhance the total sales by 35% (Kumar & Trakru, 2020). As search behavior evolves, electronic commerce brands that proactively adapt their SEO strategies for AI-driven search will gain a competitive advantage by tracking AI search, engaging AI-referred visitors, and optimizing content for AI-driven indexing.

AI enables electronic commerce platforms to analyze large data sets, enabling personalized product recommendations and trend predictions tailored to individual customers, provide 24/7 customer support, improve response time and satisfaction, predict demand, ensure optimal inventory levels, adjust prices in real time based on market trends, monitor competitor activity and customer behavior, detect fraudulent transactions, and store customer data.

Using artificial intelligence in electronic commerce can simplify operations, reduce costs and deliver highly personalized shopping experiences, driving growth and customer loyalty. As artificial intelligence technology develops, its role in electronic commerce will become even more important, offering innovative solutions to meet the demands of a fast-paced digital market. Retailers not yet adopting AI solutions need to start testing AI or risk being left behind.

As more brands turn to AI to help with product search, Amazon is looking beyond its platform to improve the customer experience. This move increases the chances for customers to find exactly what they need through familiarity with Amazon's store. Once customers are redirected to the brand's website, the brand handles all payments and orders. However, some brands can use the Amazon Prime feature, which allows customers to get the benefits of Prime shipping on their order, even if it's not sold by Amazon.

Previous updates to Amazon's search experience include visual and generative artificial intelligence. The company redesigned its customer support Chatbot and reported that customer satisfaction scores improved with a generative AI upgrade. The largest online retailer has added visual search features to its platform so customers can get faster and more accurate results.

Amazon products have an average conversion rate of 10% to 15%, Amazon Prime members have an even higher percentage, over 74% (Amazon Inc., 2024). That's because many shoppers browsing Amazon are already planning to buy something — if not right away, then soon. Another reason why so many e-retailers use Amazon as one of their platforms to sell their products. But probably the most important reason to track conversion rate is that it's one of the variables that propels ads to the top of search results. To rank ads, Amazon's algorithm takes into account conversion rates and speed of sales. As a result, if an ad has a low conversion rate, the product is less likely to appear on the first page of Amazon search results. Of course, these averages vary significantly by specialization, as with any measurement. Customers want to search and compare products before making a final purchase decision, so more expensive products often have lower conversion rates.

It's hard to say what constitutes a satisfactory conversion rate on Amazon because it varies depending on what it offers. For example, if the conversion rate is 50%, it will be satisfactory. But the situation would be different if you look at the data of this conversion in detail. Suppose there are four sessions and two orders. It is not satisfactory unless the product is sold at a very high price with a large profit margin. Let's assume a conversion rate of just 3%, based on 3,000 sessions and 90 product conversions. It is satisfactory, because it has a large number of sessions and sales.

5. CONCLUSION

The average conversion rate is a percentage that measures how effective an electronic store is at persuading users to take a desired action. The average conversion rate after the pandemic tends to increase, because habits have changed, and many customers see the real advantage of online shopping. A potential more effective application of AI in web search could be product categories that have lower conversion rates. For the effective use of AI-generated product searches that should result in an increase in conversion rates, it is necessary first of all to identify and monitor AI-based searches on the electronic store. Understanding what prompts an AI search to connect to an electronic store is valuable information to better understand how users discover an electronic store and product and what their purchase intentions are. As search behavior evolves, e-commerce brands that proactively adapt their search SEO strategies to be AI-driven will gain a competitive advantage. Previous updates to Amazon's search experience include visual and generative artificial intelligence. The largest online retailer has added visual search features to its platform so customers can get faster and more accurate results. The average conversion rate of e-commerce is between 2-4%, while Amazon products have an average conversion rate of 10% to 15%.

DRAFT VERSION

LITERATURE

Adobe for Business (2025). *Average ecommerce conversion rate benchmarks*. https://business.adobe.com/blog/basics/ecommerce-conversion-rate-benchmarks

Amazon, Inc. (2024). *Amazon Anual Report.* https://ir.aboutamazon.com/annual-reports-proxies-and-shareholder-letters/default.aspx

Buchholz, K. (Jul 7, 2023). *Threads Shoots Past One Million User Mark at Lightning Speed One Million Users*. https://www.statista.com/chart/29174/time-to-one-million-users/, datum jun 2025.

Kumar, T., & Trakru, M. (2020). The colossal impact of artificial intelligence. E-commerce: statistics and facts. *International Research Journal Engeenering Technologies* (IRJET), 6, 570-572.

Oosthuizen, K., Botha, E., Robertson, J., & Montecchi, M. (2020). Artificial intelligence in retail: The AI-enabled value chain. *Australasian Marketing Journal*, 29(3), 264-273. https://doi.org/10.1016/j.ausmj.2020.07.007

Pham, V. K., Pham Thi, T. D., & Duong, N. T. (2024). A Study on Information Search Behavior Using AI-Powered Engines: Evidence From Chatbots on Online Shopping Platforms. *SAGE Open*, 14(4). https://doi.org/10.1177/21582440241300007

Saleh, R. A., & Zeebaree, S. R. M. (2025). Artificial Intelligence in E-commerce and Digital Marketing: A Systematic Review of Opportunities, Challenges, and Ethical Implications. *Asian Journal of Research in Computer Science*, 18(3), 395–410. https://doi.org/10.9734/ajrcos/2025/v18i3601

Shafi, P.M., Jawalkar, G.S., Kadam, M.A., Ambawale, R.R., Bankar, S.V. (2020). AI—Assisted Chatbot for E-Commerce to Address Selection of Products from Multiple Products. In: Dey, N., Mahalle, P., Shafi, P., Kimabahune, V., Hassanien, A. (eds) *Internet of Things, Smart Computing and Technology: A Roadmap Ahead. Studies in Systems, Decision and Control*, 266, Springer, Cham. https://doi.org/10.1007/978-3-030-39047-1_3

SpeedCommerce (2025). *2025 eCommerce Benchmarks: Average Conversion Rates By Industry & By Year*. https://www.speedcommerce.com/insights/ecommerce-benchmarks-conversion-rates-by-industry-over-by-year/

Srivastava, A. (2021). The application & impact of artificial intelligence (AI) on E-commerce. *Contemporary issues in commerce and management*, 1(1), 165-75.

Yin J, Qiu X, Wang Y. (2025). The Impact of AI-Personalized Recommendations on Clicking Intentions: Evidence from Chinese E-Commerce. *Journal of Theoretical and Applied Electronic Commerce Research*. 20(1): 21. https://doi.org/10.3390/jtaer20010021

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

EXTENDING THE INTELLIGENT ENTERPRISE MODEL (IEM) FOR SMALL AND MEDIUM-SIZED ENTERPRISES BY INTEGRATING GENERATIVE ARTIFICIAL INTELLIGENCE

DRAGAN VUKMIROVIĆ[1], ŽEJKO BOLBOTINOVIĆ[2], TIJANA ČOMIĆ[3], RANKA POPOVAC[4]

[1] University of Belgrade, Faculty of Organizational Sciences, Belgrade, dragan.vukmirovic@fon.bg.ac.rs, 0000-0003-0248-016X
[2] Tekijanka, Kladovo, zeljko@tekijanka.com, 0009-0005-4190-7447
[3] University Business Academy in Novi Sad, The Faculty of Organizational Studies "Eduka" in Belgrade, tijana.comic@vos.edu.rs, 0000-0001-6556-5879
[4] Philip Morris International, Lausanne, ranka.popovac@pmi.com, 0009-0000-7207-4488

Abstract: *This paper presents an extension of the Intelligent Enterprise Model (IEM), a prototype designed for implementing artificial intelligence in small and medium-sized enterprises (SMEs). The original model, based on Big Data technologies and business analytics, has been enhanced by integrating concepts and tools from Generative Artificial Intelligence (Gen AI). The paper elaborates on how Gen AI, and specifically the concept of generative AI agents, transforms each phase of the IEM model, from organizational reengineering to evaluation and improvement, offering a new paradigm for building corporate intelligence in the digital age.*

Keywords: *Generative Artificial Intelligence, Intelligent Enterprise Model, AI Agents, Digital Transformation, Small and Medium-sized Enterprises.*

DRAFT VERSION

1. INTRODUCTION

The subject of research in the original paper was the implementation of information and communication technologies, with a particular focus on Big Data (BD) and Business Analytics (BA). To support the small and medium-sized enterprise (SME) sector in Serbia, the original Intelligent Enterprise Model (IEM) was defined (Vukmirović et al., 2019). The basic assumption of the model was that corporate intelligence (IQcorp) could be defined as a function of data-based knowledge:

$$IQcorp = f(data, knowledge) \quad (1)$$

However, the emergence of Generative Artificial Intelligence (Gen AI) has fundamentally changed the landscape of artificial intelligence applications in business. Gen AI, as a branch of AI that deals with creating new, original content (text, images, code), transcends traditional analytics. Unlike previous systems that primarily focused on analyzing and classifying existing data (discriminative AI), Gen AI possesses the ability to create, innovate, and make autonomous decisions.

This paper aims to present how the integration of Gen AI technologies extends the original IEM model, providing a practical framework for transforming small and medium-sized enterprises in the digital economy. The new, extended model (IEM-GenAI) views AI not merely as an analytical tool but as a creative and autonomous partner in business, capable of automating complex processes and fostering innovation through the use of generative AI agents.

The theoretical framework of the paper is based on the concept of the Intelligent Enterprise, defined as an organization's ability to collect, analyze, and apply data to make informed business decisions and respond quickly to environmental changes (Pan & Tenenbaum, 2002; Srinivasan, 2016). This concept initially relied on advanced business analytics, Big Data technologies, and continuous reengineering of business processes, enabling companies to be more agile and competitive in the digital economy. However, the rapid development of generative artificial intelligence (Gen AI) introduces a new dimension to the concept of intelligent enterprise. Generative AI expands the capabilities of traditional analytics by introducing the ability to autonomously create new content and ideas, thereby significantly surpassing analytical models based solely on existing data (Brown et al., 2020; Floridi & Chiriatti, 2020).

Generative AI agents, as autonomous systems capable of independent decision-making and task execution in complex business environments, occupy a central place in this extended concept. The theoretical basis for

AI agents is rooted in the theory of autonomous agents, which posits that agents possess the ability to perceive their environment, make decisions, act, and learn from interactions with their surroundings (Franklin & Graesser, 1997).

This approach to the theoretical framework integrates traditional intelligent enterprise concepts (data, analytics, and reengineering) with the new generative and autonomous capabilities of Gen AI technologies.

2. BASIC PREMISES OF THE EXTENDED MODEL (IEM-GENAI)

The original IEM model was based on the fundamental principles of Big Data and advanced business analytics. The extended model, now called IEM-GenAI, builds upon this foundation by introducing two key components from the domain of generative artificial intelligence (Gen AI):

- Generative Models (LLMs, etc.): Instead of an exclusive focus on analyzing unstructured data, IEM-GenAI leverages the capabilities of generative models, such as ChatGPT (OpenAI), Gemini (Google), and Claude (Anthropic), to create content, personalize communication, and automate tasks.
- Generative AI Agents: A concept that transforms AI from a passive tool (API) into a proactive executor. AI agents are autonomous systems capable of perceiving their environment, making decisions, and executing complex tasks to achieve predefined goals.

Generative models represent a distinct class of artificial intelligence models that can generate new, original content based on previously analyzed data. The most prominent representatives of this group are large language models (LLMs), such as the GPT series developed by OpenAI (Brown et al., 2020; Achiam et al., 2023). These models, which utilize deep learning techniques on vast datasets of textual data, can generate coherent texts, images, program code, or even video content, thereby surpassing traditional analytical approaches based solely on the classification or prediction of existing data (Goswami et al., 2024).

Generative AI agents represent an advanced class of autonomous AI systems that not only generate content but can also independently interact with the environment, make decisions, and perform complex activities. Unlike simple API-based AI tools, generative AI agents possess proactivity, autonomy, adaptability, and continuous learning, making them particularly suitable for dynamic and complex business environments (Hettiarachchi, 2025).

This expands and redefines the basic formula of the original IEM model, which was:

$$IEM = f(data, model, reengineering) \tag{2}$$

into the form:

$$IEM{-}GenAI = f(data, generative_model, AI_agent, reengineering) \tag{3}$$

The new model recognizes that enterprise intelligence depends not only on data analysis but also on the ability to generate new insights and act autonomously based on them. The iterative approach, described in the original paper as an "infinite learning process," remains crucial, but is now further strengthened by the ability of generative AI agents to learn and adapt to real-time changes.

3. IMPLEMENTATION PHASES OF THE EXTENDED MODEL (IEM-GENAI)

As illustrated in Figure 1, the extended IEM-GenAI model retains the original six iterative phases — self-evaluation, needs assessment, reengineering, implementation, evaluation, and improvement — but significantly transforms each of them through the integration of generative artificial intelligence. This transformation is not limited to the use of advanced analytical tools. However, it involves the introduction of generative models and autonomous AI agents, which enhance or fully automate key business processes.

In the following subsections, we describe how each of the six phases evolves from its original form in the IEM model to its extended version under the IEM-GenAI framework, demonstrating the shift from data-driven analysis to generative intelligence and autonomous execution.

DRAFT VERSION

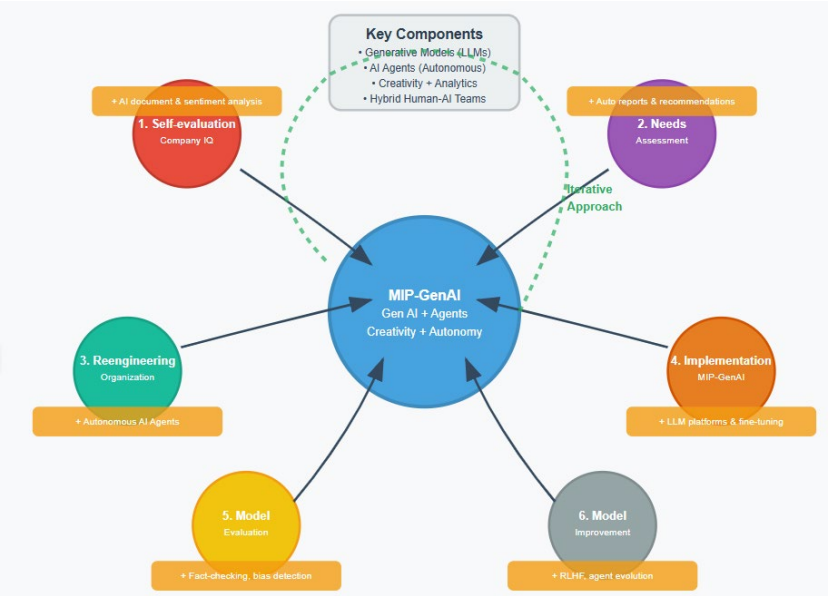


Figure 1. IEM-GenAI Model: Extension of the Intelligent Enterprise Model through Generative AI Integration

The six iterative phases — self-evaluation, needs assessment, reengineering, implementation, evaluation, and reporting — are illustrated in Figure 1. This new IEM-GenAI model unifies data, generative models, and AI agents within a single iterative process of continuous enterprise intelligence development.

Phase 1: Company IQ Self-Evaluation

- *Original Approach:* Assessing company IQ through management surveys on how "smarter and faster to react than a human" the company is.
- *Extension with Gen AI*: In addition to surveys, Gen AI enables deeper and faster analysis. Tools like ChatGPT or Gemini can analyze internal documents, meeting minutes, and business communications to identify key themes, sentiment, and implicit problems automatically. AI agents can simulate business scenarios and measure the speed and quality of team reactions, providing a more objective assessment of "company IQ."

Phase 2: Enterprise Needs Assessment

- *Original Approach*: A binary decision (YES/NO) on continuing implementation, motivated by the awareness that competitors are already using advanced analytics.
- *Extension with Gen AI*: Gen AI tools can actively assist in this phase. By analyzing business data (sales, supply chains, customer support, marketing), Gen AI can generate a detailed report on potential areas for improvement, quantify possible savings through automation, or propose new business opportunities based on generative capabilities (e.g., hyper-personalization of products). The decision is no longer based solely on the fear of falling behind, but on concrete, data-driven proposals generated by AI.

Phase 3: Reengineering the Existing Organization

- *Original Approach*: Transforming business processes by transitioning to a new technological paradigm and improving inter-departmental collaboration.
- *Extension with Gen AI*: This phase undergoes the most significant transformation with the introduction of generative AI agents. Reengineering no longer implies merely changing processes but also introducing autonomous agents that can take over entire workflows. For example, an AI agent can independently manage the customer support process, from initial contact to problem resolution, or manage the supply chain by predicting demand and automatically generating orders. The organization transforms into a hybrid model where human teams collaborate with AI agents.

Phase 4: IEM-GenAI Implementation

- *Original Approach*: Selecting an appropriate platform (BD, BA) and model, with options for in-house development or outsourcing.

DRAFT VERSION

- *Extension with Gen AI*: Technology selection now expands to Gen AI platforms such as OpenAI (ChatGPT), Google (Gemini), Anthropic (Claude), and others. Instead of just analytical models, specific Gen AI tools and agents are now implemented:
  - For sales: Smart assistants that help qualify clients and generate offers.
  - For supply chains: Gen AI agents have revolutionized supply chain management. They can autonomously predict demand with greater accuracy, considering complex variables and generating optimized procurement plans. Additionally, they can automatically generate and send purchase orders to suppliers, monitor delivery status in real-time, and identify potential disruptions in the supply chain long before they occur, proposing alternative solutions (Mindover, 2023).
  - For marketing: AI agents for generating SEO content, social media posts, and personalized email campaigns.
  - For operations: Agents for inventory optimization or dynamic pricing. The implementation steps table has been adjusted to include the selection and fine-tuning of suitable LLMs.

Phase 5: Model Evaluation

- *Original Approach*: Due to the nature of BD data, evaluation was limited and came down to a posteriori interpretation of results against achieved business outcomes. The problem of bias was also recognized.
- *Extension with Gen AI*: Evaluation challenges become even more complex, but also more defined. Gen AI introduces problems of hallucinations (generating inaccurate information), lack of transparency ("black box" problem), and deeper ethical bias. IEM-GenAI must include a more robust evaluation framework:
  - Structured Fact-checking: Every output used for decision-making must be verified, using procedures such as the "GenAI Response Verification Guide."
  - Bias and Fairness Measurement: Using technical metrics and human evaluation to detect gender, cultural, or other biases.
  - Comparative Consultation: Using multiple different Gen AI tools for the same task to compare results and reduce the risk of relying on a single, potentially erroneous, model.

Phase 6: Model Improvement

- *Original Approach*: The model is improved through a continuous influx of new data, making the "machine smarter" and overcoming the quality degradation of static models (concept drift).
- *Extension with Gen AI*: This principle is even more pronounced. IEM-GenAI is improved through:
  - Fine-tuning: Additional training of pre-trained models on specific company data to improve their relevance and accuracy for tasks.
  - Reinforcement Learning from Human Feedback (RLHF): A mechanism where human users directly rate and correct AI model responses, thereby actively "teaching" the model what desirable and undesirable outcomes are. This is a key technique for reducing hallucinations and bias.
  - Evolution of AI Agents: Agents are continuously improved through the analysis of the success of executed actions and the adaptation of their strategies.

In summary, each implementation phase in the original IEM model has not only been technologically upgraded but conceptually redefined through the integration of generative AI. By embedding generative models and autonomous agents into core business processes, IEM-GenAI introduces a new layer of enterprise intelligence—one that is not only reactive but also creative, adaptive, and proactive. This shift transforms the enterprise from a data-driven organization into a generative system capable of continuous learning and evolution, ultimately positioning it to thrive in dynamic and uncertain business environments.

4. EXAMPLES OF METRICS AND BENCHMARKS FOR IEM-GENAI EVALUATION

Evaluating the success of IEM-GenAI model implementation requires clearly defined metrics and comparability with relevant industry standards (benchmarks). The purpose of these metrics is to provide objective indicators of progress, justification of investments, and opportunities for further improvements.

The following are hypothetical examples of metrics and target values, formulated based on insights from relevant professional literature and best practices in the application of generative artificial intelligence. Sources dealing with theoretical aspects of AI output quality (Floridi & Chiriatti, 2020), evaluation of large language model performance (Brown et al., 2020), business effects of Gen AI application in small and medium-sized enterprises (McKinsey, 2023), as well as recommendations for assessing risk, compliance, and user experience in AI implementation (Gartner, 2024) have been considered.

DRAFT VERSION

The values shown are indicative and may vary depending on the industry, organizational maturity, process complexity, and technical capabilities. Their primary role is to serve as a starting point for SMEs aiming to monitor the progress of IEM-GenAI implementation and establish a meaningful internal evaluation system.

To operationalize this evaluation, Table 1 summarizes selected key **AI agent performance metrics**, focusing on automation efficiency, accuracy, and user satisfaction. These metrics help assess the degree to which generative AI agents contribute to process automation and customer experience.

Table 1. AI Agent Performance Metrics

Metric	Definition	Target / Benchmark
Task Completion Rate (TCR)	% of tasks completed without human intervention	>85% (routine tasks); >70% (complex tasks)
Mean Time to Resolution (MTTR)	Avg. time to resolve a task or issue	50–70% reduction vs. manual processes
Error Rate	% of errors in AI-generated outputs	<5% (critical); <10% (non-critical)
Customer Satisfaction Score (CSAT)	(Positive ratings / Total ratings) × 100	>80% target; Benchmark: 75–85% (traditional), 80–90% (AI-enhanced)

In parallel, Table 2 outlines business performance metrics, which offer a broader perspective on the financial and operational impact of IEM-GenAI adoption. These indicators enable SMEs to track ROI, cost efficiency, and productivity gains over time.

Table 2. Business Performance Metrics

Metric	Definition	Target / Benchmark
AI Implementation ROI	(Savings + Additional Revenue – Costs) / Costs × 100	150–200% within 2–3 years
Cost Per Task (CPT)	Avg. cost per completed task	60–80% reduction vs. manual processes
Tasks Automated Ratio	% of tasks automated	Year 1: 30–40%; Year 2: 50–70%
Employee Productivity Index	Number of tasks completed per employee	25–40% increase

Due to space limitations, this section presents only a selected subset of the most relevant metrics across core dimensions. While not exhaustive, the examples in Tables 1 and 2 aim to illustrate how SMEs can begin structuring a practical evaluation framework for IEM-GenAI implementation. By aligning empirical targets with insights from academic and industry sources, the proposed metrics provide a solid foundation for tracking performance and driving continuous improvement.

Future efforts, whether through academic research or internal company initiatives—may expand these metrics to include aspects such as ethical risk, content quality, user experience, or the long-term impact of innovation. This ensures that the evaluation of generative AI integration remains comprehensive, evidence-based, and aligned with evolving business priorities.

DRAFT VERSION

5. CONCLUSION

The original Intelligent Enterprise Model (IEM) laid solid foundations for the application of artificial intelligence in small and medium-sized enterprises, with a focus on the power of data and analytics. The extended model, IEM-GenAI, evolves this concept by integrating Generative Artificial Intelligence. This integration transforms IEM from an analysis system into a system for creation and autonomous action.

By introducing generative models and, most importantly, AI agents, enterprises gain a partner capable of automating complex workflows, personalizing client interactions in an unprecedented way, and generating innovative ideas. However, this new power also brings new responsibilities. Challenges such as hallucinations, lack of transparency, bias, and data protection become central issues requiring rigorous evaluation procedures and constant human oversight.

The goal remains the same: to build an intelligent company that is smarter and faster. However, with Gen AI, "intelligence" is no longer just analytical, but also creative and operational. IEM-GenAI, therefore, represents a significant step forward in realizing the vision of a knowledgeable enterprise, prepared for the challenges and opportunities of the 21st-century digital economy. Just like the original IEM, this extended model remains a prototype that will continue to be developed and tested, confirming that in the world of artificial intelligence, the learning process never ends.

For future research, two key directions emerge. The first direction involves a detailed empirical evaluation of the IEM-GenAI model in the small and medium-sized enterprise sector, conducted through concrete case studies and measurements of the model's impact on business performance. The second direction involves the development of specific generative AI tools and agents tailored to different industries within the SME sector, enabling easier application and faster realization of benefits from artificial intelligence, particularly in environments with limited resources and technical capacities.

LITERATURE

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://arxiv.org/abs/2005.14165

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681-694.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Gartner. (2023). Top Strategic Technology Trends for 2024: AI Trust, Risk and Security Management (AI TRiSM). Retrieved from https://www.gartner.com/en/articles/gartner-top-10-strategic-technology-trends-for-2024

Goswami, A., Kaur, G., Tayal, S., Verma, A., & Verma, M. (2024, August). Analyzing the efficacy of Deep Learning and Transformer models in classifying Human and LLM-Generated Text. In 2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA) (pp. 1-5). IEEE.

Hettiarachchi, I. (2025). Exploring generative AI agents: Architecture, applications, and challenges. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 8(1), 105-127.

McKinsey & Company. (2023). The economic potential of generative AI: The next productivity frontier. Retrieved from https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

Mindover. (2023). Why should supply chain managers invest in AI? ERP Software Blog. Preuzeto sa https://www.erpsoftwareblog.com/cloud/2023/06/why-should-supply-chain-managers-invest-in-ai/

Pan, J. Y. C., & Tenenbaum, J. M. (2002). An intelligent agent framework for enterprise integration. IEEE Transactions on Systems, man, and cybernetics, 21(6), 1391-1408.

Srinivasan, V. (2016). The intelligent enterprise in the era of big data. John Wiley & Sons.

Vukmirović, D., Čomić, T., Bolbotinović, Ž., Dabetić Đ and Jovanović Milenković, M. (2019). MIP - prototip modela inteligentnog preduzeća, Zbornik radova - SYMOPIS 2019, Editrori: Martić, M., Makajić-Nikolić, D. i G. Savić. Univerzitet u Beogradu. Fakultet organizacionih nauka. ISSN: 978-86-7860-363-7. str. 536-54.

DRAFT VERSION

DATA SCIENCE

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

A MODIFIED DBSCAN ALGORITHM FOR TRAFFIC ACCIDENT CLUSTERING

JORDAN ATANASIJEVIĆ[1], DANIJELA PROTIĆ[2], DIMITRIJE KOLAŠINAC[3], STEFAN IVANOVIĆ[4],

[1] Center for Applied Mathematics and Electronics, Belgrade, Serbia, atanasijevicjordan@gmail.com, 0009-0003-0034-9974
[2] Center for Applied Mathematics and Electronics, Belgrade, Serbia, danijelaprotic318@gmail.com, 0000-0003-0827-2863
[3] Center for Applied Mathematics and Electronics, Belgrade, Serbia, dimitrije.kolasinac@vs.rs, 0009-0008-9505-2482
[4] Center for Applied Mathematics and Electronics, Belgrade, Serbia, stefanivanovic1012@gmail.com, 0009-0000-1788-2095

Abstract: *Clustering is a machine learning approach used in wide range of scientific areas. The Density-Based Spatial Application with Noise (DBSCAN) algorithm creates clusters based on density. A modified DBSCAN (M-DBSCAN) algorithm is an upgraded version of the DBSCAN algorithm that incorporates weights as parameters to provide more precise cluster detection. Weights correspond to severities of traffic accidents. Road accidents resulting in minor injuries weight 1, road accidents resulting in serious injuries weight 10, and road accidents resulting in fatalities weight 85. The "Opasna mesta" application is developed to enable quick and convenient data management from the black spot database. The results of the experiments conducted on data on traffic roads from the Republic of Serbia reveal that M-DBSCAN provides better clustering results than DBSCAN algorithm.*

Keywords: *Density-based clustering, DBSCAN, modified-DBSCAN, traffic accidents.*

1. INTRODUCTION

In recent decades, the problems that follow the storage, processing and transmission of data have become a major research task and a significant challenge in daily business practice. Data analysis plays an important role in understanding different phenomena and can be used in a variety of research areas. Machine learning (ML) algorithms, which relate to statistical analysis focus on prediction, optimization and data analysis, are adaptive in a dynamic environment and can be evaluated in order to give correct results, even if the input parameters are (un)known. There are four basic types of ML model evaluation: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Clustering is an unsupervised machine learning technique that combines comparable data in clusters, according to their features. Cluster analysis can be used in various problem-solving methods (Xu et al., 2005). For this reason, many cluster analysis tools have been created, each of which has its own strengths and weaknesses, due to the complexity of the information (Rodriguez et al., 2019). A Density-Based Spatial Application with Noise (DBSCAN) algorithm identifies the clusters based on local data density (Singh et al., 2022). Its core parameters include maximum radius within which two points are considered neighbor in the same cluster ε, and the minimum number of points in the maximum distance, *MinPts*, required within that radius to define dense region to cluster formation. In this article, we introduce a modified DBSCAN algorithm (M-DBSCAN), which extends original framework by enhancing its applicability to multidimensional data. In addition to the spatial identifiers data which are used within DBSCAN, M-DBSCAN integrates both spatial identifiers and additional dataset attributes (numerical or textual) by transforming them into numerical values suitable for clustering. Furthermore, the weights are used alongside density-based criteria, for more precise distinction of clusters.

The spatial distribution of traffic accidents across a traffic network is not uniform. Road segments having a higher number of accidents, over a defined period, relative to the average of similar segments are considered dangerous. "Black spots" refer to specific areas or locations where traffic accidents involve fatalities, serious injuries and property damage are concentrated. Yadav et al. (2022) define a black spot as a 500-meter road stretch where either five traffic accidents resulted in fatalities and serious injuries occurred, or where 10 fatalities have been recorded within three-year period.

In this study the M-DBSCAN algorithm is applied to traffic accident data from the Republic of Serbia. To enhance clustering precision the algorithm assigns different weights to accident severity. Fatal accidents are weighted 85, serious injuries are weighted 10, and minor injuries accidents are weighted 1. The results show

DRAFT VERSION

that the M-DBSCAN algorithm is more precise and context-aware in high-risk traffic zones identification, compared to the standard DBSCAN algorithm.

2. RELATED WORK

The effectiveness of DBSCAN and in identifying high-risk zones and patterns within traffic networks, based on DBSCAN algorithm modification is proposed in Qui et al. (2016), Topcuoglu et al. (2022), and Li and Chen (2024), and Athanesious (2021). The researches underscore DBSCAN usage for traffic safety analysis, particularly when enhanced with domain-specific modifications and integrated with geospatial data. Qiu et al. (2016) proposed a modified DBSCAN algorithm for accident-prone locations detections on expressways in China. The authors emphasize spatial clustering accuracy and parameter sensitivity. In Topcuoglu et al. (2022), the authors analyzed the speed-related traffic accidents in Turkey, identifying clusters of dangerous road segments. Li and Chen (2024) examined road safety across Southeast Asia, and algorithm's adaptability in multi-criteria decision-making. In addition, Athanesious (2021) applied a DBSCAN-based anomaly detection framework, in the domain of traffic surveillance, to identify irregular vehicle trajectories from real-time monitoring video-data. Moreover, Khan et al. (2014) discuss the DBSCAN robustnest to choice of input parameters.

Based on geospatial coordinates of traffic accidents database Wang et al. (2024) identified locations where traffic accidents tend to accumulate spatially, forming black spots. Black spots are also defined by Yadav et al. (2022) in terms of traffic accidents and/or fatalities recorded within three-year period.

3. DBSCAN ALGORITHM

A DBSCAN is a density-based clustering algorithm that identifies clusters of arbitrary shapes and sizes within a dataset. The DBSCAN algorithm groups points that are close to each other within a given radius, ignores isolated points, and recognizes clusters regardless of their form. The algorithm does not require the number of clusters to be determined a priori. However, despite its robustness to noise, adaptability to non-convex cluster shapes, and independence from predefined cluster counts, the DBSCAN performance is sensitive to the choice of input parameters, and computationally complex, that can be significant challenge for large-scale datasets (Khan et al., 2014). The DBSCAN algorithm operates based on the maximum distance between two points (one considered to be neighbor to the other) ε, and minimal number of other points *MinPts* required to form dense region. Each data point can be classified as: core point (have at least *MinPts* within a given distance ε), border point (lies within range of core point but do not have *MinPts*), or noise point (does not belong to any cluster). Users control how the algorithm defines clusters by adjusting *MinPts* and ε.

The DBSCAN algorithm is a step-wise process of examining neighborhood of each point in the dataset to identify clusters based on the density of data points (Kumar, 2024), as follows:

1) Specify parameters *MinPts* and ε
2) Select an arbitrary unvisited initial point
3) Examine the neighborhood
4) Expand the cluster if the point is core point
5) Repeat the process until all points have been processed
6) Finalize clusters
7) Address remained noise points

According to the Ester et al. (1996), ε is empirically best determined when *MinPts* =4. To identify the cluster, the DBSCAN algorithm starts at the initial point *p* and recursively evaluates all points *q* that are density-reachable from *p* under the given parameters. If *p* is a central point, this procedure yields the cluster under the conditions ε and *MinPts*. If *p* is a boundary point, there are no density-reachable points, and the DBSCAN algorithm proceeds with the next candidate.

The DBSCAN algorithm is highly effective anomaly and noise detection, clustering radar reflection, mapping disease epicenters, customer segmentation, storm tracking, and geospatial event detection of global positioning system (GPS) – based locations, including the assessment of traffic safety factors that influence the occurrence of traffic accidents and accumulation of traffic accidents with similar characteristics.

4. MODIFIED DBSCAN ALGORITHM

When consider locations with a high risk of accidents, there are no standardized methods to identify black spots on the roads. The huge amount of data is a great challenge, although there is a number of mechanisms that enable the best decision possible. In modified DBSCAN, weights are assigned to each accident based on

DRAFT VERSION

severity and frequency. These weights influence the density estimation by adjusting the effective neighborhood radius per point. The algorithm is based on a parameterized mathematical model for identification of clusters (i.e. areas) around black spots, using data of the geographical position in which the accidents occurred and type of accidents, which can include injuries and/or fatalities. Unlike the DBSCAN, the M-DBSCAN operates under the assumption that, in addition to spatial identification, a weight can be assign to the point for which cluster is being sought (Eq. 1).

$$w_i = \alpha \cdot S_i + \beta \cdot F_i, \tag{1}$$

where S_i is the normalized severity score, F_i is the local frequency of similar accidents. The parameters α and β control the relative influence of each factor. This formulation allows the algorithm to prioritize high-impact of frequency occurring incidents during clustering. Eq. 1 aggregates weights to normalize the distribution but each instance receive its own weight.

Based on geospatial coordinates of traffic accidents database (publicly available data set downloaded from https://data.gov.rs/sr/datasets/podatsi-o-saobratshajnim-nezgodama-po-politsijskim-upravama-i-opshtinama/, 2025) it was possible to identify locations where traffic accidents tend to accumulate spatially, forming black spots (Wang et al., 2024). The coordinates along with accident-specific attributes serve as inputs to the M-DBSCAN algorithm that includes both density and severity. Unlike standard DBSCAN, the M-DBSCAN does not require the number of clusters to be known in advance**,** which is ideal for large, unstructured datasets. The severities of road accidents are defined by the International Road Workers' Association (PIARC, 2008), as follows: road accidents involving minor injuries – weight 1, road accidents involving serious injuries – weight 10, and road accidents with fatalities – weight 85. The clusters identified around black spots are determined as follows:

$$WNI = 85 \cdot n_{fatalities} + 10 \cdot n_{serious_injuries} + n_{minor_injuries}, \tag{2}$$

where WNI is the Weighted Number of traffic-accidents Indicator within a specified radius, and $n_{fatalities}$, $n_{serious_injuries}$, and $n_{minor_injuries}$ represent numbers of road accidents with fatalities, seriously injured persons, and participants with minor injuries, respectively.

The M-DBSCAN algorithm proceeds through the following steps:

1. *Parameter initialization*: cluster radius *r*, cumulative cluster radius $r_{sum} = 0$, threshold *k*, weights assigned to each data point, cumulative distance sum within potential cluster *Suma*=0, number of internal clusters b_{uk}=0, number of external clusters b_{sk}=0, and the incremental cluster radius value *d*=0, which will be updated in subsequent iterations.
2. *The algorithm calculates the sum of weighted values within each cluster* defined by *r*, and checks whether any previously identified clusters exist. If overlapping clusters are detected, they are merged until only one remains.
3. *Weight assignment* based on severity and frequency attributes.
4. *Pairwise distance calculation* for every data point to all other points in the data set.
5. *Cluster evaluation*. Calculate the sum of weighted values within the initially defined cluster. Clusters that satisfy condition $Suma \geq k$ are identified and subsequently analyzed in the following step.
6. *Clusters comparison and selection*. Compare analyzed cluster with those meeting threshold condition, in descending order of their total weighted values. If two clusters to overlap, the less independent cluster (i.e. the one with the lower *Suma*). There are two scenarios in decisions
   a) *The center of a cluster lies within the radius of the other cluster*. The cluster with the lower weighted value is rejected and the internal cluster count is updated as: $b_{uk} = b_{uk} + 1$.
   b) *Two clusters overlap, but the center of one cluster lies outside the radius of the other cluster*. The cluster with the lower total value is discarded. The cumulative radius is updated $r_{sum} = r_{sum} + d - r$, and the external cluster count is incremented: $b_{sk} = b_{sk} + 1$.
7. *Iteration and stopping criteria*. If not overlap is detected than compute $d = r_{sum}/b_{sk}$. The algorithm then iterates through all clusters, repeating steps 4 and 5 until clusters with the largest distances are identified.

5. EXPERIMENTAL RESULTS

To demonstrate the advantages of the M-DBSCAN algorithm, the experiments were conducted using open database on traffic accidents in Serbia. The geostatistical and spatial analysis were applied to open data, based

DRAFT VERSION

on the ArcGIS Pro environment. A detailed overview of the ArcGIS Pro is available at https://data.gov.rs while the structure of the open dataset is described in the work of Gladović & Deretić (2018).

To facilitate practical data management, custom application named *Opasna mesta* was developed. Its main interface includes six combo-boxes (*Opšti podaci o putu*, *Opasna mesta*, *Saobraćajne nezgode*, *Saobraćaj*, *Izveštaji* and *Pomoć*) and three command buttons *(Dijagram saobraćajnih nezgoda*, *Radovi na putu* and *Odjava*). The application retrieves data from all recorded traffic accidents via national open data portal, covering incidents across Serbia from 2015 to the present. Users can filter results by city, search radius, minimum *WNI* value, and time period based on attributes stored in a relational database.

Using geospatial coordinates, the M-DBSCAN algorithm identifies coordinates with concentrated occurrence of traffic accidents. It begins by extracting spatial coordinates and accident types to assign severity weight to each location within a 300m radius, focusing on period from 2022 to 2024. Hazardous road segments are ranked in descending order, according to the *WNI* (see Eq. 2). If $WNI \geq 85$, the algorithm proceeds with in-depth analysis of this cluster. According to the defined procedure, all points meeting or exceeding defined threshold undergo further evaluation. Pairwise distances between points are then calculated and, if the distance is less than 300 meters, the weighted values are added to the cluster sum. In the next step, the cluster radius is incrementally expanded based on the average value of surrounding clusters. In this study, the maximum radius was 479 meters; once the radius exceeds 500m, the algorithm terminates. Figure 1 illustrates the results for municipality of Arilje. The standard DBSCAN identified a single black spot (dashed red circle) whereas M-DBSCAN detected two distinct black spots in the same area, demonstrating its enhanced sensitivity and clustering precision.

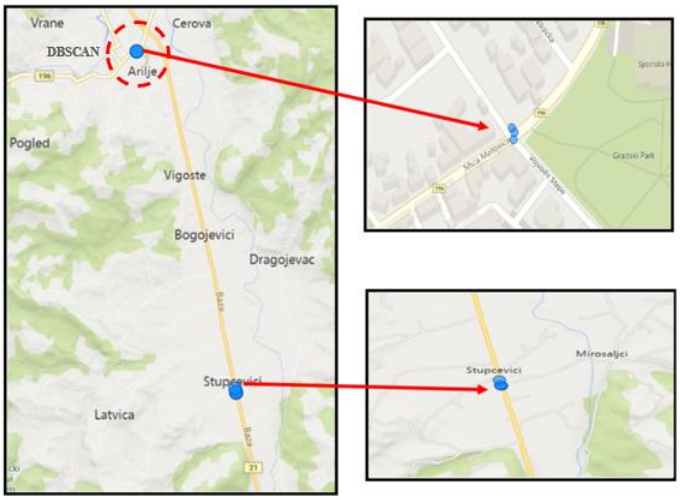


Figure 1: Black spots in the area of Arilje (2022-2024)

As it is noted before, the M-DBSCAN algorithm generates a highly detailed spatial map due to the increased number of clusters it identifies. Figure 2 illustrates this effect by presenting both the centroids (left) and the corresponding clusters of black spots detected in the Belgrade area. This example highlights the algorithm's enhanced sensitivity in distinguishing multiple high-risk zones within dense urban environments, offering a more granular view compared to standard clustering.

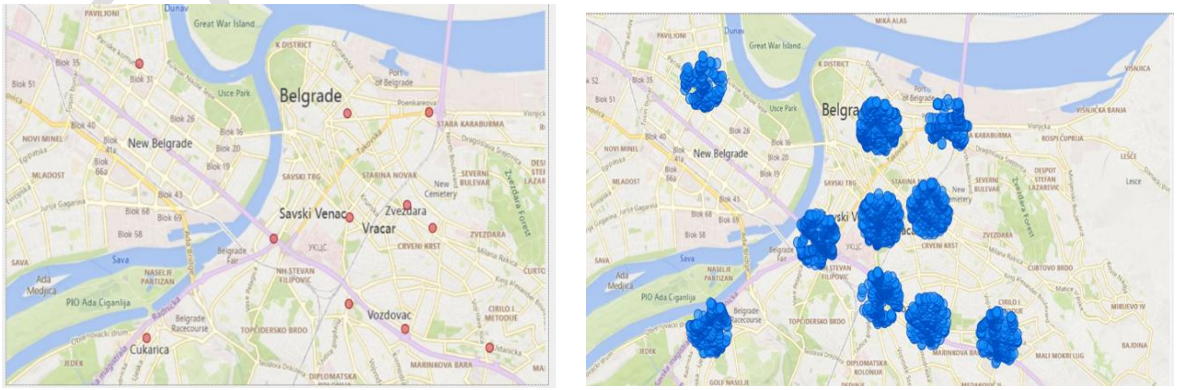


Figure 2: Centroids (left) and black spots identified in Belgrade area

6. CONCLUSION

Clustering techniques are useful for extracting knowledge from a large amount of geospatial data. The DBSCAN algorithm is density-based algorithm that forms clusters on the basis of non-dense and dense regions. The most important parameters of the DBSCAN algorithm are the maximum distance between two points in the same cluster and the minimum number of points in the maximum distance between two points to form a cluster. The M-DBSCAN algorithm is a modified version of the DBSCAN algorithm. The algorithm is based on the assumption that the points for which clusters are sought, in addition to spatial identification, have weights that refer to severity of road accidents. Traffic accidents are weighted according to severity: incidents resulting in minor injuries carry a weight of 1, accidents involving serious injuries are assigned a weight of 10, while accidents with fatalities are given the highest weight of 85. The identified black spots are ranked according to the value of the *WNI*. All spots whose *WNI* are equal to or greater than the specified threshold are considered potential black spots and are subject to further analyses. The key advantage of the algorithm lies in its ability to operate without prior specification of the number of clusters, while it achieves more precise cluster identification by including weights as additional parameters for clustering.

In order to enable a simple and comfortable data management, a user-friendly application called *Opasna mesta* was developed. It is shown that the M-DBSCAN algorithm generates a highly detailed spatial map due to the increased number of clusters it identifies. The algorithm's enhanced sensitivity is useful in distinguishing multiple high-risk zones within given environments.

In our further work we will consider using the M-DBSCAN in two directions. First, we intend to apply this algorithm in the area that are not related to the spatial clustering. The idea is to use the different types of publicly available data. The other direction is to use the M-DBSCAN algorithm in various scientific areas, such as data protection in preserving sensitive datasets (medical records, financial logs) to protect secure similar computation for clustering user behavior without revealing the user activity patterns, and AI-driven feature selection to detect anomalies without exposing the raw traffic. The intention is also to integrate the M-DBSCAN algorithm with artificial intelligence, to improve algorithms for feature selection, such as Artificial Rabbit Optimization (ARO)-DBSCAN.

REFERENCES

Athanesious, J. J. (2021). Anomaly detection using DBSCAN in traffic video surveillance. https://www.researchgate.net/publication/356789012_Anomaly_Detection_Using_DBSCAN_in_Traffic_Video_Surveillance

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Inkdd*, *96*(34), 226-231.

Gladović, P., & Deretić, N. (2018). Otvoreni podaci i baze podataka u analizi saobraćajnih nezgoda u gradu Beogradu. *Tehnika*, *73*(2), 247-253.

Li, J., & Chen, F. (2024). A hybrid EWM–PROMETHEE II–DBSCAN model for road safety monitoring. *Frontiers in Public Health*, *12*, 1413031.

Khan, K., Rehman, S., Aziz, K., Fong, S., & Sarasvady, S. (2014). DBSCAN: Past, present and future. *Proceedings of the 5th International Conference on the Applications of Digital Information and Web Technologies*, 17-19 February 2014, Bangalore, India. https://doi.org/10.1109/ICADIWT.2014.6814687

Kumar, R. (2024). A guide to the DBSCAN clustering algorithm. https://www.datacamp.com/tutorial/dbscan-clustering-algorithm

PIARC. (2008). *Road safety manual. Recommendation from the World Road Association*. Chapter 7.

Podaci o saobraćajnim nezgodama po policijskim upravama i opštinama. (2025). https://data.gov.rs/sr/datasets/podatsi-o-saobratshajnim-nezgodama-po-politsijskim-upravama-i-opshtinama/

Qiu, C., Xu, H., & Bao, Y. (2016). A modified DBSCAN algorithm for identifying traffic accident prone locations. *In IDEAL 2016 – Intelligent Data Engineering and Automated Learning*, 106–114. Springer, Cham.

Rodrigurez, M., Comin, C., Casanova, D., Bruno, O., Armancio, D., da F. Costa, L., & Rodriguez, F. (2019). Clustering algorithms: A comparative approach. *PloS One*, *14*(1), e0210236. https://doi.org/10.1371/journal.pone.0210236

DRAFT VERSION

Singh, H., Girdhar, A., & Dahya, S. (2022). A literature survey based on DBSCAN algorithms. *Proceedings of the 6th International Conference on Intelligent Computing and Control Systems*, 25-27 May 2022, Madurai, India, 751-758. https://doi.org/10.1109/ICICCS53718.2022.9788440

Topcuoglu, B., Memisoglu Baykal, T., & Tuydes Yaman, H. (2022). GIS-based DBSCAN and NNH clustering for speed-related traffic accident analysis. *ISPRS Archives, XLVIII-4/W1-2022*, 487–492.

Wang, Y., Zhao, Q., Xiu, W. (2024). Blackspot identification of abnormal driving behavior based on improved DBSCAN algorithm. *Proceedings of the 3rd Conference on Electronic Information Engineering, Big Data, and Computer Technology*, 13181, 292-299.

Xu, R., & Wunsch, D. (2005). Survey on clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645-678. https://doi.org/10.1109/TNN.2005.845141

Yadav, D. K., Ghodmare, S. D., & Kumar, N. N. (2022). Mitigation of Blackspots on Highways by the Application of Safe System Approach. *Proceedings of Materials Today*, *52*, 1228-1235. https://doi.org/10.1016/j.matpr.2021.11.042

APPENDIX

Data on traffic accidents from 2022 to 2024 are available at: https://data.gov.rs/sr/datasets/podatsi-o-saobratshajnim-nezgodama-po-politsijskim-upravama-i-opshtinama/,

The M-DBSCAN algorithm implementation and user-selected parameters from the relational dataset are available upon request form the corresponding author.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

CLUSTER ANALYSIS OF CONSUMER ATTITUDES TOWARDS ECO-FRIENDLY PRODUCTS IN WESTERN BALKANS: AN OPERATIONS RESEARCH APPROACH

BOJAN KITANOVIKJ[1], VIOLETA CVETKOSKA[2], ALEK PACHEMSKI[3]

[1] Ss. Cyril and Methodius University in Skopje, Faculty of Economics – Skopje, North Macedonia, bojan.kitanovikj@eccf.ukim.edu.mk, 0000-0002-1968-8955
[2] Ss. Cyril and Methodius University in Skopje, Faculty of Economics – Skopje, North Macedonia, vcvetkoska@eccf.ukim.edu.mk, 0000-0002-0472-5522
[3] Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering, North Macedonia, alek.pachemski@students.finki.ukim.mk

Abstract: *The study sets out to reveal actionable patterns in consumer attitudes toward eco-friendly products across the seven Western-Balkan economies, thereby supporting SDG-aligned policy and business decisions. Employing 141 Eurostat time-series indicators from 2014-2022 that capture environmental, socio-economic and behavioural dimensions we experiment with t-SNE and rank variables through Random-Forest importance and clustering the countries with a triad of techniques – K-means, hierarchical agglomeration and K-Nearest-Neighbours network graphs, each benchmarked with silhouette scores and cross-method checks. A three-cluster solution proves most informative: Albania consistently forms a high-risk/low-priority segment; Croatia oscillates in a medium-risk, transitional tier; while Bosnia and Herzegovina, Kosovo, Montenegro, North Macedonia and Serbia group as low-risk/high-priority adopters, with Montenegro emerging as a bridging hub and Serbia showing an isolated profile. These insights underline how EU integration, regional linkages and country-specific socio-economic contexts shape sustainable consumption, and they equip governments with levers such as targeted financing for Albania, knowledge-sharing platforms anchored in Montenegro and bespoke CSR incentives for Serbia, while giving firms a data-driven segmentation map for marketing eco-products and a replicable analytics workflow for tracking SDG progress in small-sample regions.*

Keywords: *Operations research clustering, Consumers, Eco-friendly products.*

1. INTRODUCTION

Sustainable consumption is a key concept in the modern economy, encompassing ecological, social and economic aspects of behaviour on consumers. Sustainable consumption aims to reduce the negative impact on the environment while maintaining economic and social benefits (Wang et al., 2019). According to the Sustainable Development Report (Sachs et al., 2021), four of the six determined priorities emphasize the need for sustainable consumption, focusing on reducing environmental damage caused by industry and consumers. The research shows that engagement in sustainable consumption practices is more common among individuals with higher income and educational levels (Verain et al., 2015).

In recent years, the global focus on sustainability and environmental responsibility has significantly influenced consumer behaviour and corporate strategies. The Balkan region, characterized by diverse economic landscapes and different levels of development, presents a unique context for researching consumer attitudes towards eco-friendly products, as the adoption of these products remains uneven across these countries (Marić & Vukmirović, 2018). Understanding these behaviours is crucial for decision-makers, businesses and stakeholders that aim to promote sustainable consumption and production patterns in the region (Lukman et al., 2016; Licastro & Sergi, 2021). Considering this, the transition to a green economy addresses environmental challenges and at the same time offers economic opportunities, such as job creation in sustainable industries and strengthening on competitive advantages on businesses which adopt eco-friendly practices (Thies et al., 2019). By conducting a cluster analysis of consumer behaviour in the Balkans, this study aims to identify distinct patterns and prioritize regions for targeted interventions through Random-Forest importance and clustering the countries with a triad of techniques – K-means, hierarchical agglomeration and K-Nearest-Neighbours network graphs.

DRAFT VERSION

2. DATA AND METHODOLOGY

The data used in this study include a partial collection of socio-economic and environmental indicators covering seven Balkan countries (Serbia, North Macedonia, Albania, Bosnia and Herzegovina, Croatia, Kosovo, and Montenegro) in the period from 2014 to 2022. This period encompasses significant political changes and economic transformations that likely influenced consumer behaviour and the adoption of eco-friendly products (Nikolić et al., 2022). Additionally, this period was the most consistent in terms of available data from Eurostat.

Specifically, the dataset contains 141 time-series characteristics collected exclusively by Eurostat, ensuring high reliability and standardization. on the variables. The key features in the dataset have been carefully selected to cover the different dimensions of sustainability and behaviour on consumers. A total of 18 data sets were used, of which the following are the most significant: 1) municipal waste management, 2) household energy consumption, 3) organic farming area, 4) e-government activities, 5) poverty rates, 6) import on goods and services, 7) satisfaction from life.

By integrating these different indicators, the study created a holistic model that aims to capture the complex interplay between different determinants of consumer behaviour. Moreover, the time-sequential character of the data allowed for the identification of temporal trends and displacements in the priorities to consumers, allowing for a better understanding of how consumer behaviour evolves in response to political changes, economic development, and environmental initiatives (Lukman et al., 2016). Additionally, the data was used to validate and compare different clustering methodologies, including K-Means, K-Nearest Neighbours (KNN), and network analysis graphs, which provides stability and reliability of clustering results (White & Lee, 2009).

To gain detailed insights into consumer behaviour, a factor analysis was conducted to identify the key factors which affect sustainable behaviour. Furthermore, the cluster analysis allowed for the categorization of behaviour according to the level of sustainability. This approach identifies consumer movements as “non-sustainers,” “reducers,” “product-oriented,” and “supporters.” The methodology includes a systematic approach to data preprocessing, selection of relevant features, application of techniques for reduction on dimensionality and implementation of different clustering algorithms for analysing consumer behaviour towards eco-friendly products in the Balkans. Given the complexity and multidimensionality of the data, a combination of statistical and machine learning algorithms was used to ensure reliable and meaningful clustering results.

Moreover, before the data was analysed, we first pre-processed the data. To address missing values, the mean imputation method was applied. Given the diverse range of feature values, standardization was necessary to standardize the data and yes everything reduce the impact of different scales on clustering algorithms. Two standardization techniques were applied – StandardScaler and MinMaxScaler. The two scaled data sets were stored separately to facilitate comparative analysis.

The dataset contained 141 time series on characteristics, what results with high dimensionality that can hinder the performance and interpretability of clustering algorithms. To address this, dimensionality reduction techniques were applied to extract the most informative components of the data. First, a Principal Component Analysis (PCA) was used to reduce on dimensionality of the data, retaining 95% from the variance. This technique identifies the main components that capture the maximum variability in the data, effectively summarizing the data set with a smaller number of features. Then, although primarily used for visualization, t-Distributed Stochastic Neighbour Embedding (t-SNE) was experimented with to explore the possibility of visualizing high-dimensional data in a lower-dimensional space, as it is particularly effective at capturing local structures and detecting clustering patterns.

To refine the clustering process and focus on the most influential variables, feature selection was performed using Random Forest, a robust ensemble-based machine learning method known for its ability to rank features according to their contribution to the accuracy of predictions. Random Forest was applied to the dataset to evaluate the importance on each characteristic in the prediction of consumer behaviour towards eco-friendly products (Figure 1). The characteristics were ranked according to their importance ratings, as the top-ranked features were selected for precise clustering. This step ensures that the clustering algorithms are guided by the most relevant and influential variables, thus improving accuracy and interpretation of the resulting clusters.

The following clustering techniques were applied: 1) k-means clustering – the primary algorithm used to categorize countries into different groups based on their consumer behaviour patterns, 2) KNN and network graphs analysis – for constructing Euclidean distances between countries based on foundation on their features and understanding the connections between nodes, 3) hierarchical clustering – as an alternative to k-means, was used to evaluate consistency on results, building a hierarchy of clusters through agglomerative

DRAFT VERSION

(bottom-up) or divisive (top-down) approaches, allowing for the identification of nested clusters and providing a more detailed understanding of the data structure. The k-Means algorithm divides the data set into k clusters, minimizing the sum of squares within the clusters, ensuring that countries in the same cluster show similar patterns of behaviour. Given the small number of countries, k-means was supplemented with hierarchical clustering to validate the robustness of the results. Also, several validation techniques were applied to the clustering results, including silhouette analysis, cross-validation with hierarchical clustering, and network graph consistency.

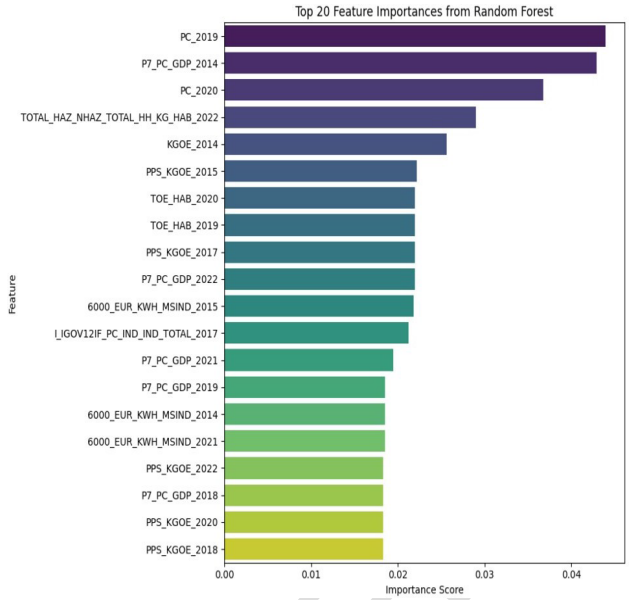


Figure 1: Random forest of the 20 best characteristics

Moreover, the initial focus was on identifying three clusters, representing the groups with low, medium and high risk/priority. The choice on three clusters was motivated out of necessity, to catch nuanced differences between countries, going beyond a simple binary classification (high risk vs. low risk).

Interpreting the numerical labels of the clusters was essential to turn the clustering results into useful insights. Each cluster was analysed to map the numerical labels into descriptive categories – "Low risk/High priority", "High risk/Low priority," and "Medium risk/Medium priority". For example, the countries identified in the cluster "Low risk/ High priority" showed strong performance in sustainable practices and awareness by consumers but may require targeted interventions to maintain or improve their sustainability initiatives and vice versa. It is important to emphasize that these clusters do not reflect the real world, but are groupings based on the similarity of the data.

3. RESULTS AND FINDINGS

The results represent an analysis of consumer behaviour towards eco-friendly products in the Balkans, using multiple clustering methodologies to categorize the seven countries of interest. The results of the clustering obtained from hierarchical clustering, k-means clustering, KNN and network graphs analysis reveal different patterns and groupings.

Hierarchical clustering. By expanding the analysis to three clusters, a more nuanced categorization emerged: 1) Low risk/High priority: Kosovo (XK), Bosnia and Herzegovina (BA), North Macedonia (MK), Montenegro (ME) and Serbia (RS) were grouped together, what indicates relatively better sustainability practices and greater prioritization of eco-friendly products in these countries. This suggests a collective potential for improvement and investment in sustainability initiatives; 2) Medium risk/Medium priority: Croatia (HR) was redistributed in this cluster, reflecting its intermediate status, despite EU membership, perhaps due to ongoing transitions or areas requiring further improvements; 3) High risk/Low priority: Albania (AL) is alone in this cluster, highlighting significant challenges in the adoption on eco-friendly practices and consumer behaviours compared to its regional peers.

DRAFT VERSION

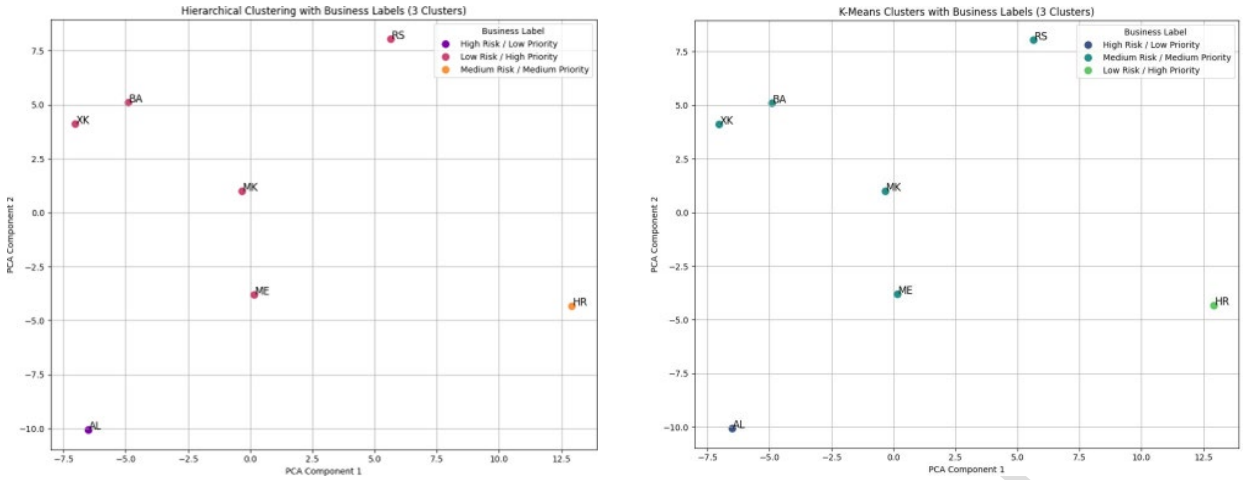


Figure 2: Hierarchical clustering (left) and k-means clustering (right) results

K-means clustering. K-Means clustering was applied for additional validation, with initial focus on three clusters to capture different patterns of consumer behaviour: 1) High risk/Low priority: Albania (AL) constantly remained in this cluster across different methodologies; 2) Low risk/High priority: Croatia (HR) was recategorized into this cluster, indicating that k-means may be more sensitive to certain feature distributions, highlighting Croatia's relative stronger performance in sustainability indicators; 3) Medium risk/Medium priority: the remaining countries – Bosnia and Herzegovina (BA), North Macedonia (MK), Montenegro (ME), Serbia (RS), and Kosovo (XK) – were grouped in this category, indicating on moderate levels on risk and prioritizing the adoption of eco-friendly consumer behaviours. The divergence between the results highlights the impact of algorithmic differences on cluster formation, especially in small samples where each data point significantly affects the outcome on clustering (Jain, 2010).

KNN analysis. KNN Analysis identified the closest neighbours, visually representing the relations and closeness between Balkan countries based on their consumer behaviour towards eco-friendly products (Figure 3).

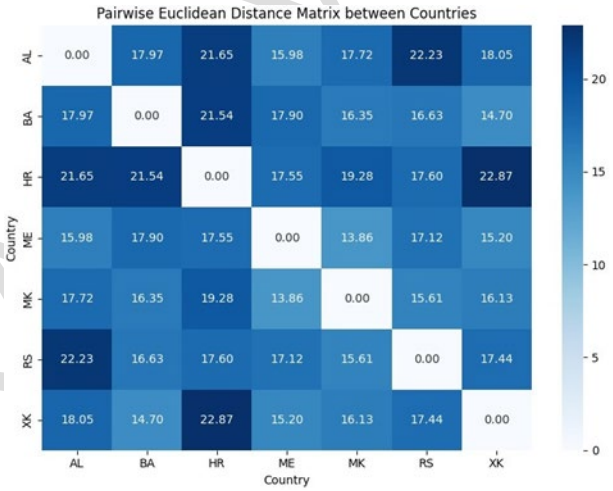


Figure 3: Matrix on Euclidean distance between neighbours using KNN analysis

The network graph built based on the KNN results reveals the following key insights. Regarding the connections between nodes, Montenegro stands out as a central hub, connecting with Albania, Croatia, and North Macedonia, indicating that Montenegro shares similar consumer behaviour patterns with these countries, acting as a bridge within the network. Kosovo and Bosnia and Herzegovina form a closely related pair, which suggests strong similarities in their sustainability practices and consumer preferences. Serbia remains isolated, indicating different consumer behaviour that does not match close with none another country in the dataset.

In terms of self-connections, Montenegro, Serbia, and Kosovo show self-connections, where the closest neighbour are themselves with insignificant distances. This implies that these countries have unique profiles that differ significantly from their neighbours. This suggests that they may be looking for specialized

DRAFT VERSION

strategies for sustainability which will consider their unique patterns of consumer behaviour and socio-economic contexts.

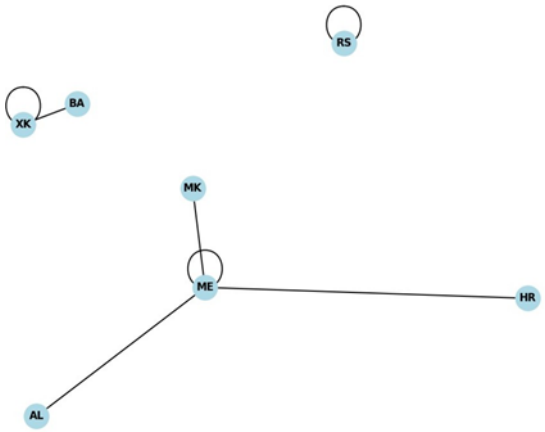


Figure 4: KNN Network Graph

The network graph effectively groups countries into three groups: 1) Group 1: Kosovo and Bosnia and Herzegovina, 2) Group 2: Montenegro, Croatia, North Macedonia, and Albania, reflecting diverse patterns on behaviour, but with basic similarities enabled by the central role in Montenegro, and 3) Group 3: Serbia, standing alone, emphasizing its own approach to sustainability and consumer preference.

4. CONCLUSION

By clustering 141 Eurostat time-series indicators, this study mapped sustainable-consumption patterns across the seven Western Balkan countries, consistently flagging Albania as high-risk/low-priority country, positioning Croatia as a shifting medium-risk case influenced by EU integration, highlighting Montenegro as a regional bridge, and revealing Serbia's distinct consumer profile. The convergence of k-means, hierarchical, and KNN network analyses strengthens confidence in these insights, yet the narrow country sample and reliance on quantitative data limit generalizability and overlook qualitative motivations. Policymakers and businesses should tailor interventions, such as improving financing, infrastructure, and awareness campaigns, to each country's circumstances while aligning with Sustainable Development Goals, thereby fostering broader adoption of eco-friendly products and laying ground for expanded, mixed-method research on sustainable behaviour in diverse socio-economic contexts.

LITERATURE

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651-666.

Licastro, A., & Sergi, B. S. (2021). Drivers and barriers to a green economy. A review of selected balkan countries. *Cleaner Engineering and Technology*, *4*, 100228.

Lukman, R. K., Glavič, P., Carpenter, A., & Virtič, P. (2016). Sustainable consumption and production–Research, experience, and development–The Europe we want. *Journal of cleaner production*, *138*, 139-147.

Marić, R., & Vukmirović, G. (2018). Sustainability in International Food Retailing: The Case Of Western Balkans. *Trade Perspectives*, *59*.

Nikolić, T. M., Paunović, I., Milovanović, M., Lozović, N., & Đurović, M. (2022). Examining Generation Z's attitudes, behavior and awareness regarding eco-products: A Bayesian approach to confirmatory factor analysis. *Sustainability*, *14*(5), 2727.

Sachs, J., Kroll, C., Lafortune, G., Fuller, G., & Woelm, F. (2021). *Sustainable development report 2021*. Cambridge University Press.

Thies, C., Kieckhäfer, K., Spengler, T. S., & Sodhi, M. S. (2019). Operations research for sustainability assessment of products: A review. *European Journal of Operational Research*, *274*(1), 1-21.

Verain, M. C., Dagevos, H., & Antonides, G. (2015). Sustainable food consumption. Product choice or curtailment?. *Appetite*, *91*, 375-384.

DRAFT VERSION

Wang, C., Ghadimi, P., Lim, M. K., & Tseng, M. L. (2019). A literature review of sustainable consumption and production: A comparative analysis in developed and developing economies. *Journal of cleaner production*, *206*, 741-754.

White, L., & Lee, G. J. (2009). Operational research and sustainable development: Tackling the social dimension. *European Journal of Operational Research*, *193*(3), 683-692.

DRAFT VERSION

E-BUSINESS

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

THE ROLE OF S-COMMERCE IN HIGHER EDUCATION: ANALYZING THE PRESENCE OF SERBIAN FACULTIES ON TIKTOK

SARA KOSTIĆ[1], JELENA SPAJIĆ[2], BOJANA MILIĆ[3], DUNJA BOŠKOVIĆ[4], DANIJELA LALIĆ[5]

[1] Novi Sad School of Business, Novi Sad, sara97kostic@gmail.com, 0000-0002-5079-1096
[2] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, stankovicj@uns.ac.rs, 0000-0003-2890-969X
[3] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, bojana.milic@uns.ac.rs. 0000-0001-6327-6424
[4] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, dunja.vujicic@uns.ac.rs, 0000-0003-3313-661X
[5] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, danijelalalic@uns.ac.rs, 0000-0002-1277-5116

Abstract: *TikTok, as a key tool of S-Commerce, offers new opportunities for promotion and interaction in higher education. An analysis of the presence of state faculties in Serbia on TikTok reveals that only 18% of the 89 faculties have official accounts, with the University of Belgrade leading with nine active profiles. The results highlight a significant untapped potential of this platform. A strategic approach and exchange of best practices are proposed to enhance digital presence, improve visibility, and foster deeper connections with younger generations..*

Keywords: *S-Commerce, TikTok, higher education, social networks, promotion.*

DRAFT VERSION

1. INTRODUCTION

The integration of social media into commerce, commonly referred to as social commerce (*S-Commerce*), has become increasingly relevant in higher education marketing. Platforms such as TikTok offer new opportunities for universities to enhance visibility, engage prospective students and build meaningful connections with digital native audiences.

S-Commerce merges social interaction with commercial activity, enabling institutions to communicate directly with users and foster engagement through content-driven strategies (Hajli, 2015). TikTok in particular, has rapidly emerged as a high-engagement platform due to its short-form video format and viral potential (Influencer Marketing Hub, 2024). While originally perceived as an entertainment-focused app, it is now being adopted by various organizations, including educational institutions, as a tool to reach younger demographics.

Despite the widespread use of social media in higher education marketing, TikTok remains a relatively underexplored channel in academic research (Wang & Herrando, 2019). This paper aims to provide a focused literature review on the application of S-Commerce in promoting higher education institutions (HEIs), followed by an analysis of TikTok presence among faculties within Serbian universities, examining its potential through the lens of S-Commerce principles.

2. S-COMMERCE: THEORETICAL BACKGROUND

The concept of social commerce (*S-commerce*) represents a distinct form of electronic commerce that incorporates elements of social interaction into the purchasing process. Rather than serving as passive content consumers, users of social media platforms actively shape demand by commenting, sharing, and recommending products (Kaplan & Haenlein, 2010). At its core, S-Commerce leverages social networks as transactional environments, enabling dynamic, two-way communication between brands and consumers.

According to Zhang and Benyoucef (2016), S-Commerce integrates social components such as peer recommendations and content sharing with traditional e-commerce functions to deliver a more engaging and personalized user experience. It is distinguished from conventional e-commerce by three key dimensions: social interaction, commercial intent and user engagement. Social interaction forms the foundation of S-Commerce, fostering trust through user-generated reviews and social proof (Wang & Herrando, 2019). Hajli et al. (2017) also investigate the relationship between trust and purchase intentions in social networking sites, discovering that trust plays a substantial role in consumers' decisions to engage in S-commerce transactions.

Shin et al. (2020) delve into service quality management in S-commerce, analyzing how service design qualities affect consumer satisfaction and decision-making processes. Their research found that hedonic value surpasses utilitarian value in impacting user satisfaction and responsiveness towards social commerce service offerings. Hajli et al. (2017) also emphasize the significant role of social technologies in shaping interactions among consumers, businesses and organizations in the S-commerce landscape.

Platforms like TikTok and Instagram facilitate viral dissemination of product-related content through comments and shares (Huang & Benyoucef, 2013). „In 2025, sales through social networks accounted for an estimated 17.11 percent of total online sales" (Statista, 2025). Live commerce has grown in popularity in recent years. Its prevalence is only expected to increase as companies utilize livestreaming technologies more and more for promotional and marketing purposes. Digital shoppers benefit from live commerce because it offers attractive discounts as well as inspiration and ideas. In 2022, Facebook was the leading social network platform where internet users purchased products during live streaming events. As with social commerce, Asian countries have paved the way for this highly interactive shopping experience. In 2023, over seven in ten consumers engaged in live shopping in China, India, Thailand, and United Arab Emirates (Statista, 2025).

This comprehensive perspective underlines the collaborative essence of S-commerce, suggesting that successful strategies must integrate social and technological dimensions effectively.

3. THE ROLE OF S-COMMERCE IN HIGHER EDUCATION PROMOTION

S-Commerce is gaining prominence in the education sector, where it serves as a strategic tool for institutional promotion, student recruitment and audience engagement. By merging educational content with social media interactivity, universities can significantly enhance their visibility among younger demographics. The key applications of S-Commerce in higher education include:

- **Program Promotion**: Platforms such as TikTok, Instagram and Facebook allow institutions to creatively showcase academic programs, student life and campus activities. Short videos, interactive posts and visual storytelling foster familiarity and interest. TikTok, in particular, facilitates viral dissemination through lecture snippets, behind-the-scenes content and educational challenges (Influencer Marketing Hub, 2024).
- **Student Engagement**: S-Commerce enables real-time interaction with current and prospective students through comments, polls and live streams. Institutions utilize Instagram Live for Q&A sessions on admissions, while TikTok campaigns featuring student content creators enhance authenticity and deepen engagement (Huang & Benyoucef, 2013).
- **Educational Resource Sharing**: Universities increasingly use social platforms to distribute tutorials, academic tips and research highlights, contributing to educational accessibility and community outreach (Zhang & Benyoucef, 2016). This approach not only attracts students but also reframes the institutional role in public knowledge dissemination.
- **Brand Building**: Social media supports institutional branding by combining informative and entertaining content. Participation in viral trends allows universities to align with contemporary discourse and strengthen public perception (Stephen & Toubia, 2010).

As SproutSocial (2023) notes, "thanks to social media, every aspect of the college experience can be shared," offering a level of transparency and interaction that traditional media cannot match. Recent data confirms the strategic value of social media in education. Around 68% of high school students use social platforms to research universities, underscoring the shift toward digital-first exploration (SproutSocial, 2023). Moreover, 80% of alumni organizations identify social media and digital content as key tools for maintaining alumni relations. Importantly, 41% of educational institutions report a direct link between their social media strategy and increased enrollment, highlighting its role as a core, not peripheral, element of modern marketing (SproutSocial, 2023).

Social Media Strategies Summit Blog (2024) identifies effective uses of TikTok in higher education marketing. The article "19 Great Examples of TikTok Content for Higher Education" outlines best practices:

- *Humanizing Leadership*: Harvard College features Dean Rakesh Khurana answering personal questions to create a more approachable image.
- *Trend Engagement*: Washington University in St. Louis leverages trending memes and challenges to connect with a wider audience.
- *Tradition Showcasing*: The University of Alabama shares viral time-lapse videos of students forming the university logo, reinforcing community and tradition.

DRAFT VERSION

- *Humor and Relatability*: The University of Michigan uses humor to align with TikTok's entertainment-centric culture and appeal to Gen Z.

These strategies illustrate how TikTok can serve as a cost-effective, high-engagement tool for fostering authentic communication, strengthening institutional identity and reaching a digital-native audience.

A study by Escamilla-Fajardo et al. (2021) even explored the pedagogical value of TikTok in higher education by examining its integration into a corporal expression sport course. The research demonstrated that TikTok, when used as a learning tool, significantly enhanced student engagement, motivation and creativity. Students were not only more actively involved in the learning process but also developed essential digital competencies such as video editing and content creation.

The alignment of video content with promotional strategies is particularly relevant to platforms like TikTok, which have redefined social media interactions through short-form video content designed for virality and engagement (G.U. & F.O., 2023). According to Marešová et al., social media provides extensive feedback mechanisms that enable universities to tailor their marketing efforts effectively based on analytics gathered from user interactions (Marešová et al., 2020).

Furthermore, the literature suggests that universities can benefit from employing influencer marketing strategies on platforms such as TikTok. Successful collaborations between influencers and educational brands can yield better engagement metrics, as followers tend to trust influencer recommendations more than traditional advertisements (Ibáñez-Sánchez et al., 2021). Influencers can amplify the reach and authenticity of educational content, making it more relatable and appealing to prospective students.

Importantly, the impact of TikTok is also reflected in the motivational effects on students. Research has shown that engagement with TikTok can influence students' academic performance and learning motivation, with students perceiving social media presence as beneficial to their learning processes (Hill et al., 2024). Thus, as higher education institutions craft their marketing strategies, they must account for the multifaceted influences of TikTok, balancing promotional content with authentic engagement that resonates with students' realities and aspirations as they navigate their educational journeys (Qin et al., 2022).

4. THE ANALYSIS OF THE PRESENCE OF SERBIAN FACULTIES ON TIKTOK

According to the Ministry of Education of the Republic of Serbia, there are nine state universities, including the University of Belgrade, University of Novi Sad, University of Niš, University of Kragujevac, University of Arts in Belgrade, University of Defence, State University of Novi Pazar, University of Priština (temporarily located in Kosovska Mitrovica), and the University of Criminal Investigation and Police Studies (Ministarstvo prosvete, 2024). Additionally, Serbia has ten private universities (Edukacija, 2024). Given the differences in organization and funding models, this study focuses solely on state universities.

The analysis included all nine Serbian universities, specifically their constituent faculties. The research process was conducted during the December 2024, and involved the following steps:

- **TikTok Search**: Keywords included full faculty names and common abbreviations (e.g., "Fakultet tehničkih nauka Novi Sad" for the Faculty of Technical Sciences and "Ekof", etc.) to identify official TikTok profiles.
- **Account Verification**: Identified accounts were examined to determine their official status by reviewing usernames, content and links to institutional websites.
- **Categorization**: Faculties were grouped into three categories: (1) those with official TikTok accounts, (2) those without any presence, and (3) those represented by student-generated informal content. Only faculties with verified official accounts were included in the analysis.
- **Content Analysis**: Official accounts, such as the Faculty of Fine Arts in Belgrade (@flubeograd), were assessed based on the type of content they published: informational, promotional or educational.

The University of Belgrade, the largest state university in Serbia, comprises 31 faculties, organized into four academic fields: social sciences and humanities, medical sciences, natural and mathematical sciences, and technical and technological sciences (University of Belgrade, 2024).

Table 1 presents the faculties of the University of Belgrade that maintain official TikTok accounts.

Table 1: Faculties of the University of Belgrade with Official TikTok Presence

The name of HEI	TikTok name	Content type
Faculty of Economics (Ekonomski fakultet)	@ekofbg	Informational, educational, promotional

DRAFT VERSION

Faculty of Political Sciences (Fakultet političkih nauka)	@fpn_beograd	Informational
Faculty of Special Education and Rehabilitation (Fakultet za specijalnu edukaciju i rehabilitaciju)	@parlamentfasper	Informational, educational
Faculty of Philosophy (Filozofski fakultet)	@filozofski_fakultet_bg	Informational
Faculty of Physics (Fizički fakultet)	@fizicki_fakultet	Educational
Faculty of Civil Engineering (Građevinski fakultet)	@gradjevniskifakultet	Informational
Faculty of Chemistry (Hemijski fakultet)	@hemijskifakultet	No content on the account
Faculty of Transport and Traffic Engineering (Saobraćajni fakultet)	@saobracajni_fakultet	Informational, educational, entertaining
Faculty of Forestry (Šumarski fakultet)	@sumarskifakultetbeograd	Informational, educational

The University of Novi Sad, founded in 1960, is one of the leading higher education institutions in Serbia and the region. The university consists of 14 faculties, which are located in four cities: Novi Sad, Subotica, Zrenjanin and Sombor. The faculties cover a wide range of disciplines, from natural and technical sciences to social and artistic fields, providing education to more than 50,000 students. (University of Novi Sad, 2024).

Table 2: Faculties of the University of Novi Sad with Official TikTok Presence

The name of HEI	The TikTok name	Content type
Faculty of Agriculture (Poljoprivredni fakultet, Novi Sad)	@poljo.ns	Informational, educational
Faculty of Technical Sciences (Fakultet tehničkih nauka, Novi Sad)	@ftn_ns	Informational, educational, entertaining
Faculty of Civil Engineering (Građevinski fakultet, Subotica)	@gfsubotica	Informational
Faculty of Education (Pedagoški fakultet, Sombor)	@pedagoski_fakultet_so	Informational, entertaining

The University of Niš was founded in 1965 and is one of the key higher education institutions in southeastern Serbia. The University stands out for its contribution to education, science and the development of the local community, and it has a total of 14 faculties. Faculties cover a wide range of scientific disciplines, including technical, natural-mathematical, social-humanistic, medical and artistic fields (Univerzitet u Nišu, 2024).

Table 3: Faculties of the University of Niš with Official TikTok Presence

The name of HEI	The TikTok name	Content type
Faculty of Techology (Tehnološki fakultet, Leskovac)	@tehnoloski.fakultet	Informational
Faculty of Education (Pedagoški fakultet, Vranje)	@pedagoski.fakultet	Informational

The University of Kragujevac, a state HEI established on May 21, 1976, traces its roots to the Lyceum of the Principality of Serbia founded in 1838, the first institution of higher education in modern Serbia (University of Kragujevac, 2024). Today, it comprises 12 faculties across six cities in central Serbia. A keyword-based search on TikTok revealed that none of these faculties maintain an official presence on the

DRAFT VERSION

platform. While individual users and students occasionally share related content, there are no verified institutional accounts.

The University of Arts in Belgrade, founded in 1957, is a key pillar of higher education in the arts in Serbia. It consists of four faculties covering music, visual and applied arts, and dramatic arts, including theater, film and television (University of Arts in Belgrade, 2024). Among its faculties, only the Faculty of Fine Arts (*Fakultet likovih umetnosti*) maintains an official TikTok account (@flubeograd), sharing informative content related to academic activities and programs.

In contrast, the research found no official TikTok accounts for the University of Defence in Belgrade, the State University of Novi Pazar, the University of Priština (temporarily based in Kosovska Mitrovica), or the University of Criminal Investigation and Police Studies.

5. DISCUSSION AND CONCLUSION

The emergence of S-commerce in higher education marks a strategic shift toward more interactive, engaging and student-centered communication. TikTok, as one of the most popular platforms among younger audiences, offers a unique opportunity for universities to promote their programs, interact with prospective students, and build recognizable digital brands.

Recent studies (Marešová et al., 2020; Ibáñez-Sánchez et al., 2021; G.U. & F.O., 2023) highlight the multidimensional role of trust, user behavior, service quality and social interaction in shaping the effectiveness of S-commerce. These insights call for a deeper understanding of user dynamics on social platforms, not only in business but increasingly in educational contexts.

This study found that only 16 out of 89 public faculties in Serbia (approximately 18%) have official TikTok accounts. The University of Belgrade leads in digital presence, with nine faculties actively using the platform. These faculties, such as the Faculty of Economics, Faculty of Transport and Traffic Engineering and Faculty of Forestry, demonstrate a creative, strategic approach by producing diverse content types, ranging from informative and educational to promotional and entertaining.

In contrast, faculties from the Universities of Kragujevac, Novi Pazar and Priština have no official presence on TikTok, indicating a missed opportunity to reach and engage younger audiences. A similar situation was observed at the University of Arts in Belgrade, where only the Faculty of Fine Arts maintains a verified account. This isolated case of good practice highlights the untapped potential within other faculties at the same institution.

Faculties that are active on TikTok increase their visibility, enhance their digital footprint and attract new generations of students. Those that remain absent risk reduced competitiveness in the increasingly digital landscape of higher education.

These findings underscore the need for modernizing communication strategies in academia through the lens of S-Commerce. Based on the research results, the following recommendations are proposed:

- **Strategic content development**: Institutions should adopt tailored strategies that align with the logic of S-Commerce: short videos, interactive formats, educational challenges and engaging content suited to younger audiences.
- **Encouraging digital onboarding**: Faculties currently absent from TikTok should be supported in launching official accounts and building content that blends educational and entertaining elements.
- **Knowledge sharing**: Successful examples, such as the Faculty of Economics, University of Belgrade, and the Faculty of Technical Sciences, University of Novi Sad, should be used as benchmarks. Sharing best practices across institutions may accelerate digital transformation and effective use of social platforms.

By integrating traditional academic values with modern digital tools and S-Commerce principles, universities can significantly enhance their visibility, attract prospective students and strengthen connections with current ones. In summary, TikTok's integration into higher education offers more than just a communication tool, it fosters community, motivation and academic interest through culturally relevant content. This positions it as a powerful medium for educational marketing in the digital age.

DRAFT VERSION

ACKNOWLEDGMENT

The results presented in this paper are part of the research within the project "Improving the quality of teaching in the Department's study programs through the implementation of the results of scientific research in the field of Industrial Engineering and Management", Department of Industrial Engineering and Management, Faculty of Technical Sciences in Novi Sad, University of Novi Sad, Republic of Serbia.

LITERATURE

Edukacija. https://fakulteti.edukacija.rs/univerziteti, (Accessed on: December 22, 2024)

Escamilla-Fajardo, P., Alguacil, M., & López-Carril, S. (2021). Incorporating TikTok in higher education: Pedagogical perspectives from a corporal expression sport sciences course. *Journal of Hospitality, Leisure, Sport & Tourism Education, 28*, 100302. https://doi.org/10.1016/j.jhlste.2021.100302

G.U., N. and F.O., N. (2023). Uses and gratifications of tiktok platform among university undergraduates. *African Journal of Social Sciences and Humanities Research, 6(6),* 139-153. https://doi.org/10.52589/ajsshr-66fabncr

Hajli, N. (2015). Social commerce constructs and consumer's intention to buy. *International Journal of Information Management, 35(2),* 183-191.

Hajli, N., Sims, J., Zadeh, A., & Richard, M. (2017). A social commerce investigation of the role of trust in a social networking site on purchase intentions. *Journal of Business Research, 71,* 133-141. https://doi.org/10.1016/j.jbusres.2016.10.004

Hill, K., Xie, J., Gallo, K., Wood, S., Parlow, M., Hynes, J., & Stewart, S. (2024). The role of a major social media platform on students' academic performance: perception versus reality. *European Journal of Interactive Multimedia and Education, 5(1)*, https://doi.org/10.30935/ejimed/14135

Huang, Z., & Benyoucef, M. (2013). From e-commerce to social commerce: A close look at design features. *Electronic Commerce Research and Applications*, *12(4),* 246-259.

Ibáñez-Sánchez, S., Flavián, M., Casaló, L., & Belanche, D. (2021). Influencers and brands successful collaborations: a mutual reinforcement to promote products and services on social media. *Journal of Marketing Communications, 28(5),* 469-486. https://doi.org/10.1080/13527266.2021.1929410

Influencer Marketing Hub. (2024). TikTok Marketing Statistics: The Ultimate Marketer's Guide. https://influencermarketinghub.com (Accessed on: December 1, 2024)

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons, 53(1),* 59-68.

Marešová, P., Hruska, J., & Kuca, K. (2020). Social Media University Branding. *Education Sciences*, *10(3),* 74. https://doi.org/10.3390/educsci10030074

Ministarstvo prosvete (2024). https://prosveta.gov.rs/prosveta/visoko-obrazovanje/, (Accessed on: December, 5, 2024.)

Qin, Y., Omar, B., & Musetti, A. (2022). The addiction behavior of short-form video app TikTok: The information quality and system quality perspective. *Frontiers in Psychology, 13.* https://doi.org/10.3389/fpsyg.2022.932805

Shin, N., Park, S., & Kim, H. (2020). Consumer satisfaction–based social commerce service quality management. *BRQ Business Research Quarterly, 24(1),* 34-52. https://doi.org/10.1177/2340944420916098

Social Media Strategie Summit (2024). https://blog.socialmediastrategiessummit.com/10-great-examples-of-tiktok-content-for-higher-education/ Accessed on December 25, 2024.

Sproutsocial (2023). Higher Education Social Media Benchmarks for 2023. URL: https://sproutsocial.com/insights/data/higher-education-social-media-benchmarks/, (Accessed on December 25, 2024.)

Statista (2025). Social commerce as percentage of total e-commerce worldwide from 2018 to 2029. URL: https://www.statista.com/statistics/1251145/social-commerce-share-worldwide/#statisticContainer, Accessed on June 6, 2025.

Stephen, A. T., & Toubia, O. (2010). Deriving value from social commerce networks. *Journal of Marketing Research*, *47(2),* 215-228.

Wang, Y., & Herrando, C. (2019). Does engagement with social media influence consumer purchasing behavior? An empirical study of S-Commerce platforms. *Computers in Human Behavior, 101,* 238-245.

Zhang, L., & Benyoucef, M. (2016). Consumer behavior in social commerce: A literature review. *Decision Support Systems*, 86, 95-108.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

Review article

SMART RAILWAY AS A SMART CITY SERVICE

ПАМЕТНА ЖЕЛЕЗНИЦА КАО СЕРВИС ПАМЕТНОГ ГРАДА

DANIJELA STOJANOVIĆ[1], NENAD STANISAVLJEVIĆ[2,3], PETAR LUKOVAC[3]

[1] Institute of Economic Sciences, Belgrade, danijela.stojanović@ien.bg.ac.rs, 0000-0002-9728-8586
[2] Infrastructure of Serbian Railways, Belgrade, nenad.stanisavljevic@srbrail.rs, 0000-0003-1723-7103
[3] University of Belgrade, Faculty of Organizational Sciences, Belgrade, petar.lukovac@fon.bg.ac.rs, 0000-0003-4561-8886

Rezime: *Železnički sistem predstavlja tehnički, saobraćajni i poslovni podsistem velikih razmera, koji je po svojoj prirodi kompleksan i multidisciplinaran. Kao energetski efikasan i ekološki prihvatljiv vid transporta, železnica beleži najniži nivo emisije gasova sa efektom staklene bašte u poređenju sa drugim transportnim modalitetima. Razvoj savremenih informaciono-komunikacionih tehnologija (IKT), Interneta inteligentnih uređaja (IoT) i veštačke inteligencije (AI) inicirao je temeljne promene i u železničkom sektoru, u pravcu digitalne transformacije i modernizacije sistema. U radu se analizira koncept „pametne železnice" kao ključnog infrastrukturnog segmenta pametnih gradova, kroz prikaz strukture, funkcija i primera implementacije u urbanim sredinama. Poseban akcenat stavljen je na uporednu analizu globalnih praksi i identifikaciju razvojnih prioriteta. Cilj istraživanja je da se proceni strateški značaj i komparativne prednosti pametnih železnica u kontekstu razvoja pametnih gradova. Zaključuje se da dobro planirana i integrisana pametna železnica može značajno doprineti održivom urbanom razvoju i unapređenju kvaliteta života građana.*

Ključne reči: *pametna železnica, pametni grad, informaciono-komunikacione tehnologije* (*IKT*), *internet inteligentnih uređaja (IoT), održivi transport, urbana mobilnost*

Abstract: *The railway system is a large-scale, complex, and multidisciplinary technical, transportation, and business subsystem. As an energy-efficient and environmentally friendly mode of transport, railways record the lowest levels of greenhouse gas emissions compared to other transport modalities. The development of modern information and communication technologies (ICT), the Internet of Things (IoT), and artificial intelligence (AI) has triggered substantial changes in the railway sector, steering it toward digital transformation and system modernization. This paper analyzes the concept of the "smart railway" as a key infrastructural component of smart cities by examining its structure, core functions, and selected urban implementation cases. Special emphasis is placed on comparative analysis of global practices and the identification of development priorities. The aim of the research is to assess the strategic relevance and comparative advantages of smart railways within the broader smart city ecosystem. The paper concludes that a well-planned and integrated smart railway system can significantly contribute to sustainable urban development and enhance citizens' quality of life.*

Keywords: *smart railway, smart city, information and communication technologies (ICT), Internet of Things (IoT), sustainable transport, urban mobility*

DRAFT VERSION

1. INTRODUCTION

Contemporary trends in the development of information technologies and digitalization have fundamentally transformed social systems and numerous industrial sectors, including railway transport (Bogdanović et al., 2021; Stojanović & Domazet, 2020). The emergence of the Industry 4.0 paradigm has enabled the integration of knowledge and resources from various sectors—public administration, business, education, civil society, and individual actors—with the aim of joint participation in innovation processes and the improvement of complex systems (Stojanović et al., 2021).

Within this context, the railway sector gains particular significance as a complex, multidisciplinary technical, technological, transport, and business system. In modern conditions, thanks to the development of the Internet of Things (IoT), mobile technologies, blockchain, and other digital solutions, the prerequisites have been created for the railway system to respond more efficiently to challenges that exceed the capabilities of traditional, often costly and time-consuming, technical approaches (Kupriyanovsky et al., 2020; Stanisavljević, Stojanović, & Bogdanović, 2023; Thurner & Gershman, 2014).

Through the concept of open innovation and the application of crowdsourcing, it has become possible to include various stakeholders—from citizens and passengers, to the academic community and industry experts—in processes of identifying problems, proposing solutions, and co-creating innovations in the railway sector (Radenković et al., 2020; Stanisavljević et al., 2022). This contributes to greater flexibility, efficiency, and rationality in the management and technical operation of railways.

The modernization and optimization of the railway system represent key factors in achieving a sustainable, safe, and competitive transport environment. According to data from the International Union of Railways (UIC), railway transport is highly energy-efficient and environmentally friendly, emitting significantly lower levels of harmful gases compared to other modes of transport.

In this light, the implementation of advanced technologies and the development of intelligent railway management systems become essential for building an integrated and smart railway ecosystem that ensures greater safety, reliability, and sustainability, while simultaneously enhancing the overall user experience (Sharma & Kumar, 2017).

2. SMART RAILWAY

The exponential growth in demand for Internet of Things (IoT) technologies in recent years has driven significant advances in sensors, radio access, network protocols, and hardware/software platforms designed for large-scale applications (Jo et al., 2018). IoT has found wide application across various industries, and particularly in the railway sector, where it is used to enhance operations management, maintenance, video surveillance, and safety at level crossings (Singh et al., 2022).

Modern high-resolution sensors, such as high-speed cameras and LiDAR systems, enable real-time monitoring of railway infrastructure and rolling stock. These technologies support early detection of technical faults and potential risks, allowing for predictive maintenance and reduced operational costs. At the same time, data analytics powered by artificial intelligence and machine learning allows for the optimization of train scheduling and operational efficiency (Wang et al., 2022). Automated control systems, including advanced train control mechanisms, further improve safety by preventing collisions and derailments.

In the context of digital transformation, a railway system that integrates advanced information and communication technologies, sensor networks, IoT devices, wireless communication, real-time data processing, and AI algorithms is referred to as a Smart Railway (Ai et al., 2020; Z. Yu et al., 2023). The main objective of such a system is to enhance the management of trains, infrastructure, and logistics while simultaneously improving energy efficiency, safety, and operational reliability.

A smart railway is composed of the following interconnected elements: smart trains and intelligent control systems, IoT sensors embedded in tracks and stations, integrated control centers, digitized railway stations, automated signaling and traffic lights, video surveillance and safety systems, visual interfaces for citizen interaction, sustainable energy infrastructure, interoperability, and system connectivity.

Given the continued evolution of digital technologies, railway transport is expected to become even more intelligent, flexible, and integrated within smart cities and multimodal transport networks. Therefore, research and development in the field of smart railways is a strategic priority for achieving sustainable mobility, safety, and efficient transportation in the near future (Da Fonseca-Soares et al., 2023).

3. SMART CITY

Thanks to information and communication technologies, a smart city improves operational efficiency, shares information with the public, and ensures better quality of urban services and citizens' well-being. The goal of a smart city is to enhance the quality of life through the use of smart technologies and data analytics, while simultaneously optimizing its functions and ensuring economic growth.

Smart cities regulate and manage, among other things, the entire transportation infrastructure, including the railway system, to provide the highest quality services using minimal resources, thus meeting set goals (Sabory et al., 2021). Initiatives for using smart technologies are launched jointly by local governments and citizens.

The concept of smart cities dates back to the 1960s and 1970s when the U.S. Bureau of Community Analysis began using databases, aerial photography, and cluster analysis to collect data, direct resources, and

DRAFT VERSION

issue reports to guide services, mitigate disasters, and prevent poverty. Today, we are witnessing the third generation of smart cities, characterized by community engagement and inclusion. For example, in Vienna, Austria, citizens can invest in local solar plants and take part in initiatives related to gender equality and housing accessibility. In Vancouver, Canada, over 30,000 residents participated in the creation of the "Greenest City Action Plan." Among today's global leaders in smart city development are Singapore, Kansas City, Barcelona, and Dubai.

One of the most important foundations of a smart city is smart transport. Smart cities and their industrial facilities are most efficiently supplied through road and railway transport (Miksa Hell & János Varga, 2018). Railways are the backbone of urban mobility, and in future smart cities, railway stations must offer easy access, multiple ticket booking and payment channels, and seamless integration with other modes of transport (Fantechi et al., 2021). The development of IoT has enabled the broader society to contribute to solutions across various domains, including railways. The application of open innovation facilitates the development of smart rail transport within smart cities (Stanisavljević, Stojanović, Miletić, et al., 2023).

4. SMART RAILWAYS AS A SERVICE

According to studies on smart city standards, transportation is one of the core services that defines a smart city. Smart railways, as the future of rail transport, relying on smart technologies and continuous innovation, will be the foundation of sustainable and safe mobility (Cruz et al., 2018; Fadel & Abu-Eisheh, 2024).

Cities are seeking to implement smart city strategies aimed at significantly improving the quality of life, particularly in terms of mobility and reducing CO_2 emissions. In this context, railways play a critical role. The location of stations and their accessibility to other transport modes are crucial for the use of rail. Therefore, the concept of a smart city must integrate rail transport (Bubelíny et al., 2021).

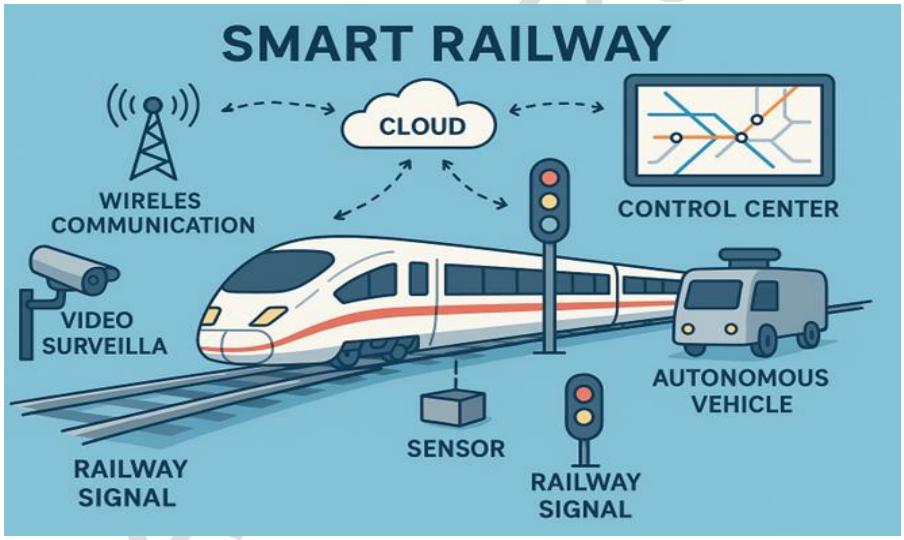


Figure 1: Smart Railway System Concept

Figure 1 presents the key components of a smart railway system that enable safe, efficient, and automated rail traffic management. The system is based on the integration of various technologies: video surveillance ensures continuous monitoring of infrastructure and surroundings; wireless communication enables real-time data exchange between vehicles, infrastructure, and the control center; sensors placed along the tracks detect field conditions and send data to the cloud, while railway signals and autonomous vehicles support automated control and timely responses to traffic situations. The control center, connected to all components, manages the entire system based on the analysis of collected data, relying on technologies such as IoT, AI, and cloud computing. This integration enhances the safety, efficiency, and sustainability of the railway system within the context of a smart city.

Efficient, long-term development of railway infrastructure is one of the key components in improving the performance of transport systems and quality of life, as demonstrated by the example of Prague. This case shows that not only advanced traffic control and signaling systems, but also efficient service concepts, can significantly enhance the railway system's efficiency and network capacity (Drábek & Michl, 2017).

Although public transport is central to urban mobility, current trends in Slovakia show a preference for individual transport. By addressing citizens' mobility needs through smart city concepts, local governments can promote greater use of public transport while improving air quality and overall quality of life (Bubelíny & Kubina, 2021).

DRAFT VERSION

Studies conducted in China have shown that smart city pilot projects positively impact regional economic growth through infrastructure and market effects. The rapid development of Chinese high-speed rail in the 21st century has reshaped the regional innovation landscape (X. Yu & Wan, 2022; Zhou et al., 2023).

In Sydney, the railway system was assessed as a component of the smart city based on indicators related to transportation and urban mobility across 12 municipalities within a 35 km radius of the city center. New rail infrastructure has provided significant technological benefits to suburbs, improving urban mobility (Gharehbaghi et al., 2023).

Mobility—both physical and economic—is a key factor in smart city development. The introduction of smart concepts into transport infrastructure will influence population growth and business demands. Railway stations that can provide strong economic development, environmental performance, and punctuality are recognized as core components of urban development (Shaltout, 2021).

Digital rail is the most important global trend in the development of rail transport. The "smart station" is a part of the digital railway concept. In Russia, this includes building new or reconstructing existing station complexes according to the smart station model (Nikonova et al., 2022).

Finnish Railways has defined train stations as strategic locations for implementing smart city concepts. These stations are becoming part of broader urban development efforts that align with economic and innovation policies, linking stations with smart city goals and open innovation frameworks (Hynynen & Kolehmainen, 2019).

Rail transit hubs play a critical role in modern urban development discourse as focal points of social, economic, and environmental interest. A railway station can effectively form a "city within a city," often referred to as a Railpolis. In India—a country facing complex urban challenges—such Railpolis models are expected to align with sustainable development goals and help cities become globally competitive (Banerjee, 2024).

Smart transportation is a key factor in smart city development, and its competitiveness depends on aligning management approaches with global best practices. Researchers have analyzed European and Russian approaches to smart transport governance to develop systematic standards (Bolgova et al., 2022).

However, studies of smart city concepts during the reconstruction of three railway stations in the Netherlands show that practical results are still insufficient. Although stakeholders were aware of the opportunities smart cities offer, knowledge sharing and innovation were hindered by limited acceptance and understanding, as well as institutional constraints such as data privacy regulations (de Wijs et al., 2016).

With the increasing demand for and adoption of smart city initiatives, it is essential to plan railway infrastructure adequately. If properly integrated, it can have a major positive impact. Given the complexity involved, systems thinking provides a useful framework for railway transport planning (Okonta et al., 2023).

5. CONCLUSION

The development of the smart railway concept is the result of dynamic transformations in the fields of information and communication technologies, the Internet of Things (IoT), and digital infrastructure. The modernization and implementation of advanced railway traffic management systems contribute to the enhanced safety, efficiency, reliability, and sustainability of the railway system, thereby improving the overall user experience. In this context, railways are positioned as an environmentally friendly and mobile mode of transport, with smart railways emerging as one of the foundational elements of the future smart city.

As demonstrated in this paper, the integration of smart railways into the smart city concept enables the achievement of key urban and societal goals—improved quality of life, increased mobility, economic growth, and optimal resource utilization. Railway stations, in particular, stand out as central infrastructural hubs within the broader frameworks of sustainable and digitally supported urban development.

However, the development and implementation of smart railways are accompanied by certain challenges: high initial implementation and maintenance costs, complex integration with existing systems, the need for a high level of cybersecurity and data protection, and potential resistance from certain demographic groups to technological innovations.

Despite these challenges, global trends in digital transformation and the widespread adoption of smart city initiatives indicate that the future of railways is inextricably linked with further digitalization and open innovation. In this regard, infrastructural and urban development plans for smart railways must be aligned with the strategic directions of smart city development. Only then can synergistic effects be achieved, enabling the full potential of the smart railway as a vital service of the smart city to be realized.

DRAFT VERSION

ACKNOWLEDGMENT

This research received no external funding.

LITERATURE

Ai, B., Molisch, A. F., Rupp, M., & Zhong, Z. D. (2020). 5G key technologies for smart railways. *Proceedings of the IEEE*, *108*(6), 856–893. https://doi.org/10.1109/JPROC.2020.2988595

Banerjee, I. (2024). Railopolis: City within A City. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, *17*(3), 443–463. https://doi.org/10.1080/17549175.2022.2074523

Bogdanović, Z., Stojanović, M., Radenković, M., Labus, A., & Despotović-Zrakić, M. (2021). Mobile Operator as the Aggregator in a Demand Response Model for Smart Residential Communities. *Lecture Notes on Data Engineering and Communications Technologies*, *79*, 58–67. https://doi.org/10.1007/978-3-030-79206-0_5

Bolgova, E., Koroleva, E., & Bolgov, S. (2022). Smart transport in a smart city: European and Russian development management track. *Transportation Research Procedia*, *63*, 844–852. https://doi.org/10.1016/J.TRPRO.2022.06.081

Bubelíny, O., & Kubina, M. (2021). Impact of the concept Smart City on public transport. *Transportation Research Procedia*, *55*, 1361–1367. https://doi.org/10.1016/J.TRPRO.2021.07.120

Bubelíny, O., Kubina, M., & Varmus, M. (2021). Railway Stations as Part of Mobility in the Smart City Concept. *Transportation Research Procedia*, *53*, 274–281.

Cruz, R., Jardim, J., Mira, J., & Teixeira, C. (2018). Smart rail for smart mobility. *Proceedings of 2018 16th International Conference on Intelligent Transport System Telecommunications, ITST 2018*. https://doi.org/10.1109/ITST.2018.8566842

Da Fonseca-Soares, D., Eliziário, S. A., Galvinicio, J. D., & Ramos-Ridao, A. F. (2023). Life-Cycle Greenhouse Gas (GHG) Emissions Calculation for Urban Rail Transit Systems: The Case of Pernambuco Metro. *Applied Sciences 2023, Vol. 13, Page 8965*, *13*(15), 8965. https://doi.org/10.3390/APP13158965

de Wijs, L., Witte, P., & Geertman, S. (2016). How smart is smart? Theoretical and empirical considerations on implementing smart city objectives – a case study of Dutch railway station areas. *Innovation: The European Journal of Social Science Research*, *29*(4), 422–439.

Drábek, M., & Michl, Z. (2017). Smart rail infrastructure planning for smart cities: A prague rail hub case study. *2017 Smart Cities Symposium Prague, SCSP 2017 - IEEE Proceedings*. https://doi.org/10.1109/SCSP.2017.7973866

Fadel, R., & Abu-Eisheh, S. (2024). Identification of Strategic Planning Factors to Achieve Smart Mobility for New Cities in Developing Countries Using CIB Method. *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems, ICETSIS 2024*, 1963–1967. https://doi.org/10.1109/ICETSIS61505.2024.10459709

Fantechi, A. , Gori, G. , Parri, J. , & Sampietro, S. (2021). STINGRAY Project: Smart Stations as hubs of infomobility services for Smart Cities. In *Proceedings of the 7th Cini Annual Conference on ICT for Smart Cities & Communities (I-CITIES 2021) Salerno*. CINI. https://flore.unifi.it/handle/2158/1257884

Gharehbaghi, K., McManus, K., Hurst, N., Robson, K., Pagliara, F., & Eves, C. (2023). Advanced rail transportation infrastructure as the basis of improved urban mobility: research into Sydney as a smart city. *Australian Planner*, *59*(2), 101–116. https://doi.org/10.1080/07293682.2023.2202867

Hynynen, A., & Kolehmainen, J. (2019). Smart Principles for Knowledge-based Urban Development: Case Finnish Railway Station Areas. *Architectural Research in Finland*, *3*(1), 11–30. https://doi.org/10.37457/ARF.121660

Jo, O., Kim, Y. K., & Kim, J. (2018). Internet of Things for Smart Railway: Feasibility and Applications. *IEEE Internet of Things Journal*, *5*(2), 482–490. https://doi.org/10.1109/JIOT.2017.2749401

Kupriyanovsky, V. , Voropaev, Y. , Pokusaev, O. , Klimov, A. , Dobrynin, A. , & Gapanovich, D. (2020). BIM Technologies for Tunnels Used in Subways, Railways, Highways, and Hyperloop-IFC-Driven Real-Time Systems and Disruptive Innovation. *International Journal of Open Information Technologies*, *8*(9), 70–93. http://www.injoit.org/index.php/j1/article/view/994

Miksa Hell, P. , & János Varga, P. (2018). Accurate radiofrequency identification tracking in smart city railways by using drones. *Interdisciplinary Description of Complex Systems: INDECS*, *16*(A), 333–341. https://doi.org/10.7906/indecs.16.3.5

DRAFT VERSION

Nikonova, Y. I., Игоревна, Н. Я., Proskuryakova, E. A., & Александровна, П. Е. (2022). Digitalization of railway station complexes. *Modern Transportation Systems and Technologies*, *8*(3), 31–44. https://doi.org/10.17816/transsyst20228331-44

Okonta, U., Hosseinian-Far, A., & Sarwar, D. (2023). A systemic approach to sustainable railway infrastructure planning: the case study of Milton Keynes. *Smart and Sustainable Built Environment*, *14*(2), 407–428. https://doi.org/10.1108/SASBE-03-2023-0057/FULL/PDF

Radenković, B., Despotović-Zrakić, M., Bogdanović, Z., Barać, D., Labus, A., & Naumović, T. (2020). A distributed IoT system for modelling dynamics in smart environments. *Proceedings - 2020 International Conference Engineering Technologies and Computer Science, EnT 2020*, 47–53. https://doi.org/10.1109/ENT48576.2020.00015

Sabory, N. R., Senjyu, T., Danish, M. S. S., Hosham, A., Noorzada, A., Amiri, A. S., & Muhammdi, Z. (2021). Applicable Smart City Strategies to Ensure Energy Efficiency and Renewable Energy Integration in Poor Cities: Kabul Case Study. *Sustainability 2021, Vol. 13, Page 11984*, *13*(21), 11984. https://doi.org/10.3390/SU132111984

Shaltout, R. E. (2021). Key Challenges of Smart Railway Station. In I. , et al. El Dimeery (Ed.), Design and Construction of Smart Cities. JIC Smart Cities 2019. Sustainable Civil Infrastructures (pp. 287–294). Springer, Cham. https://doi.org/10.1007/978-3-030-64217-4_33

Sharma, S. K., & Kumar, A. (2017). Ride performance of a high speed rail vehicle using controlled semi active suspension system. *Smart Materials and Structures*, *26*(5), 055026. https://doi.org/10.1088/1361-665X/AA68F7

Singh, P., Elmi, Z., Krishna Meriga, V., Pasha, J., & Dulebenets, M. A. (2022). Internet of Things for sustainable railway transportation: Past, present, and future. *Cleaner Logistics and Supply Chain*, *4*, 100065. https://doi.org/10.1016/J.CLSCN.2022.100065

Stanisavljević, N., Stojanović, D., & Bogdanović, Z. (2023). Fostering Crowd-Based Open Innovations in Serbian Railways - Preliminary Readiness Assessment. In *Mihić, M., Jednak, S., Savić, G. (eds) Sustainable Business Management and Digital Transformation: Challenges and Opportunities in the Post-COVID Era.* (Vol. 562, pp. 278–297). Springer International Publishing.

Stanisavljević, N., Stojanović, D., Miletić, A., Lukovac, P., & Bogdanović, Z. (2023). Internal Stakeholders' Readiness for Developing Smart Railway Services Through Crowd-Based Open Innovations. In J. L. , Reis, M. K. , Peter, J. A. , Varela González, & Z. Bogdanović (Eds.), *Reis, J.L., Peter, M.K., Varela González, J.A., Bogdanović, Z. (eds) Marketing and Smart Technologies. Smart Innovation, Systems and Technologies* (Vol. 337, pp. 189–206). Springer, Singapore. https://doi.org/10.1007/978-981-19-9099-1_13

Stanisavljević, N., Stojanović, D., & Petrović, L. (2022). Open Innovation and Crowdsourcing: Challenges and Opportunities for Serbian Railways. In *E-business technologies conference proceedings* (Vol. 2, Issue 1, pp. 36–41). https://ebt.rs/journals/index.php/conf-proc/article/view/137

Stojanović, D., & Domazet, I. (2020). Use of Information Technologies in Educational Purposes-Case from Serbia. *Economic Analysis*, *53*(2), 68–78. https://doi.org/10.28934/ea.20.53.2.pp68-78

Stojanović, D., Stanisavljević, N., & Jovičić, E. (2021). Digital marketing techniques for promotion of "Infrastructure of Serbian Railways." In *International conference on E-business technologies (EBT)* (Vol. 1, pp. 84–88). Faculty of Organizational Sciences, University of Belgrade, Serbia.

Thurner, T., & Gershman, M. (2014). Catching the runaway train innovation management in Russian railways. *Journal of Technology Management & Innovation*, *9*(3), 158–168.

Wang, Y., Li, H. X., Meng, H. D., & Wang, Y. (2022). Dynamic characteristics of underframe semi-active inerter-based suspended device for high-speed train based on LQR control. *Bulletin of the Polish Academy of Sciences. Technical Sciences*, *Vol. 70*(nr 4), 141722. https://doi.org/10.24425/BPASTS.2022.141722

Yu, X., & Wan, K. (2022). High-Speed Rail Opening and Green Innovation—Evidence From China. *Frontiers in Environmental Science*, *10*, 901879. https://doi.org/10.3389/FENVS.2022.901879/BIBTEX

Yu, Z., Wang, H., & Chen, F. (2023). Security of railway control systems: A survey, research issues and challenges. *High-Speed Railway*, *1*(1), 6–17. https://doi.org/10.1016/J.HSPR.2022.12.001

Zhou, Y., Xiao, F., & Deng, W. (2023). Is smart city a slogan? Evidence from China. *Asian Geographer*, *40*(2), 185–202.

DRAFT VERSION

ECONOMIC MODELS AND ECONOMETRICS

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

Original article

REVISITING THE TOURISM-LED ECONOMIC GROWTH HYPOTHESIS: EMPIRICAL INSIGHTS FROM SERBIA

ALEKSANDRA ANIĆ[1], BRANISLAVA HRISTOV STANČIĆ[2], JELENA RAŠKOVIĆ[3]

[1] University of Belgrade, Faculty of Economics and Business, Belgrade, aleksandra.anic@ekof.bg.ac.rs 0000-0002-3058-0777
[2] University of Belgrade, Faculty of Economics and Business, Belgrade, branislava.hristov@ekof.bg.ac.rs 0000-0002-1285-6485
[3] University of Belgrade, Faculty of Economics and Business, Belgrade, jelena.raskovic@ekof.bg.ac.rs 0000-0001-8154-0161

Abstract: *Although the tourism–growth nexus is extensively explored in the tourism and economics literature, empirical evidence specific to Serbia remains limited. To address this gap, the present study tests the tourism-led growth hypothesis (TLGH) by examining the impact of international tourism receipts, to the real GDP, using quarterly data. The ARDL model reveals that international tourism receipts have a statistically significant positive impact on economic growth. The TLGH is confirmed by the existance of a long run cointegration relationship between international tourism receipts and gdp growth, after accounting for control variables such as total fixed assets and real effective exchange rate. This finding suggests that, in the Serbian context, policies aimed at promoting international tourism could serve as an effective instrument for stimulating sustained economic growth.*

Keywords: *TLGH, international tourism receipts, GDP, ARDL, Serbia.*

1. INTRODUCTION

The development of tourism has emerged as a strategic priority for many countries, largely due to the sector's substantial potential to contribute to economic growth (Gwenhure & Odhiambo, 2017). Its sustained relevance positions tourism as an important driver of socio-economic progress across the globe (Băndoi et al., 2020). Despite the high sensitivity of the tourism sector to external shocks, such as the COVID-19 pandemic, tourism continues to be recognized as one of the contributors to economic growth. It remains one of the most significant sectors of the global economy, accounting for approximately 10% of global GDP and 10.4% of total employment (Travel & Tourism: Economic Impact Report 2024). In the case of Serbia, tourism has experienced notable growth over the past decade, based on the country's rich natural and cultural heritage, while also incorporating the growing segments of entertainment and business tourism. Serbia has made significant improvement toward establishing itself as a regional hub for congress tourism. This multidimensional potential should be harnessed in a sustainable and strategic manner to stimulate economic growth. Accordingly, the primary aim of this study is to empirically examine the relationship between tourism and economic growth, commonly referred to in the literature as the TLGH. The existing literature on the TLGH in the context of Serbia is limited to only few studies (Mitra, 2019; Mirović et al.,2020; Hristov Stančić et al., 2022; Pagria et al., 2022). However, neither of these incorporates control variables that may also influence GDP growth. Furthermore, Mitra (2019), Mirović et al. (2020), and Pagria et al. (2022), have examined the TLGH with Serbia considered as part of broader, cross-country analyses focused on the Western Balkans. Hristov Stančić et al. (2022) has conducted a country-specific analysis exclusively focused on Serbia examining the relationship between international tourism receipts and GDP based on data that end in 2017, whereas the present research extends the time horizon and provides additional insights by examining more recent developments and also using different econometric methodology.

2. LITERATURE

The TLGH, formally introduced as the "tourism growth model" by Balaguer and Cantavella-Jordá (2002), posits that international tourism serves as a driver of long-term economic growth, establishing a unidirectional causal relationship from tourism development to economic performance. According to this hypothesis, increases in tourist arrivals and tourism-generated revenues contribute to the expansion of national output (Song & Wu, 2022). Since its introduction, the TLGH has attracted growing scholarly attention and has been

DRAFT VERSION

tested across a wide range of geographical and economic contexts, encompassing both advanced and emerging economies, as well as mature and developing tourism markets (Chiriko et al., 2025). Extensive empirical research has investigated this hypothesis in various regions, including the European Union (Matzana et al., 2022), the African continent (Chiriko et al., 2025), and sub-Saharan Africa (Baidoo et al., 2022).

A comprehensive review of the empirical findings on the tourism growth nexus by analyzing 100 relevant research articles (Brida et al., 2016) has shown that in most cases, international tourism positively influences economic growth, although some studies report divergent results. Moreover, most empirical investigations employ a bivariate modeling framework, allowing for the assessment of both tourism-led growth and growth-led tourism hypotheses (Brida et al., 2016).

Table 1: TLGH in previous empirical findings

Author(s)	Country/Countries	TLGH confirmed
Eyuboglu & Eyuboglu (2020)	9 emerging countries	Yes
Balaguer & Cantavella-Jordà (2002); Perles-Ribeset al., (2017)	Spain	Yes; Yes
Dritsakis (2004); Mavrommati et al., (2024)	Greece	Yes; Yes
Gunduz & Hatemi-J (2005), Demiroz & Ongan (2005); Katircioglu (2009); Terzi (2015).	Turkey	Yes; Yes; No; Yes
Jalil et al. (2013)	Pakistan	Yes
Nepal et al., (2019)	Nepal	Yes
Sanchez Carrera et al., (2008)	Mexico	Yes
Aktar et al., (2014)	Bangladesh	Yes
Brida & Risso, (2009)	Chile	Yes
Chiriko et al., (2025)	27 African countries	Yes

Source: Compiled by the authors.

3. EMPIRICAL MODEL AND METHODOLOGY

To test whether tourism drives economic growth (TLGH), researchers often estimate the following regression model:

$$\ln Y_t = \beta_0 + \beta_1 lnT_t + \beta_2 lnX_t + \varepsilon_t \quad (1)$$

where Y_t denotes GDP at time *t*, T_t represents tourism receipts, and X_t stands for a set of control variables.

In our analysis, we use two control variables: investment and the real effective exchange rate. Investment (total growth fixed assets) is included because it is one of the primary drivers of economic growth in many developing and transitional economies. The real effective exchange rate is also considered, as it directly influences a country's terms of trade, thereby affecting GDP growth.

We use the Autoregressive Distributed Lag (ARDL) model to test the TLGH in Serbia, due to its flexibility in dealing with time series that are integrated of different orders, I(0) or I(1). As a preliminary step, we examine the stationarity properties of the variables involved. This step is essential to ensure that the ARDL bounds testing procedure can be validly applied.

The bounds testing methodology, developed by Pesaran and Shin (1999) and Pesaran et al. (2001), is then used to determine whether a long-run equilibrium relationship exists between tourism and economic growth. This approach allows for heterogeneous short-run dynamics and provides robust results even when variables are integrated of mixed orders.

If cointegration is confirmed, the ARDL model can be reformulated into an Error Correction Model (ECM), which captures both short-run adjustments and long-run equilibrium dynamics. The baseline ARDL(p+1, q_1+1, q_2+1, q_3+1) model can be expressed as the following unrestricted conditional ECM specification:

$$\Delta y_t = \alpha_0 + \sum_{i=1}^{p} \varphi_i \Delta y_{t-i} + \sum_{j=0}^{q_1} \beta_j \Delta x_{1,t-j} + \sum_{k=0}^{q_2} \gamma_k \Delta x_{2,t-k} + \sum_{l=0}^{q_3} \delta_l \Delta x_{3,t-l} + \theta_0 y_{t-1} + \theta_1 x_{1,t-1} + \theta_2 x_{2,t-1} + \theta_3 x_{3,t-1} + e_t \quad (2)$$

DRAFT VERSION

where y_t is real GDP, x_{1t}, x_{2t}, x_{3t} are independent variables (tourism receipts, total fixed assets, real effective exchange rate) and e_t is an error term. Initially, as in the traditional ECM framework, the F-test is applied to assess the null hypothesis that there is no level relationship among the variables (H_0: $\theta_0=\theta_1=\theta_2=\theta_3=0$). Rejection of the null hypothesis indicates the existence of a long-run (level) relationship between real GDP and the explanatory variables: tourism receipts, real effective exchange rate, and fixed assets.

4. DATA AND DESCRIPTIVE STATISTICS

The data for research is obtained from Eurostat database and National bank of Serbia (NBS). The analysis is based on quarterly data spanning the period from 2010Q1 to 2024Q4. GDP and total gross fixed assets are seasonally and calendar adjusted data, chain linked volumes (2020), million euro data. International tourism receipts are in million euros. They are seasonally adjusted in EViews by using Tramo/seats option. Real effective exchange rate (REER) is expressed as the base index (2015=100), obtained from NBS. The REER is calculated by adjusting the nominal effective exchange rate for relative price levels between Serbia and its major trading partners. An increase in REER indicates a real appreciation (domestic goods become relatively more expensive, reducing competitiveness), while a decrease indicates a real depreciation (domestic goods become relatively cheaper, improving competitiveness). All data are log values.

Figure 1 presents data for international tourism receipt and GDP during the period of Q1 2010 – Q4 2024. International tourism receipt tripled in the 14 years period in Serbia. It increased from 605 mil. EUR in 2010 to 2,075 mil. EUR in 2024. International tourism receipts experienced a sharp decline of 25% in 2020 as a result of the COVID-19 pandemic. However, the reduction was temporary, since the receipts increased by 11% in 2021 compared with 2019. The increase in GDP during the whole period analyzed was 36%, whereas international tourism receipts increased by 243%.

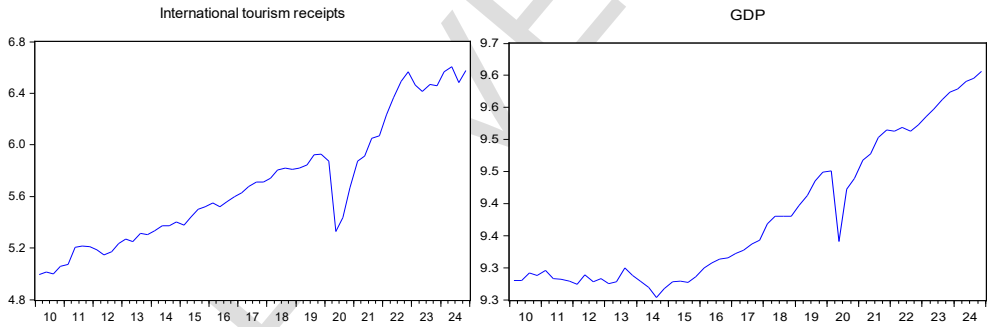


Figure 1: International tourism receipts and GDP, log values

5. ECONOMETRIC ANALYSIS

Results for Dickey-Fuller unit root test are presented in Appendix (Table 6). All four variables real GDP, tourism receipts, REER and fixed assets are integrated of order one. In second step, we proceed with ARDL approach.

The optimal ARDL model was initially selected based on the lowest AIC, starting from a maximum of 4 lags. However, since serial autocorrelation was detected at lag four and eight, the model was re-specified by increasing the number of lags, resulting in the ARDL (3, 4, 1, 0) specification, which resolved the autocorrelation issue. Based on this specification, the conditional ECM is estimated without imposing any coefficient restrictions (Table 4).

Table 2 reports the results of the ARDL bounds testing procedure. The null hypothesis of no level (long-run) relationship between real GDP, international tourism receipts, real effective exchange rate, and fixed assets is rejected at the 5% significance level. However, the result is inconclusive at the 1% level.

DRAFT VERSION

Table 2: Bounds test results

H_0 : There is no level relationship	**5% critical value**		**1% critical value**	
F statistic	I(0)	I(1)	I(0)	I(1)
5.068	3.356	4.701	4.701	6.360

Note: If the value of F statistic is above the I(1) critical value, we reject null hypothesis and conclude that the variables have long run level relationship. If the test statistic is below I(0) critical value, we fail to reject the null hypothesis. For values between two critical values, i.e. between critical values I(0) and I(1), the test is inconclusive.

Table 3 reports the estimates for the long-run level equation. The statistically significant impact of international tourism receipts on GDP supports the TLGH. A 1% increase in tourism receipts is associated with a 0.081% rise in GDP. As expected, investments in fixed assets exert the most substantial influence on GDP growth. A 1% increase in fixed assets results in a 0.225% increase in output. Moreover, the real effective exchange rate has a positive effect on GDP growth. Although a stronger REER (or appreciation) may reduce export competitiveness, it can improve terms of trade by lowering import prices relative to export prices. While this effect mainly influences nominal net exports, in some cases it can also support real GDP growth by reducing input costs and stimulating domestic consumption and investment.

Table 3: ARDL, estimate of level equation

Dep. var log GDP (*lgdp*)	**Serbia**
Log int. tourism receipts (*ltour*)	0.081***
	(0.019)
Log fixed assets (*lasset*)	0.225***
	(0.035)
Log real effective exchange rate (*lreer*)	0.251***
	(0.078)

Notes: Standard errors in parentheses. ***, ** and * denote respectively statistical significance at the 0.01, 0.05 and 0.1 levels.

The speed-of-adjustment coefficient is negative and statistically significant (Table 4), indicating that real GDP adjusts toward its long-run equilibrium path with tourism receipts, fixed assets, and the real effective exchange rate. The coefficient of –0.441 suggests that approximately 44% of any deviation from equilibrium is corrected within one period, implying moderate convergence speed.

Table 4: Estimate of ECM of the ARDL (3,4,1,0) , TLGH

Dependent variable $\Delta lgdp_t$	
z_{t-1}	-0.441***
	(0.103)
$\Delta lgdp_{t-1}$	0.147
	(0.146)
$\Delta lgdp_{t-2}$	0.207*
	(0.104)
$\Delta ltour_t$	0.070***
	(0.017)
$\Delta ltour_{t-1}$	-0.002
	(0.022)
$\Delta ltour_{t-2}$	-0.007
	(0.020)
$\Delta ltour_{t-3}$	0.023
	(0.017)
$\Delta lasset_t$	0.098**

DRAFT VERSION

	(0.039)
Dummy (Q3 2020=1)	0.054**
	(0.025)
Constant	2.645***
	(0.623)

Notes:

1) The regression follows the conditional ECM model presented in equation (2), applying the ARDL(3,4,1,0) specification. The variable z_{t-1} represents the equilibrium correction term lagged by one period, derived from the estimates reported in Table 3.
2) Adj R^2=0.849; N=56; Jarque-Bera test for normality: JB = 1.80 (0.407); Breusch-Godfrey LM test for autocorrelation: BG LM test (4) = 6.41 (0.171), (8) =13.71 (0.090), (12) 17.53 (0.131).
3) Standard errors in parentheses. ***, ** and * denote respectively statistical significance at the 0.01, 0.05 and 0.1 levels.
4) One impulse dummy variable is included to take only non-zero value 1 for Q32020 to capture the impact of the COVID-19 pandemic

6. CONCLUSION

The results support the validity of the TLGH in Serbia. Although the direct impact of international tourism receipts on real GDP is relatively modest, its statistical significance confirms that tourism is a relevant contributor to economic growth. This finding suggests that, in the Serbian context, policies aimed at promoting international tourism could serve as an effective instrument for stimulating sustained economic growth. Continued investment in the tourism sector is therefore both justified and strategically important. As traditional sources of growth, such as investments in fixed assets, gradually reach their saturation point, tourism may play an increasingly more significant role in supporting GDP growth in the years ahead.

LITERATURE

Aktar, M. A., Sadekin, M. N., & Saha, S. K. (2014). Relationship between tourist arrival and foreign exchange earnings: The case for Bangladesh. *Mediterranean Journal of Social Sciences, 5*(16), 162-165. DOI: 10.5901/mjss.2014.v5n16p162

Baidoo, F., Agbloyor, E. K., Fiador, V. O. L., & Marfo, N. A. (2022). Do countries' geographical locations moderate the tourism-led economic growth nexus in sub-Saharan Africa?. *Tourism Economics, 28*(4), 1009-1039. DOI: 10.1177/13548166211015512

Balaguer, J., & Cantavella-Jorda, M. (2002). Tourism as a long-run economic growth factor: the Spanish case. *Applied economics, 34*(7), 877-884. DOI: 10.1080/00036840110058923

Băndoi, A., Jianu, E., Enescu, M., Axinte, G., Tudor, S., & Firoiu, D. (2020). The relationship between development of tourism, quality of life and sustainable performance in EU countries. *Sustainability, 12*(4), 1628. https://doi.org/10.3390/su12041628

Brida, J. G., Cortes-Jimenez, I., & Pulina, M. (2016). Has the tourism-led growth hypothesis been validated? A literature review. Current Issues in Tourism, 19(5), 394-430. http://dx.doi.org/10.1080/13683500.2013.868414

Brida, J. G., & Risso, W. A. (2009). Tourism as a factor of long-run economic growth: An empirical analysis for Chile. *European journal of tourism Research, 2*(2), 178-185.

Chiriko, A. Y., Alemu, S. H., & Kim, S. (2025). Tourism-led growth hypothesis (TLGH) in Africa: does institutional quality matter?. *Tourism Review, 80*(2), 568-585. DOI: 10.1108/TR-04-2024-0280

Demiroz, D. M., & Ongan, S. (2005). The contribution of tourism to the long-run Turkish economic growth. *Ekonomický časopis*, 9, 880-894.

Dritsakis, N. (2004). Tourism as a long-run economic growth factor: an empirical investigation for Greece using causality analysis. *Tourism economics, 10*(3), 305-316. https://doi.org/10.5367/000000000418950

Eyuboglu, S., & Eyuboglu, K. (2020). Tourism development and economic growth: An asymmetric panel causality test. *Current Issues in Tourism, 23*(6), 659–665. https://doi.org/10.1080/13683500.2019.158886

DRAFT VERSION

Gunduz*, L., & Hatemi-J, A. (2005). Is the tourism-led growth hypothesis valid for Turkey?. *Applied Economics Letters, 12*(8), 499-504. https://doi.org/10.1080/13504850500109865

Gwenhure, Y., & Odhiambo, N. M. (2017). Tourism and economic growth: A review of international literature. Tourism: *An International Interdisciplinary Journal, 65*(1), 33-44.

Hristov Stančić, B., Đorđević, A., Kovačević, I., & Zečević, B. (2022). Tourism-led economic growth hypothesis-an empirical investigation for Serbia. *Teme*, 251-267. https://doi.org/10.22190/TEME210217014H

Jalil, A., Mahmood, T., & Idrees, M. (2013). Tourism–growth nexus in Pakistan:Evidence from ARDL bounds tests. *Economic Modelling, 35*, 185–191. http://dx.doi.org/10.1016/j.econmod.2013.06.034

Katircioglu, S. T. (2009). Revisiting the tourism-led-growth hypothesis forTurkey using the bounds test and Johansen approach for cointegration.*Tourism Management, 30*(1), 17–20. https://doi.org/10.1016/j.tourman.2008.04.004

Matzana, V., Oikonomou, A., & Polemis, M. (2022). Tourism activity as an engine of growth: lessons learned from the European Union. *Journal of Risk and Financial Management, 15*(4), 177. https://doi.org/10.3390/ jrfm15040177

Mavrommati, A., Kazanas, T. H. A. N. A. S. S. I. S., Pliakoura, A. L. E. X. A. N. D. R. A., Kalogiannidis, S. T. A. V. R. O. S., & Chatzitheodoridis, F. O. T. I. O. S. (2024). An empirical study on tourism and economic growth in Greece: an autoregressive distributed lag boundary test approach. *WSEAS Transactions on Business and Economics, 21*, 588-602.3 DOI: 10.37394/23207.2024.21.49

Mirović, V., Kalaš, B., & Pavlović, N. (2020). An empirical analysis of relationship between tourism and economic growth: panel evidence from Western Balkan countries. *5th International Thematic Monograph: Modern Management Tools and Economy of Tourism Sector in Present Era*, Belgrade, 39-52.

Mitra, S. K. (2019). Is tourism-led growth hypothesis still valid?. *International Journal of Tourism Research*, *21*(5), 615-624. https://doi.org/10.1002/jtr.2285

Nepal, R., Indra al Irsyad, M., & Nepal, S. K. (2019). Tourist arrivals, energy consumption and pollutant emissions in a developing economy–implications for sustainable tourism. *Tourism Management, 72*, 145–154. https://doi.org/10.1016/j.tourman.2018.08.025

Pagria, I., Risilia, D., & Sula, M. (2022). The impact of tourism on economic growth of some Western Balkan countries during Covid-19 period. *Calitatea, 23*(191), 353-362. DOI: 10.47750/QAS/23.191.40

Perles-Ribes, J. F., Ramón-Rodríguez, A. B., Rubia, A., & Moreno-Izquierdo, L. (2017). Is the tourism-led growth hypothesis valid after the global economic and financial crisis? The case of Spain 1957–2014. *Tourism Management, 61*, 96-109. https://doi.org/10.1016/j.tourman.2017.01.003

Sanchez Carrera, E. J., Brida, J. G., & Risso, W. A. (2008). Tourism's impact on long-run Mexican economic growth. *Economics Bulletin, 23*(21), 1-8. URL: http://economicsbulletin.vanderbilt.edu/2008/volume3/EB-07C20155A.pdf

Song, H., & Wu, D. C. (2022). A critique of tourism-led economic growth studies. *Journal of Travel Research, 61*(4), 719-729. https://doi.org/10.1177/00472875211018514

Terzi, H. (2015). Is The Tourism-Led Growth Hypothesis Tlgh Valid For Turkey?. *Doğuş Üniversitesi Dergisi, 16*(2), 165-178.

Travel & Tourism: Economic Impact Report 2024.

APPENDIX

Table A.1: Dickey-Fuller unit root test

Level: GDP	Cons, trend ADF(1) = -1.69	**I(1)**
FD: GDP	Cons, trend DF = -10.09	
Level: International tourism receipts	Cons, trend, dummy ADF(5) = -3.34	**I(1)**
FD: International tourism receipts	Cons, trend DF = -6.61	
Level: Real effective exchange rate	Cons ADF(3) = -0.02	**I(1)**
FD: Real effective exchange rate	Cons ADF(2) = -7.18	
Level: Fixed assets	Cons, trend ADF(2) = -1.85	**I(1)**
FD: Fixed assets	Cons, trend ADF(1) = -6.57	

Notes:

1) For the variable International Tourism Receipts, we performed the ADF test including a dummy variable (set to 1 for Q2 2020 and 0 otherwise) to account for a one-time structural break resulting from the impact of COVID-19
2) Critical values for 5% significance level $\tau_\mu = -2.91, \tau_t = -3.84$.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

THE CRISIS IN THE EUROPEAN AUTOMOTIVE INDUSTRY AND ITS IMPACT ON SERBIA'S ECONOMY

MIRJANA GLIGORIĆ MATIĆ[1], DUŠAN MARKOVIĆ[2],

[1] University of Belgrade, Faculty of Economics, Belgrade, mirjana.gligoric@ekof.bg.ac.rs, 0000-0002-5443-3560
[2] University of Belgrade, Faculty of Economics, Belgrade, dusan.markovic@ekof .bg.ac.rs, 0000-0003-1512-739X

Abstract: *Since the 2008 Great Recession, the EU automotive industry has faced disruptions from technological changes, the COVID-19 pandemic, and shifting global regulations. These challenges weakened financially strained EU carmakers. In response, many pursued mergers, acquisitions, and outsourcing to lower-cost labor markets on Europe's periphery. While these countries benefited from foreign direct investment, the current EU automotive crisis is slowing their growth and forcing economic strategy reconsideration. Fiat's investment around 2010 triggered significant growth in Serbia's automobile production, with exports peaking in 2013. Although motor vehicle exports later declined, the sector remains vital to Serbia's overall exports and has been bringing substantial foreign investment, driven by relatively low labor costs. However, the initiation of electric vehicle production this year carries substantial promise, but its impact on Serbia's export performance and GDP growth in the foreseeable future remains uncertain due to adverse external conditions. Given the importance of Serbia's automotive sector in production and exports, attracting additional technology-intensive foreign direct investments is essential for long-term growth, innovation, productivity gains, and deeper integration into global value chains. This would accelerate economic growth and help Serbia converge with more developed countries.*

Keywords: *Automotive industry, Europe, Supply chain, Serbia, Crisis.*

1. INTRODUCTION

The third decade of the 21st century has accelerated the shift of economic influence from the former Triad countries (USA, Western Europe, and Japan) to emerging markets led by China. Emerging markets' automotive companies, have managed to position themselves in global industries where multinational corporations from developed countries previously dominated. These companies have evolved from being suppliers of low-price and low-quality products to becoming dominant players in the global 'value for money' segment (Williamson, 2010). It once seemed that many technologically intensive and asset-heavy industries, such as the automotive sector, were protected from this shift (Abbosh et al, 2017). However, due to regulatory changes, unpredictable rising political risks, the emergence of disruptive technologies, economic stagnation in domestic markets, and the rise of emerging competitors from China, EU automotive companies are now facing growing business and financial challenges.

Deterioration of market position of EU companies is gradually affecting other members of their business ecosystem. In recent years, the companies have been increasing their investments in countries on the periphery of Europe that economically gravitate towards the European market. This strategy aims to reduce costs, improve supply chain control, and decrease dependency—particularly on China. Among these countries, Serbia has stood out due to its network of free trade agreements, relatively good infrastructure, skilled and affordable labor force (Marković, 2023). As a result, Serbia has attracted a significant number of automotive suppliers and one global car producers, Stellantis. This has led to the development of an automotive industry business ecosystem dominated by foreign subsidiaries (from Germany, China, Japan, and South Korea), while market influence of domestic companies remains negligible.

Negative trends from European market have, with a certain time lag, spread to outsourcing locations. Serbia is particularly vulnerable to this risk, as it is economically strongly connected to developed European countries and has based its growth strategy on foreign direct investments. An additional problem is that certain municipalities almost entirely depend on automotive companies that produce low-value-added vehicle components, whose demand is declining in the market.

DRAFT VERSION

This paper aims to explain the causes of the crisis in the automotive industry, possible scenarios of further development, and how it may impact Serbia's economy. The goal is to highlight the necessity of creating a strategy for development of automotive industry, which until now has evolved spontaneously. The government has focused on attracting investments from automotive suppliers investing in labor-intensive projects with low added value.

2. RESTRUCTURING OF EUROPEAN AUTOMOTIVE

The automotive industry is one of the key industries in Europe. Its significance stems from its contribution to employment, exports, innovation, and economic activity. According to the European Commission, the automotive sector in the EU generates approximately 13.8 million jobs and accounts for around 7% of the EU's GDP (European Commission, 2025). However, since the Great Recession of 2008, EU automotive industry has been in constant vibrancy, leading to changes in market structure. To analyze European automotive industry, we will use an extended version of Porter's Five Forces model (Porter, 2008), which, in addition to bargaining power of suppliers and customers, competitive rivalry, threat of new entrants and substitutes, also includes the impact of regulation and technology.

The automotive industry is one of the few industries that are asset-heavy, labor-intensive, and technology-intensive. For decades, traditional competitors from Europe, Japan, and the United States have dominated the industry. Due to these characteristics, some authors have argued that entry barriers are so high that the emergence of new competitors is unlikely (MacDuffie and Fujimoto, 2010), but since the Great Recession the situation has changed.

Due to the Great Recession of 2008 Ford was forced to sell its premium brands in Europe: Jaguar-Land Rover and Volvo. This created an opportunity for India's Tata and China's Geely to acquire valuable brands, allowing them to enter the premium segment in Europe. By combining their low-cost manufacturing expertise with acquired technology, brands, supplier networks, and management know-how, challengers from China and India successfully captured market share from European competitors (Marković, Rakita, and Filipović, 2015).

Another challenge for EU-based companies has been the rapid expansion of vehicle electrification. Although the European Commission adopted an agenda promoting the development of "green vehicles," European automakers lagged behind in embracing this technology. Meanwhile, Chinese challengers mastered electric vehicle technology and vertically integrated their supply chains, resulting in production costs in China being 30% lower than in Europe. Due to technological lag in electrification, EU companies lost 6 percentage points of market share in Europe and 5 percentage points in China between 2019 and 2022 (McKinsey & Company, 2023). Additionally, regulatory changes in the U.S. have posed further challenges for EU premium vehicle manufacturers, as the American market is highly important for them. In 2022, the Inflation Reduction Act was passed in the U.S., providing substantial financial subsidies for companies establishing manufacturing facilities within the country. On top of this, the announcement of increased tariffs on vehicles manufactured outside the U.S. in 2025 has made EU-produced cars entirely uncompetitive in this market.

European companies have not only lost ground in China and the U.S. but also in European market. Due to the war in Ukraine, European companies lost access to the Russian market, which accounted for 1.8 million vehicle sales before the pandemic (OICA, 2025). The most affected was Renault, which controlled Russia's leading brand, Lada, but was forced to transfer it to the Russian state for just 1 EUR. Additionally, sanctions policies prevented European companies from accessing cheap resources and energy from Russia, making their products even more expensive. The economic stagnation in Europe, which began with the pandemic, continued until 2025. During economic crises, consumers tend to postpone purchases of durable goods (Flatters and Willmott, 2009), leading to a decline in vehicle sales across the EU. The year 2022 was the worst for the European market since 1993, with sales reaching only 72% of pre-pandemic levels. Although sales began recovering afterward, by the end of 2024, they still accounted for only 84% of pre-pandemic sales, resulting in 3 million less vehicles sold. Alongside declining sales, the composition of vehicle sales changed, with a decreasing share of internal combustion engine (ICE) vehicles. By 2024, hybrid electric cars accounted for 31% of EU sales, battery electric cars for 14%, and plug-in hybrid electric cars for 7% (ACEA, 2025). The shift in demand structure is negatively impacting European manufacturers, as they lack expertise in electric battery production. In 2023, vehicles produced in China accounted for 20% of the battery electric vehicle segment in the EU, while Chinese brands represented 8% of the segment (ACEA, 2024).

Table 1: Sales of vehicles in EU 000 units (OICA, 2025)

2019.	2020.	2021.	2022.	2023.	2024.
18,424	14,081	14,141	13,296	15,179	15,409

DRAFT VERSION

Technological innovations have driven the development of the mobility services market, like ride-hailing, car sharing, etc. These services have become a true substitute for owning a personal vehicle. Consumer willingness to use these services is influenced by demographic factors and lifestyle, while fleet quality has a negligible impact on their decision (Hahn et al., 2020). Recognizing that this market will continue to grow in the future, many car companies have invested in this business, expecting that these services will compensate for the decline in revenue from vehicle sales.

Due to the development of digital and AI solutions, modern vehicle buyers expect them to always be connected, enhance safety, and provide various digital services. A challenge for car companies in Europe is that customers often do not perceive these services as additional and are not willing to pay extra for them (Deloitte, 2020). Another issue for EU car manufacturers is the changing structure of the value chain in the automotive industry. Some studies highlight that by the end of the second decade of the 21st century, hardware accounted for 90% of the perceived value of a vehicle, while it is expected that soon software and content will make up 60% of a vehicle's value (Xu and Liu, 2018).

Automotive industry supply chain restructuring is driven by vibrant startups working on the development of innovative digital solutions. Making business contacts with traditional automotive companies, emerging startups struggle to preserve their innovative solutions and impose them as the industry standard (Ferras-Hernández, Tarrats-Pons, and Arimany-Serrat, 2017). In this way, innovative digital suppliers are becoming increasingly important in the supply chain and gaining a better bargaining position compared to bureaucratic traditional automotive companies. An additional challenge is that innovative digital suppliers are displacing traditional suppliers from the supply chain, which may lead to a wave of consolidation.

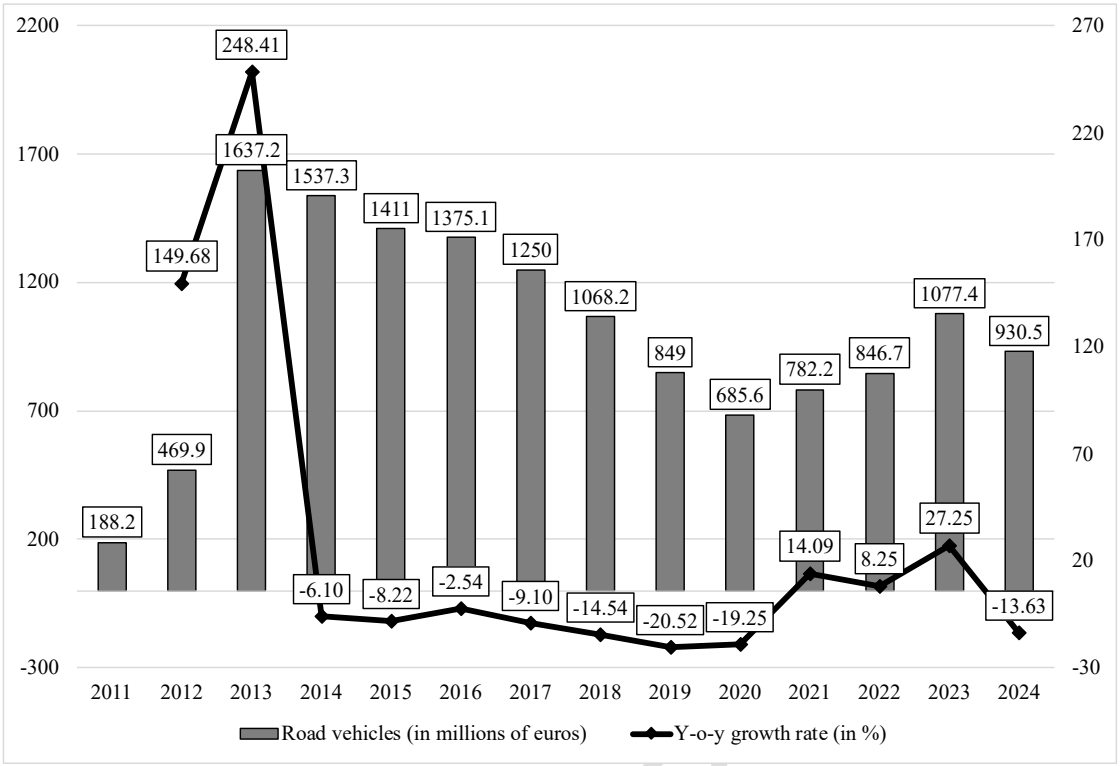
The analysis has shown that the EU automotive industry is at a turning point in its further development. Technological innovations that traditional EU competitors have not mastered, combined with unpredictable political events, have led to an increased negative impact of each of the five competitive forces on the profits of automotive companies. As a result, in 2024 EU vehicle production dropped by 19% compared to 2019 (OICA, 2025a). In the same year, over 30% of production facilities in Europe had a capacity utilization rate below 50% (Bloomberg, 2024). These negative trends in the EU automotive industry are gradually spilling over to outsourcing locations on the periphery of Europe.

3. SERBIA'S AUTOMOTIVE INDUSTRY: PRODUCTION AND EXPORT UNDER EUROPEAN DISRUPTIONS

FIAT's investments 15 years ago directly led to the beginning of a significant volume of car production and export. According to data for the export of road vehicles from the Statistical Office of the Republic of Serbia (SORS), the first "jump" in the value of the export of these products from Serbia was already noticeable in 2012. The biggest "boom" was achieved in 2013, when the export of road vehicles reached 1.64 billion euros, as much as 250% above the one in 2012. In the years that followed, there was a downward trend, until 2020, when, in the observed period, the minimum value of the export of these products of 686 million euros was reached. The start of the COVID-19 pandemic and the general global shutdown certainly partially contributed to this result in 2020. Although this export declined rapidly from 2013 to 2020, it recorded a value of over one billion euros during most of this interval (except for 2019 and 2020). The change in trend and the positive growth rate of the export of road vehicles were achieved in 2021[1], 2022 and 2023. In 2023, exports again exceeded one billion euros, which is 27% higher than the year before. However, the slow economic growth of the countries of the European Union, increased geopolitical tensions and customs conflicts have had an unfavorable effect on the export of cars from Serbia. The value of this export in 2024 was 13% lower than in 2023, i.e. amounted to 930 million euros. The fact that since the beginning of 2025 there has been no significant recovery in the economic activity of European countries, and tensions continued and even intensified, does not indicate favorable opportunities for the production and export of cars manufactured in Serbia. This is confirmed by the very low value of road vehicle exports in the first three months of 2025, which is even 40% lower than in the first quarter of 2024.

[1] In 2021, the merger of the Fiat Chrysler Automobiles (FCA) and Groupe PSA (Peugeot, Citroën, Opel, DS) automotive groups took place, creating the new multinational company Stellantis, the fourth largest car manufacturer in the world.

DRAFT VERSION



Graph 1: Serbia: Value of road vehicle exports and annual growth rate, 2011-2024

Source: Authors' representation based on data from SORS
Note: Including air-cushion vehicles

The share of the value of the road vehicles export in the total export was significantly reduced, from as much as 15% in 2013 to 10% in 2016, then to 5% in 2019. According to the latest data (for 2024), the export of road vehicles accounts for 3% of the total value of goods exported from Serbia. If we look at the entire sector of the automotive industry, the exports of the automotive industry in previous years averaged about 12% of the country's total goods exports. On average, motor vehicles accounted for about 30% of the total exports of the auto cluster, while electrical conductors, tires and other parts accounted for the remaining 70% in the period 2012–2024.[2]

Certainly, previous foreign investments in this sector and the inclusion of companies operating in Serbia in European production chains had a very positive effect on production and export results. In fact, according to NBS data, Serbia attracted 5.1 billion euros of foreign direct investment in the automotive cluster (which includes the production of motor vehicles, rubber and plastic products, batteries and other electrical equipment) in the period from 2012 to 2024, and foreign companies see relatively low labor costs in Serbia as key motive for these investments.[3] As a result, the export of this sector increased from 900 million euros in 2012 to 3.4 billion euros in 2024. The maximum result was reached in 2023, when the total export of the auto industry amounted to 3.6 billion euros.[4]

The start of serial production of the new electric model "Fiat Grande Panda" in February reflected the growth of the production of motor vehicles from the beginning of 2025. RZS data related to the growth in the production of motor vehicles, trailers and semi-trailers indicate that in 2024 a growth of 3.4% was achieved compared to 2023. Since the beginning of 2025, only in January has a lower level of production been registered in Production of motor vehicles, trailers and semi-trailers and in the Production of motor vehicles, while since February production activity has jumped sharply. In January, the production of motor vehicles, trailers and semi-trailers was 11.3% below that of January last year, i.e. 7.4% less compared to the average production of this area in 2024 (Table 1). Data for motor vehicle production indicates that it was particularly low compared to the average in 2024 - 73% lower. Already in February, there was a noticeable change in trend and a sudden

[2] NBS. *Inflation Report*, May 2025, p. 36.
[3] The average labor cost in the processing industry in Serbia in 2024 was 10.5 euros per hour, which is among the lowest in Europe (e.g. in Hungary and Poland the costs were about 15 euros per hour, while in the Czech Republic and Slovakia they were about 18 euros per hour). NBS. *Inflation Report*, May 2025, p. 38.
[4] NBS. *Inflation Report*, May 2025, p. 36.

DRAFT VER

increase in the production of motor vehicles, precisely as a direct consequence of the started production of electric cars. Growth indices indicate that the volume of motor vehicle production compared to the average level of production in 2024 was 9 times higher in February, almost 19 times in March and 22 times in April. (see Table 1). According to the data for the first four months of 2025, the year-on-year growth in the production of motor vehicles, trailers and semi-trailers, in accordance with this trend of growth in the production of motor vehicles, is 10% compared to the same period of the previous year, and 18% compared to the average production in 2024.

Table 1: Motor vehicle production, January-April 2025

	January 2025	February 2025	March 2025	April 2026	January-April 2025
	Year-on-year indices				
Manufacture of motor vehicles, trailers and semi-trailers	92.6	109.8	120.3	111.1	110.7
	2024=100				
Manufacture of motor vehicles, trailers and semi-trailers	88.7	121.8	134.7	124.4	117.7
Manufacture of motor vehicles	26.6	986.5	1886.8	2250.3	-

Source: SORS

In Serbia, a further increase in the production and export of cars, as well as other products of the automotive sector, is planned. Despite the challenges in the auto industry in Europe and the world, the inflow of FDI into this sector in Serbia continues. In addition to European companies, there are investments from companies from Asia too - primarily China, Japan and South Korea. Electric car production begins at the "Stellantis" factory in Kragujevac, and earlier production of the hybrid model is planned, as well as the possibility of production of the "Citroen C3" model. In addition, at the end of 2023, serial production of car tires was launched at the "Ling Long" factory in Zrenjanin, with an expected increase in volume during 2025. All this, barring major delays in production and export, will have a positive effect on the export of road vehicles, the export of the entire automotive sector, and will also be reflected in the growth of the gross domestic product.

The automotive sector of Serbia is currently facing uncertainty regarding the possibility of planned production and growth in the current external circumstances of unfavorable trends in the international market. The problems that the auto industry in Europe is currently facing, as well as the new 25% tariffs that the US has imposed on cars and parts, are likely to reduce the exports of the Serbian auto cluster. However, the production of a new electric model and the increased production of car tires should mitigate these negative effects. Although these production activities are expected to contribute positively to exports and accelerate GDP growth in the second half of 2025 compared to the beginning of the year, the question arises whether this potential will be realized. Despite the significant potential of starting the production of electric cars in Serbia, it is particularly uncertain to what extent this production will be able to contribute to Serbia's export performance and GDP growth in the coming years. The basis for such doubts arises from unfavorable external conditions, including geopolitical tensions and economic instability and turbulence in important export areas, above all in the European market, primarily Germany, since the local supplier sector is dominantly linked to that market. All this together represents a risk and imposes restrictions on the realization of the full potential of this new industrial initiative in Serbia.

4. CONCLUSION

EU automotive industry has been facing constant disruption since the Great Recession in 2008. After the recession, negative trends have been underlined by technology disruptions, Covid-19 pandemic and radical changes in global regulation, resulting in erosion of competitiveness of financially exhausted EU automotive companies. Companies have tried to deal with these trends, realizing certain M&A and outsourcing some activities to cheap labor force locations in the European periphery. These countries have experienced significant economic benefits from inflow of foreign direct investments in automotive industry, but due to the current crisis in EU automotive industries, these economies face a decreasing rate of economic growth and must reconsider their economic model.

Problems in the automotive industry in Europe left a mark on the production and export of the Serbian auto industry, which, thanks to significant inflows of foreign direct investments, continued to strengthen its importance in the production and export sector. Despite certain unfavorable effects, this sector remains crucial for the domestic economy. Although intensive development is predicted in the coming years, the question is to what extent can these plans be realized considering adverse external factors, including geopolitical tensions, currently weakened performance of the automotive industry in Europe and global economic uncertainty.

DRAFT VERSION

LITERATURE

Abbosh, O., Nunes, P., Savic, V., & Moore, M. (2017). The big squeeze: How compression threatens old industries. *MIT Sloan Management Review*.

ACEA (2024) https://www.acea.auto/fact/fact-sheet-eu-china-vehicle-trade-2024/ accessed June 2025

ACEA (2025) https://www.acea.auto/pc-registrations/new-car-registrations-0-8-in-2024-battery-electric-13-6-market-share/ accessed June 2025

Bloomberg (2024) https://www.bloomberg.com/news/articles/2024-09-11/vw-s-german-factory-closures-highlight-wider-problems-for-europe-s-carmakers accessed June 2025

Deloitte (2020) *Who is going to pay for the future of mobility?*. https://www2.deloitte.com/content/dam/insights/us/articles/automotive-news_industry-4-0-in-automotive/6470_whos-going-to-pay/6470_whos-going-to-pay.pdf. Accessed July 09th 2020

European Commission (2025) https://single-market-economy.ec.europa.eu/sectors/automotive-industry_en accessed June 2025

Ferras-Hernández X, Tarrats-Pons E, Arimany-Serrat N (2017) Disruption in the automotive industry: A Cambrian moment. *Business Horizons* 60(6): 855-863

Flatters P, Willmott M (2009) Understanding the Post – Recession Consumer. *Harvard Business Review* 87(7): 106-112

Hahn R, Ostertag F, Lehr A, Buttgen M, Benoit S (2020) "I like it but I don't use it": Impact of car-sharing business models on usage intentions in the sharing economy. *Business Strategy and Environment* 29(3): 1404-1418

MacDuffie, J. P. & Fujimoto, T. (2010). Why Dinosaurs Will Keep Ruling the Auto Industry. *Harvard Business Review*, 88(6), 23-25.

Marković, D. (2023). Uticaj Covid-19 pandemije na finansijske performanse auto-dobavljača u Srbiji. *Ekonomske ideje i praksa*, (48). 37-49

Marković, D., Rakita, B., & Filipović, D. (2016). Strategic importance of Cross-border acquisitions for emerging market multinationals. *Neostrategic Management: An International Perspective on Trends and Challenges*, 189-201.

McKinsey & Company (2023) https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/a-road-map-for-europes-automotive-industry#/ accessed June 2025

National Bank of Serbia (NBS). *Inflation Report*, May 2025

OICA (2025) https://www.oica.net/category/sales-statistics/ accessed June 2025

OICA (2025a) https://www.oica.net/category/production-statistics/ accessed June 2025

Porter, M. (2008). The Five Competitive Forces that Shape Strategy. *Harvard Business Review*, 86(1), 79-83.

Statistical Office of the Republic of Serbia (SORS), https://www.stat.gov.rs accessed June 2025

Williamson, P. J. (2010). Cost innovation: preparing for a 'value-for-money'revolution. *Long Range Planning*, 43(2-3), 343-353.

Xu J, Liu X (2018) Technology is changing what a premium automotive brand looks like. *Harvard Business Review*. https://hbr.org/2018/05/technology-is-changing-what-a-premium-automotive-brand-looks-like. Accessed July 2025

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Review article

DEMAND RATIONING RULES: BEYOND EFFICIENCY AND PROPORTIONALITY

BOJAN RISTIĆ[1], NIKOLA NJEGOVAN[2]

[1] University of Belgrade, Faculty of Economics and Business, Belgrade, bojan.ristic@ekof.bg.ac.rs, 0000-0002-9883-8914
[2] University of Belgrade, Faculty of Economics and Business, Belgrade, nikola.njegovan@ekof.bg.ac.rs, 0000-0002-9531-323

Abstract: *This paper examines demand rationing mechanisms in oligopoly models beyond the classical efficient and proportional rules, with a focus on capacity-constrained price competition. Efficient rationing assumes that consumers with the highest willingness to pay are served first by the lower-priced firm. In contrast, proportional rationing assigns all consumers an equal chance to buy. We introduce a general framework parameterised by α, encompassing both classical rules as special cases and enabling the modelling of switching-cost-based rationing. Such a framework captures the impact of consumer inertia, search frictions, or contractual obligations on residual demand. Our analysis reveals that different rationing rules have a significant effect on equilibrium outcomes and firms' strategic choices, particularly in terms of capacity and pricing. Departures from the efficient rule may weaken or amplify competitive intensity, thus reshaping market power. These insights directly affect competition protection policy, particularly in the areas of horizontal restrictive agreements, abuse of dominant position, and merger control. The findings advocate for a more nuanced integration of rationing rules into oligopoly models and policy tools.*

Keywords: *capacities, Bertrand-Edgeworth competition, rationing rules, switching costs.*

1. INTRODUCTION

Demand rationing rules are crucial in oligopoly models, particularly in the context of capacity-constrained price competition, as they determine how market demand is allocated among competing firms. This allocation has a direct impact on firms' strategic incentives, equilibrium outcomes, and overall market efficiency. The key question arises: What kind of demand does the more expensive firm face compared to its cheaper rival? Thus, at the heart of the matter lies the shape of the residual demand curve that a relatively more expensive firm faces.

Rationing mechanisms have long played a central role in models of imperfect competition where supply cannot meet demand due to binding capacity constraints. Classical Bertrand-Edgeworth frameworks typically assume that excess demand is allocated via proportional or efficient rationing, which is treated as exogenous to the model. In short, under efficient rationing, the cheaper firm sells to those consumers with the highest willingness to pay—at least up to the limit of its own capacity—while the more expensive firm is left to serve only the lower end of the market. On the other hand, under the proportional rationing rule, each consumer has an equal chance of being allocated to either the cheaper or the more expensive product. According to Dixon (1987), this makes the rule probabilistic, in contrast to the efficient rationing rule, which is deterministic.

This work builds on the classical literature on capacity-constrained competition, notably that of Edgeworth (1897), Levitan & Shubik (1972), and Kreps & Scheinkman (1983). These models typically examine the implications of Bertrand's competition under capacity constraints by treating rationing mechanisms as exogenous to the model. This paper poses the following question: What if rationing rules differ from the standard *efficient* or *proportional* mechanisms while remaining exogenous to the competition model? More precisely, what factors might shape such a rule? What would the market outcome be under competition between firms with capacity constraints, and what policy implications would this have? In the Bertrand-Edgeworth setting, where rationing becomes necessary due to excess demand, a price war does not necessarily lead to the Bertrand paradox. Indeed, in the seminal model of Kreps and Scheinkman (the KS model), a two-stage game involving capacity choice followed by price competition yields a Cournot-type equilibrium. However, this outcome crucially depends on the assumption of efficient rationing, and the model proves to be non-robust to changes in the rationing mechanism. Davidson & Deneckere (1986) offer a widely cited critique of the KS framework, demonstrating that the choice of rationing rule is central to its validity. If a different rationing rule replaces the efficient one, *ceteris paribus*, the outcome of the two-stage game no longer aligns

DRAFT VERSION

with that of the Cournot model. They prove this by using proportional rationing, which they associate with Edgeworth's original setup. Complementing this theoretical insight, Jacobs & Requate (2016) provide experimental evidence showing that the choice of rationing rule significantly affects the shape and outcome of residual demand, with proportional rationing leading to higher prices and profits, thereby challenging standard equilibrium predictions.

Furthermore, Davidson & Deneckere (1986) characterise these two rationing rules, efficient and proportional, as extremes, suggesting that a continuum of other, more nuanced rationing rules lies between them. In contrast, Herk (1993) introduces an additional assumption in shaping residual demand for high-priced firms: consumers face high switching costs when moving from one firm to another (Examples include: early termination fees as a form of contractual lock-in, loyalty programs, data or infrastructure incompatibility in tech industries, behavioural inertia of consumers caused by habitual usage or search costs, etc.). This introduces a new form of rationing rule that may result in residual demand that falls outside the range defined by the classical extremes. Consumer switching costs shape this rationing rule. Accordingly, we might refer to it as *switching-cost-based rationing*. In that case, it turns out that competitors' capacities and prices are not the only decisive factors in shaping the residual demand faced by the higher-priced firm.

Before attempting to answer the questions posed, it is essential to consider what constitutes efficient rationing and which alternative rules might plausibly shape the residual demand faced by the higher-priced firm. Accordingly, this paper is structured into three interrelated sections following the introductory remarks. The following two chapters discuss how alternative demand allocation rules might be formulated alongside efficient rationing, as the more commonly employed approach in capacity-constrained oligopoly models. The final chapter concludes the discussion, summarising the observed implications of different rationing mechanisms on economic policymaking, with a particular emphasis on competition protection policy.

2. LOGIC OF EFFICIENT RATIONING

In a capacity-constrained duopoly competition, if one firm sets a higher price than the other, its demand depends on consumer preferences and the market share that the lower-priced firm can and is willing to serve. The rationing rule determines how excess demand (beyond the capacity of the cheaper firm) is distributed, which becomes the residual demand faced by the other firm. The number of possible rationing rules is infinite. In this paper, in addition to the efficient rationing rule, the focus will also be on its two alternatives: the proportional rule, a classical example rooted in Edgeworth's work, and the so-called switching-cost rule. However, this chapter focuses on the economic meaning of efficient rationing. With efficient rationing, the higher-priced firm serves only consumers with a low willingness to pay, as the lower-priced firm captures the top portion of market demand up to available capacities, thereby maximising consumer welfare in such circumstances. Therefore, this indicates an entirely rational demand spillover in cases of limited capacity. Consider a situation where two direct competitors on a homogeneous product market (firms i and j) compete for market demand defined by $p = 1 - q$. The capacities of firms i and j (denoted k_i and k_j, respectively) are assumed to be constrained, such that neither firm can serve the entire market when in the position of the lower-priced seller. Thus, the limit on "cheaper buying" is effectively imposed by the capacity of the cheaper firm. If a firm i were to undercut its competitor, the sales it faces would be

$$q_i = \min\{k_i,\, q(p_i)\} = \min\{k_i,\, 1 - p_i\}\,, \tag{1}$$

and on the other hand, the sales of firm j, as the higher-priced competitor, would amount to

$$q_j = \min\left\{k_j,\, \max\left[0,\, q(p_j) - k_i\right]\right\} = \min\left\{k_j,\, \max\left[0,\, 1 - p_j - k_i\right]\right\}. \tag{2}$$

Finally, if the two firms were to charge identical prices, the sales of firm i would be

$$q_i = \min\left\{k_i,\, \left(\frac{k_i}{k_i + k_j}\right) q(p_i)\right\} = \min\left\{k_i,\, \left(\frac{k_i}{k_i + k_j}\right) (1 - p_i)\right\}. \tag{3}$$

Following expression (3), each firm would satisfy a share of total market demand at the same price proportional to its capacities relative to the total available capacities on the market, but without exceeding its individual capacity constraint. Note that in the case of equal capacities, expression (3) simplifies to $1/2(1 - p_i)$, indicating that demand is equally divided between the two competitors at a uniform price. Expression (2) is key when

DRAFT VERSION

discussing the rationing rule, as it determines the residual demand the higher-priced firm faces. Its graphical representation is given in Figure 1(a) (the blue line, shown as the demand function shifted in parallel by the capacity of the cheaper firm).

As Vives points out, the efficient rationing rule eliminates the possibility of resale, which is a logical feature since it assumes that individuals with the highest willingness to pay are served first at a lower price (Vives, 1999, p. 126). In contrast, Tirole insightfully notes that the residual demand defined by the efficient rule is precisely the one that would emerge if consumers could freely resell the good among themselves—i.e. if they could conduct costless arbitrage (Tirole, 1988, p. 214). This interpretation enables individuals with a lower reservation price, motivated by resale opportunities, to access the cheaper product, resell it, and reappear in the pool of consumers that the higher-priced firm faces. However, the resale mechanism that would realise the efficient rule presupposes a fully efficient exchange market, which Tirole considers an overly strong assumption (Tirole, 1988, p. 213).

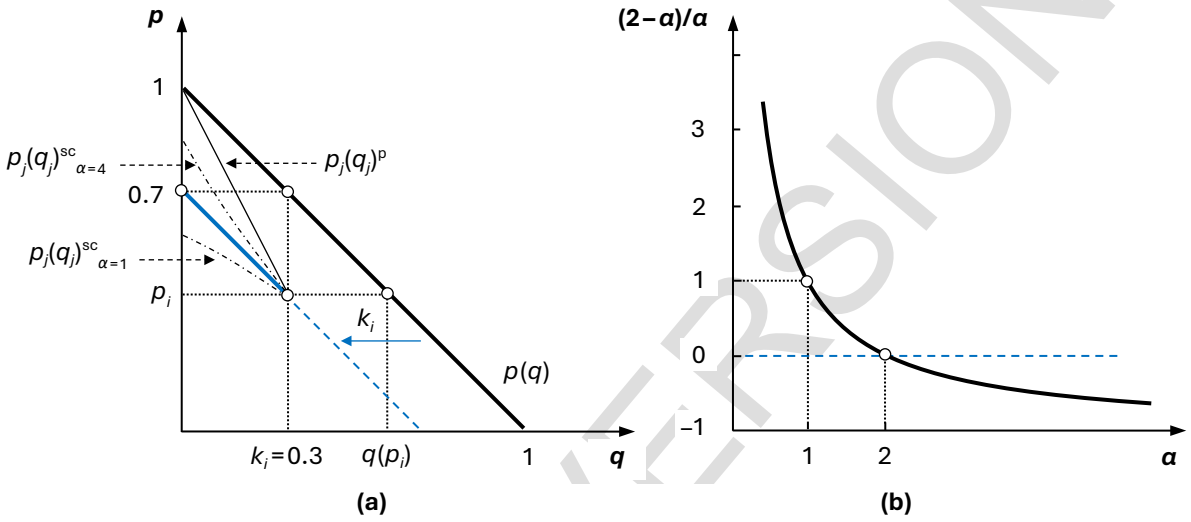


Figure 1: Alternative rationing rules

Consumers willing to pay more are served before those willing to pay less. Why? Does this necessarily mean that the seller must be able to distinguish between buyers based on their willingness to pay? These are some of the critical questions for understanding the context in which it is realistic to expect efficient rationing. Suppose a buyer's willingness to pay decreases with their geographical distance from the seller. In that case, those with the highest reservation prices would be located closest to the seller and, as a result, would be better informed about the timing and terms of sale. This would increase the likelihood that they are the first to be served when the product becomes available. In such a setting, the seller does not need to know which individual buyer has the highest reservation price—this will naturally be the one who shows up first at the shop door.

On the other hand, a buyer's position in the queue for the cheaper product may be interpreted as the result of individual market effort, and this effort need not be correlated with the consumer's reservation price. For example, if participating in the market requires sacrificing free time (commonly considered a normal good), it is plausible that individuals with lower income levels, and thus lower reservation prices, may exert greater market effort due to more available free time. As a result, individuals with lower reservation prices may get cheaper products sooner than those with higher ones. If resale is not the motive behind waiting in line, then the resulting allocation of goods will differ from that implied by the efficient rationing rule. As Dixon (1987) argues, under such conditions, the residual demand may deviate from that shaped by efficient rationing. This insight invites a further question: What types of residual demand might lie *above* or *below* the "blue line" from Figure 1(a), indicating weaker or stronger consumer responsiveness to price undercutting? Or, what are the alternatives to efficient rationing?

3. ALTERNATIVES TO EFFICIENCY

Let's begin this departure from efficiency by considering the *proportional rule* applied in the original Edgeworth (1925) model. The situations in which a firm is a lower-priced competitor or where prices are equal are interpreted in the same way as under the efficient rule—these correspond to expressions (1) and (3),

DRAFT VERSION

respectively. However, in the case where firm j is the higher-priced competitor, the residual demand it faces will take the form

$$q_j = \min\left\{k_j,\ \max\left[0,\ q(p_j) - k_i\frac{q(p_j)}{q(p_i)}\right]\right\}$$
$$= \min\left\{k_j,\ \max\left[0,\left(1-\frac{k_i}{q(p_i)}\right)q(p_j)\right]\right\} = \min\left\{k_j,\ \max\left[0,\left(1-\frac{k_i}{1-p_i}\right)(1-p_j)\right]\right\}. \quad (4)$$

The residual demand faced by firm j is obtained by multiplying the total market demand at the higher price ($p_j > p_i$) by a constant—that is, the proportionality factor $[1 - k_i/(1-p_i)]$. This factor reflects the share of total demand that *cannot* be served at the lower price and, therefore, remains available to be served at the higher price. The defining feature of the proportional rationing rule, and its principal distinction from the efficient one, is that every consumer with a sufficient willingness to pay can purchase from either firm based on the "first come, first served" principle. For example, when $k_i = 0.3$ and $p_i = 0.4$, the inverse residual demand curve shaped by the proportional rule for firm j is given by $p_j = 1 - 2q_j$, as shown in Figure 1(a) as $p_j(q_j)^{\mathrm{P}}$. It represents the original market demand rotated around the maximum willingness to pay, $p(0) = 1$.

Since the Edgeworth rule is not deterministic, as the efficient, the higher-priced firm cannot know with certainty which contingent of consumers will turn to it. However, it does know that they are not necessarily those with the lowest reservation prices. Consequently, the proportional rule provides the higher-priced firm with a more advantageous residual demand curve than the efficient one, allowing for softer competitive pressure in the oligopoly market structure.

In Ruebeck (2011), the formation of residual demand is critically examined through the lens of consumer choice, challenging traditional assumptions about its derivation. Thus, the residual demand is derived through a simulation of consumer arrivals at the firm's door based on a modified version of the proportional rationing rule. As in the case of the standard proportional rule, consumers are assumed to arrive at sellers randomly, while both the higher- and lower-priced firms have equal chances of being visited. This stems from the assumption that consumers lack information about which firm offers the lower price until they find out through experience, and they will purchase if their willingness to pay exceeds the price. In contrast, suppose that the willingness to pay is below the price. In that case, the individual will redirect to the other seller or forgo the purchase altogether if the second price also exceeds her willingness to pay. Naturally, introducing restrictions on the operation of the canonical Edgeworth rule would alter the nature of the resulting residual demand. It becomes clear that the issue of demand allocation is fundamentally empirical in nature. In such a context, introducing a rationing rule into the economic model is "a substitute for a complete analysis of consumer behaviour" (Tirole, 1988, p. 214).

Davidson & Deneckere (1986) suggest that efficient and proportional rationing represent boundary cases, framing a continuum of intermediate, more context-specific rules. In contrast to this view, Herk (1993) introduces an additional assumption: consumers in particular markets face significant switching costs when moving from one firm to another (a kind of "first come, first served", but with the burden of switching costs). Residual demand can then surpass the bounds defined by the classical extremes. In such cases, residual demand would depend on the intensity of switching costs (therefore, the *switching-cost rule*), which in turn defines the strategic room for manoeuvre available to the higher-priced firm. As in Herk (1993), for linear market demand, sales for higher-priced firms can be formalised as

$$q_j = \min\left\{k_j, \max\left[0,\ q(p_j) - k_i\left(\frac{q(p_i)}{q(p_j)}\right)^{\frac{2-\alpha}{\alpha}}\right]\right\},\ \text{for } \alpha \in (0,\infty). \quad (5)$$

As $\alpha \to 0$, the exponent in expression (5), $(2-\alpha)/\alpha$, tends toward infinity; conversely, as $\alpha \to \infty$, the exponent approaches -1. The function is positive for $\alpha < 2$ and negative for $\alpha > 2$. This behaviour of the exponent within the domain of definition of parameter α is illustrated in Figure 1(b). In this way, Herk (1993) used the parameter α to represent the impact of switching costs on the formation of residual demand, and consequently, their influence on the intensity of price competition between duopolists. In such a framework, when $\alpha = 2$, the rule corresponds to efficient rationing, while for $\alpha = \infty$, it reflects Edgeworth's proportional rule—as represented by expressions (2) and (4), respectively. For $\alpha = 1$, as in Herk (1993), the residual demand

DRAFT VERSION

is depicted in Figure 1(a) as $p_j(q_j)^{sc}{}_{\alpha=1}$. Given the capacity and price of the cheaper firm previously used (namely, $k_i = 0.3$ and $p_i = 0.4$), we obtain a nonlinear concave function: $q_j = (1 - p_j) - 0.3 \times [0.6/(1 - p_j)]$ which corresponds to $p_j(q_j)^{sc}{}_{\alpha=1}$, and lies below the residual demand curve associated with the efficient rule. Above the "blue line," for $\alpha = 4$, *ceteris paribus*, we would obtain a nonlinear convex function of the form: $q_j = (1 - p_j) - 0.3 \times [0.6/(1 - p_j)]^{-0.5}$.

In general, for $\alpha < 2$ within the domain of definition of α, one observes a family of concave residual demand functions that can be regarded as relatively less favourable for the higher-priced firm compared to the benchmark (efficient rationing rule). Conversely, for $\alpha > 2$, the corresponding family of convex residual demand functions is relatively more favourable to the higher-priced firm.

It is worthwhile to examine the market circumstances associated with different values of the parameter α, particularly in the context of consumer switching costs. Table 1 below provides a numerical example illustrating the behaviour of residual demand for values of α: 1, 2, 4 i ∞, based on discrete price increases by the higher-priced firm, using the same parameterisation as in the example shown in Figure 1(a).

Table 1: Consequences of different α

α	p_j: 0.4 → 0.45		p_j: 0.5 → 0.55	
	Δq_j	Arc ε	Δq_j	Arc ε
1	0.077	- 2.513	0.090	- 9.947
2	0.050	- 1.545	0.050	- 3.000
4	0.037	- 1.125	0.036	- 1.813
→ ∞	0.025	- 0.741	0.025	- 1.107

Based on the example presented in Table 1, it can be concluded that as the parameter α increases, the residual demand becomes more inert to price changes, as reflected in the values of the arc price elasticity of demand (Arc ε). Seemingly, higher values of α are more favourable to the higher-priced firm. When the magnitude of α is interpreted in the context of switching costs, lower values of the parameter α (below 2) indicate low switching costs—consistent with high price elasticity—whereas higher values (above 2) suggest the presence of significant switching costs. Therefore, α is governing the slope and position of the residual demand curve for firm j, thus affecting its pricing incentives and market power. The benchmark of full efficiency is related to $\alpha = 2$, where we have complete deterministic sorting by willingness to pay, where only the difference in prices matters.

4. CONCLUDING REMARKS

Our discussion on alternative demand rationing rules contributes to the field of oligopoly theory from at least three perspectives: first, in terms of the theoretical domain and functioning of economic models; second, by enhancing our understanding of how residual demand is formed for the higher-priced firm; and finally, by offering implications for the conduct of competition protection policy, given that oligopolistic markets are typically at the centre of competition authorities attention.

Thus, the considerations presented in this paper contribute to the discourse on demand-rationing rules beyond the framework defined by the classical rationing principles typically employed in models, namely, the efficient and proportional rules. This approach encompasses all key rationing forms as special cases of a more general form, thereby opening space for theoretical and empirical investigations of firm behaviour in oligopolistic markets. The focus of analyses of oligopolistic conduct, both from a modelling and an empirical perspective, could be situated within the context of the parameter α and the intensity of the lock-in caused by switching costs. Even in homogeneous product markets, the price differences and mode of customer rationing under capacity constraints need not be decisive for customer switching, which may significantly influence firms' strategic behaviour within competition models, such as the framework of the KS model.

Departures from the efficient rationing rule can invalidate the standard conclusion of the KS model—namely, that the outcome of the Cournot model necessarily emerges from a two-stage game where duopolists first choose capacities and then compete in prices. For instance, under any rationing rule more favourable to the higher-priced firm than the efficient one, if both firms hold Cournot-level capacities and one sets the Cournot price, the other would find it profitable to set a price above the Cournot level. Being aware of the probability that its rival will choose a price from the Edgeworth range above its own, each firm may hold some reserve capacities as a precautionary measure. Firms would do so because it allows them to increase output above the Cournot level and capture a larger market share by undercutting their competitors. A more favourable

DRAFT VERSION

residual demand compared to the efficient rule also implies greater room for price fluctuations for the higher-priced firm and, consequently, higher potential profits, which would induce firms to choose capacities exceeding the Cournot output level.

Conversely, in the case of residual demand that is less favourable to the high-priced firm than that predicted by the efficient rule, the market outcome is likely to be less competitive. Caution would dictate that firms in equilibrium hold lower capacity than what is expected under the Cournot equilibrium.

Finally, discussing the demand rationing rules may have essential implications for enforcing competition protection policy. The assessment of market power is particularly relevant in cases involving the abuse of dominance, as well as in the context of horizontal restrictive agreements and merger control.

Market power may persist due to hidden structural favouritism or behavioural frictions that are not readily observable through conventional indicators such as price-cost margins or the Herfindahl-Hirschman Index. Therefore, it would be helpful to develop empirical methods for estimating the value of the parameter α across different relevant markets. For instance, by comparing observed and predicted residual demand flows or by employing consumer surveys to assess brand loyalty, search frictions, and behavioural inertia. By logic, values of α above two may indicate significant market power necessary for claims related to abuse of dominant position, as well as greater cartel stability. A firm that undercuts the cartel price gains little additional demand, as many customers remain locked in with the higher-priced firm.

In the context of merger control, it is crucial to recognise that mergers may affect not only market concentration but also the structure of demand allocation. Accordingly, merger simulation models should incorporate rationing rules and switching costs while examining whether a merger may lead to implicit changes in α (through mechanisms such as shared loyalty programmes, unified branding strategies, contractual bundling, etc.). In any case, conditional merger approvals should include behavioural measures that ensure consumer choice neutrality based on the knowledge of the efficient rationing rule and its alternatives.

LITERATURE

Davidson, C. & Deneckere, R. (1986). Long-Run Competition in Capacity, Short-Run Competition in Price, and the Cournot Model. *The RAND Journal of Economics*, 17(3), 404-415. https://doi.org/10.2307/2555720

Dixon, H. (1987). The General Theory of Household and Market Contingent Demand. *The Manchester School*, 55(3), 287-304. https://doi.org/10.1111/j.1467-9957.1987.tb01303.x

Edgeworth, Y. F. ({1897} 1925). *The Pure Theory of Monopoly*. In: *Papers Relating to Political Economy*, Vol. I, Chapter E, (pp. 111-142). New York: Burt Franklin.

Farrell, J. & Klemperer, P. (2007). Coordination and lock-in: Competition with switching costs and network effects. In: M. Armstrong and R. Porter (Eds.), *Handbook of Industrial Organization*, Vol. 3, (pp. 1967-2072). Elsevier. https://doi.org/10.1016/S1573-448X(06)03031-7

Herk, F. L. (1993). Consumer Choice and Cournot Behavior in Capacity-Constrained Duopoly Competition. *The RAND Journal of Economics*, 24(3), 399-417. https://doi.org/10.2307/2555965

Jacobs, M. & Requate, T. (2016). Demand rationing in Bertrand-Edgeworth markets with fixed capacities: An experiment, Economics Working Paper, No. 2016-03, Kiel University, Department of Economics, Kiel. https://hdl.handle.net/10419/125820

Klemperer, P. (1987). Markets with consumer switching costs. *Quarterly Journal of Economics*, 102(2), 375-394. https://doi.org/10.2307/1885068

Kreps, M. D. & Scheinkman, A. J. (1983). Quantity Precommitment and Bertrand Competition Yield Cournot Outcomes. *The Bell Journal of Economics*, 14(2), 326-337. https://doi.org/10.2307/3003636

Levitan, R. & Shubik, M. (1972). Price Duopoly and Capacity Constraints. *International Economic Review*, 13(1), 111-122. https://doi.org/10.2307/2525908

Ruebeck, S. C. (2011). Consumer Search, Rationing Rules, and the Consequence for Competition. In: J. Salerno, S. Jay Yang, D. Nau and Sun-Ki Chai (Eds.), *Social Computing, Behavioral-Cultural Modelling and Prediction* (pp. 155-162). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-19656-0_24

Tirole, J. (1988). *The Theory of Industrial Organization*. The MIT Press, Cambridge Massachusetts.

Vives, X. (1999). *Oligopoly Pricing: Old Ideas and New Tools*. The MIT Press, Cambridge Massachusetts.

DRAFT VERSION

ENVIRONMENTAL MANAGEMENT AND NATURAL RESOURCE MANAGEMENT

DRAFT VERSION

SYM-OP-IS 2025
52. Simpozijum o operacionim istraživanjima
Palić, 7 - 10.9.2025.

UNIVERZITET U BEOGRADU
FAKULTET ORGANIZACIONIH NAUKA

Originalni članak

TREND ANALYSIS OF PRECIPITATION AND AIR TEMPERATURE OBSERVED AT THE METEOROLOGICAL STATION BELGRADE

JELISAVETA VAKANJAC[1], ALEKSANDAR TANASKOVIĆ[2], JUGOSLAV NIKOLIĆ[3], BORIS VAKANJAC[4], VESNA RISTIĆ VAKANJAC[5],

[1] Ministry of Defence Republic of Serbia, Birčaninova 5, Belgrade, jelisaveta.vakanjac@gmail.com,
[2] University of Belgrade, Faculty of Mining and Geology, Đušina 7, Belgrade, aleksandar.tanaskovic@rgf.bg.ac.rs, 0009-0005-2668-3588, vesna.ristic@rgf.bg.ac.rs, 0000-0002-2040-8252,
[3] Republic Hydrometeorological Service of Serbia, Kneza Višeslava 66, Belgrade, jugoslav.nikolic@hidmet.gov.rs,
[4] Military Geographical Institute "General Stevan Bošković", Mije Kovačevića 5, Belgrade, and IRASA Academy borivac@gmail.com 0000-0002-9429-6943
[5] University of Belgrade, Faculty of Mining and Geology, Đušina 7, Belgrade, and IRASA Academy, 0000-0002-2040-8252,

Abstract: *The meteorological station in Belgrade is the station with the longest series of observations of certain meteorological parameters. Lyceum professor, Vladimir Jakšić, on January 1, 1948, on his estate in Senjak, was the first, on his own initiative, who started with systematic air temperature measurements, he observed the sky and recorded meteorological phenomena and all changes on the Sava River. On his initiative, six years later, a network of meteorological stations was established on the territory of what was then Serbia, which included 20 stations that year, and another 7 stations were included a year later. Meteorological station Belgrade, in view of its observation series, has the longest time series of individual meteorological parameters that can be analyzed in order to confirm or deny existing climate changes. For the purposes of this work, data were collected on monthly and annual values of air temperature and precipitation for the observation period 1946-2023. year. Time series trend analyzes were performed. The obtained results of the applied analyzes are presented in the paper.*

Keywords: *Air temperature, precipitation, trend analysis, weather station Belgrade.*

Rezime: *Meteorološka stanica Beograd je stanica sa najdužim nizom osmatranja pojedinih meteoroloških parametara. Profesor Liceja, Vladimir Jakšić je na svom imanju na Senjaku 1. januara 1948. godine prvi u tadašnji Srbiji, samoinicijativno, otpočeo sa sistematskim merenjima temperature vazduha, osmatrao je stanje neba i beležio meteorološke pojave i promene na reci Savi. Na njegovu incijativu, šest godina kasnije, uspostavlja se mreža meteoroloških stanica na teritoriji Srbije koja je te godine obuhvatila 20, a godinu dana kasnije je uključeno još 7 stanica. Dakle, meteorološka stanica Beograd u Srbiji, obzirom na svoj osmatrački niz, poseduje najduže vremenske serije pojedinih meteoroloških parametara koji se mogu analizirati u cilju potvrđivanja ili negiranja postojećih klimatskih promena. Za potrebe izrade ovog rada prikupljeni su podaci o mesečnim i godišnjih vrednostima temperatura vazduha i padavina za osmatrački period 1946-2023. godina. Izvršene su analize trenda vremenskih serija. Dobijeni rezultati primenjenih analiza prikazani su u radu.*

Ključne reči: *Temperatura vazduha, padavine, analiza trenda, meteorološka stanica Beograd*

DRAFT VERSION

1. INTRODUCTION

Since 1654, when Ferdinand II de' Medici established the first international network of meteorological stations-Rete Medicea, scientists have strived to standardize measurements in order to compare climate conditions across Europe. This network, under the auspices of the Academy of Experiments, included 11 cities and used uniform thermometers and barometers. In the 18th century, meteorological observations became linked to medical research. In 1776, the Royal Academy of Medicine in France established a network of observers to study the effects of weather on health, while the Palatine Meteorological Society in Germany exchanged data with dozens of institutions across Europe. During the 19th century, meteorological observations became increasingly organized and professionalized. National meteorological services were established in France, Germany, and England, and Matthew Fontaine Maury standardized maritime

measurements. These initiatives led to the founding of the International Meteorological Organization in 1873, marking the beginning of the era of global data exchange and synoptic meteorology (Pepin, 2020; https://catalogo.museogalileo.it/multimedia/ReteMeteorologicaMediceaBis.html).

At the time, Serbia kept pace with Europe. Meteorological measurements in Serbia began as early as 1848, when Vladimir Jakšić regularly recorded weather data in the Senjak area of Belgrade. His diaries span 42 years of observations, with one missing period of ten years. In 1857, Jakšić established a network of 27 meteorological stations — one of the densest in Europe at that time — and authored the first methodological Observer's Manual, published in 1856. The Meteorological Observatory in Belgrade began operations on July 13, 1887, in the Vračar district, and was relocated in 1891 to its current site on Bulevar Oslobođenja. The founder and first director was Professor Milan Nedeljković (Bilak, 2013). The official institution responsible for meteorological observations in Serbia today is the Republic Hydrometeorological Service of Serbia (RHMZ). According to data from the Meteorological Yearbook for 2023, climatological observations were conducted at 55 meteorological stations across the country. These observations included 51 meteorological phenomena and the measurement of 12 meteorological elements.

Climate data collected in the past is crucial for comparing today's climate phenomena with those from before the Industrial Revolution. According to the World Meteorological Organization's Global Annual to Decadal Climate Update (2025–2029), average global temperatures are expected to be 1.2°C to 1.9°C higher than pre-industrial levels (1850–1900). There is also an 80% chance that at least one year will be hotter than the current record holder (2024), and an 86% likelihood that one year will surpass the 1.5°C warming threshold. This trend increases the risk of heatwaves, extreme rainfall, droughts, melting of ice sheets, and rising sea levels (https://wmo.int/publication-series/wmo-global-annual-decadal-climate-update-2025-2029).

2. METHODOLOGY

For the purposes of this study, all available data on average monthly and annual air temperatures, as well as monthly and annual precipitation sums were collected and systematically organized for the period from 1946 to 2023. The study conducted a trend analysis of the collected data for the entire period as well as for selected 30-year intervals (first period: 1961–1990 and second period: 1991–2020), and statistical parameters were calculated for both the entire dataset and the selected intervals. Trend analysis of long-term meteorological data is often used as the simplest analysis that indicates existing climate changes (Hakdar et al., 2023, Rohman and Wiyono, 2022, Mann et al., 2023, Mainuddin et al., 2022, Islam Bhuyan et al. 2018).

3. RESULTS AND DISCUSSION

The intensity of climate change - its very existence and the extent of human influence - continues to be debated both within scientific circles and at the political level. It is evident that the climate of our planet is not static; it has always changed and will continue to do so. Since the planet's formation, these climate changes have occurred in response to natural causes such as atmospheric, oceanic, terrestrial, cryospheric, and biospheric factors (Anđelić, 2012). Unlike the mentioned changes driven by natural forces, those occurring over the last century - often traced back to the Industrial Revolution - can be partially attributed to human influence. Today, when we speak of 'climate change', we refer to temperature variations over the past 100 years caused by human activities rather than natural processes. However, distinguishing between natural climate shifts and human-induced changes remains a challenge, making it difficult to quantify their respective contributions.

Precipitation was analyzed for the period 1946–2023, with monthly and annual total data obtained from the Belgrade meteorological station. Figure 1 presents a graphical depiction of the annual precipitation sums observed at the station, while Table 1 provides statistical parameters for monthly and annual sums over the analyzed period. For the observed period (1946–2023), the average annual precipitation amounted to 692.0 mm. The wettest year was 2014, with a record 1,095.1 mm, while the driest year was 2000, with only 367.7 mm recorded (see Figure 1 and Table 1). The trend in annual precipitation is slightly positive. In addition to total annual precipitation, seasonal precipitation totals were calculated following the same principle. Table 2 contains their statistical summaries, while Figure 2 displays their yearly values. As with total annual precipitation, seasonal trends exhibit negligible slopes - mostly slightly positive, except for summer, where the trend is faintly negative. If trend equations are taken as reference models (see Figures 1 and 2), they can be used to estimate expected precipitation values for specific future time periods (such as 2050 and 2100), as shown in Table 3.

DRAFT VERSION

Table 1: Statistical parameters of average monthly and annual precipitation totals, meteorological station Belgrade, period 1946–2023 (Source of average monthly and annual precipitation totals precipitation totals: Republic Hydrometeorological Service of Serbia)

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	TOTAL
P_{avg}	48.8	44.1	47.3	54.2	73.1	94.0	65.6	52.7	52.4	47.0	55.9	57.1	692.0
St.dev.	26.7	27.8	31.4	25.5	45.2	50.6	48.7	35.9	35.8	35.9	28.8	34.6	133.8
Cv	0.5	0.6	0.7	0.5	0.6	0.5	0.7	0.7	0.7	0.8	0.5	0.6	0.2
Max	112.6	127.8	144.7	157.9	280.4	212.2	262.5	168.1	183.7	184.9	131.5	178.7	1095.1
Min	4.2	2.3	1.8	3.8	8.7	15.4	2.9	4.5	1.0	0.0	5.0	0.8	367.7

Table 2: Statistical parameters of average seasonal precipitation totals, meteorological station Belgrade, period 1946–2023

	Spring	Summer	Autumn	Winter
P_{avg}	174.5	212.2	155.3	150.1
St.dev.	60.6	79.6	60.9	57.1
Cv	0.3	0.4	0.4	0.4
Max	414.4	417.8	346.7	325.0
Min	74.0	56.2	40.2	45.9

Table 3: Trend equations of analyzed total precipitation (annual and seasonal) and calculation of expected total precipitation for time intervals 2050 and 2100

Period	Trend equation	Time intervals		Difference in precipitation over 100 years
		2050	2100	
Total period	P = 0.4173x-135.58	719.9	740.7	41.7
Spring	P = 0.1894x-201.29	187.0	196.4	18.9
Summer	P = -0.0091x+230.24	211.6	211.3	-0.9
Autumn	P = 0.2396x-320.31	170.9	182.8	24.0
Winter	P = 0.0407x+69.261	152.7	154.7	4.1

Based on the results shown in Table 3, it can be concluded that the annual precipitation increase will be only 4.2 mm, meaning that over a 100-year period, this value amounts to 41.7 mm. When distributing this value across seasons, it can be observed that precipitation will increase by 2.4 mm in autumn and 1.9 mm in spring, while winter and summer seasons will see almost no changes. Therefore, based on the obtained analyses, significant variations in total precipitation - both annual and seasonal - should not be expected. What can be concluded from the visualization of these data (Figures 1 and 2) is that extreme values have been more pronounced in the last 20 years.

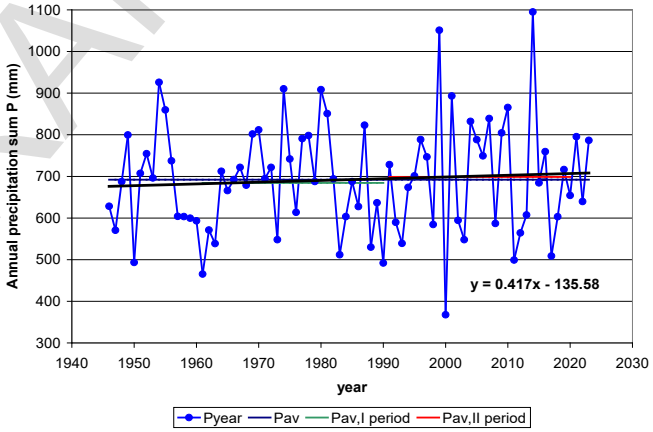
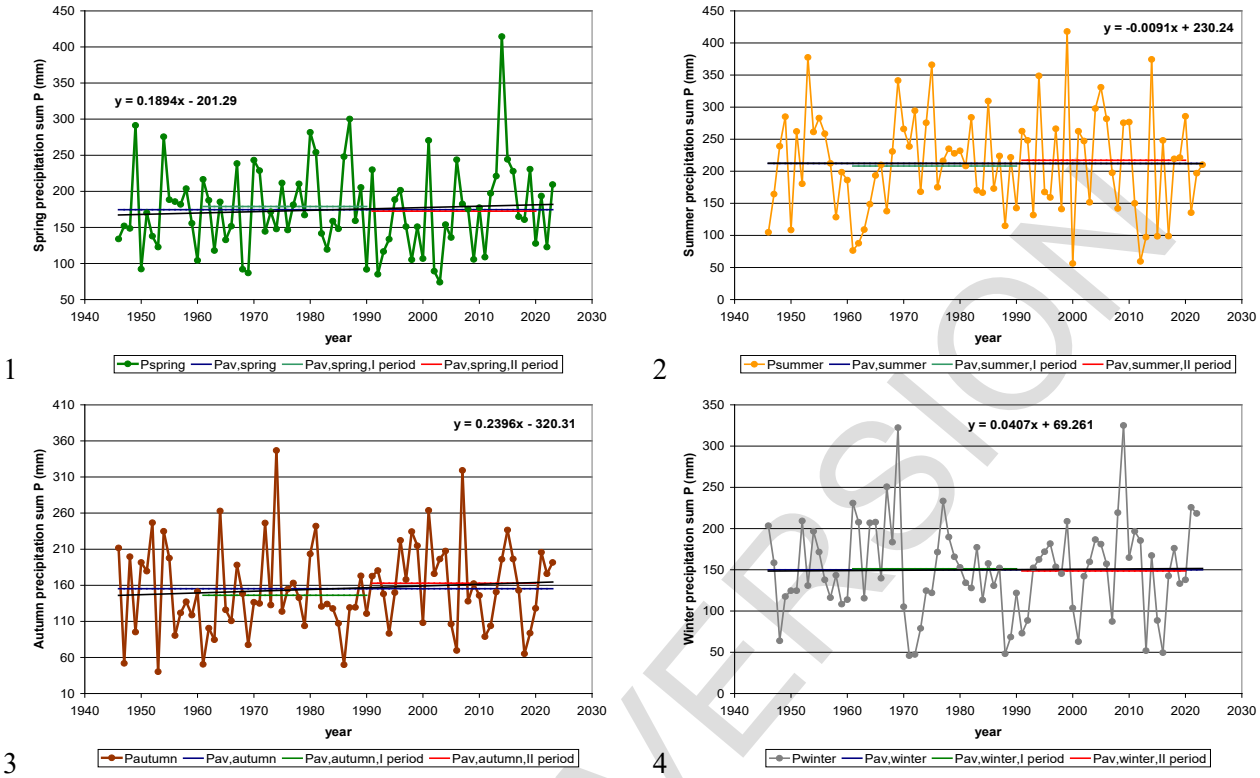


Figure 1: Diagram of annual precipitation totals observed at the meteorological station Belgrade for the observation period 1946–2023.

Average air temperatures were analyzed for the same period, with monthly and annual values obtained from the Belgrade meteorological station. Figure 3 presents a graphical depiction of the average annual air temperatures observed at this station, while Table 4 provides statistical parameters for monthly and annual values over the analyzed period. For the studied period (1946–2023), the average annual air temperature was 12.6°C. The warmest year was 2023, with a record 14.9°C, while the coldest year on average was 1956,

DRAFT VERSION

when the annual mean temperature was 10.5°C (see Figure 3 and Table 4). The trend in annual air temperatures is notably positive. In addition to average annual air temperatures, seasonal averages were calculated following the same principle. Table 5 contains their statistical summaries, while their visualization in the form of a diagram is provided in Figure 4. As with annual air temperatures, seasonal trends show a pronounced positive slope. If trend equations are taken as reference models, they can be used to estimate expected reference values for specific future time periods (2050 and 2100—see Table 6).



1 2 3 4

Figure 2: Diagram of seasonal precipitation totals observed at the meteorological station Belgrade for the observation period 1946–2023; 1 - spring, 2 - summer, 3 - autumn, 4 – winter.

Table 4: Statistical parameters of average monthly and annual air temperatures, meteorological station Belgrade, period 1946–2023 (Source of average monthly and annual air temperatures: Republic Hydrometeorological Service of Serbia)

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	TOTAL
T_{avg}	1.1	3.1	7.5	12.8	17.6	22.5	22.9	22.6	18.3	12.8	7.5	2.9	12.6
St.dev.	2.8	3.3	2.5	1.9	1.7	2.0	1.8	2.0	1.8	1.8	2.3	2.2	1.0
Cv	2.4	1.1	0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.3	0.8	0.1
Max	7.6	9.1	11.8	18.2	21.5	27.0	27.0	26.8	22.6	17.6	12.4	7.0	14.9
Min	-6.3	-7.2	1.4	8.2	13.6	17.5	19.8	18.1	14.1	8.7	1.4	-2.1	10.5

Table 5: Statistical parameters of average seasonal air temperatures, meteorological station Belgrade, period 1946–2023

	Spring	Summer	Autumn	Winter
T_{avg}	12.6	22.7	12.9	2.4
S	1.3	1.7	1.3	2.0
Cv	0.1	0.1	0.1	0.8
Max	15.5	26.8	16.4	6.4
Min	9.7	19.4	10.0	-3.2

Table 6: Trend equations of analyzed air temperatures (annual and seasonal) and calculation of expected values for time intervals 2050 and 2100

Period	Trend equation	Time intervals		Difference in temperatures over 100 years
		2050	2100	

DRAFT VERSION

Total period	T = 0.0304x-47.807	14.5	16.0	3.0
Spring	T = 0.0249x-36.844	14.2	15.4	2.5
Summer	T = 0.0483x-73.192	25.8	28.2	4.8
Autumn	T = 0.0208x-28.472	14.2	15.2	2.1
Winter	T = 0.0340x-65.125	4.6	6.3	3.5

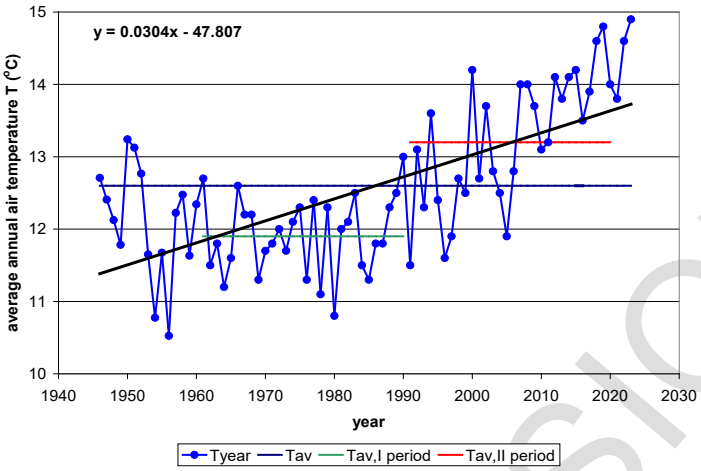


Figure 3: Diagram of average annual air temperatures observed at the meteorological station Belgrade for the observation period 1946–2023.

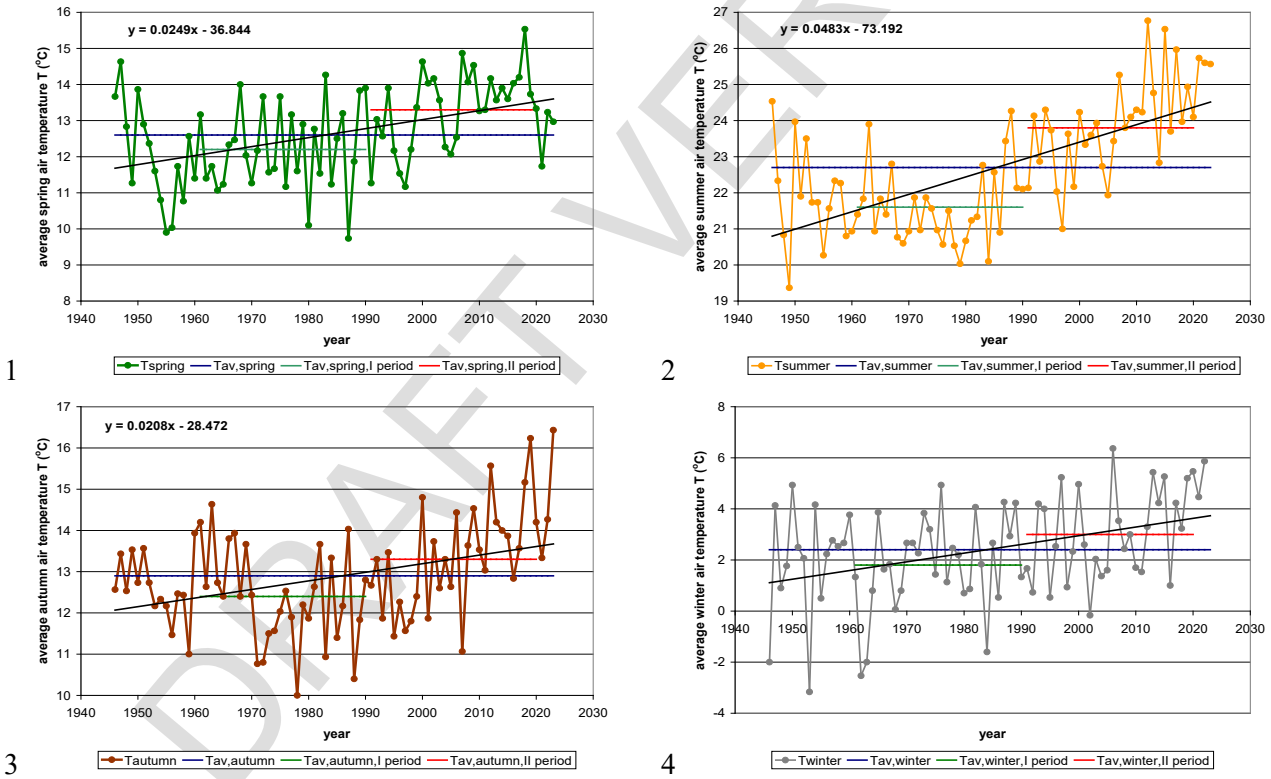


Figure 4: Diagram of average seasonal air temperatures observed at the meteorological station Belgrade for the observation period 1946–2023; 1 - spring, 2 - summer, 3 - autumn, 4 – winter.

Based on the results presented in Table 6, it can be concluded that, over a 100-year period, the temperature increase is concerning, amounting to 3°C annually. A similar trend is observed across seasons, with the highest rise expected during summer (4.8°C) and winter (3.5°C). Spring shows a slightly lower increase of 2.5°C, while autumn exhibits the smallest rise at 2°C. Therefore, based on the conducted analyses, an increase in air temperature is expected both annually and seasonally, with the most significant warming occurring during summer and winter.

DRAFT VERSION

4. CONCLUSION

If the current trends in precipitation and air temperature at both annual and seasonal levels - based on the 78-year analyzed period - continue, it can be concluded that total precipitation will not undergo significant changes by the end of the century. However, extreme weather events will become increasingly frequent and pronounced. Regarding air temperature, it is evident that climate change has led to a rise in temperatures. These ongoing changes indicate a further increase in temperature, which will result in more frequent absence of snow cover. Specifically, groundwater recharge primarily depends on snowmelt and spring precipitation. However, due to rising temperatures and a general lack of snowfall, this component will play a diminished role in groundwater recharge. Considering that over 75% of Serbia's population relies on groundwater for water supply, this issue will become a significant challenge in the near future. A notable example illustrating these concerns occurred in 2024, when water had to be delivered by tankers to livestock grazing on Stara Planina. This was due to the unprecedented drying up of numerous springs that had never run dry before. The primary cause was the absence of snowfall and a prolonged drought during summer, coupled with exceptionally elevated temperatures. Surface flows whose low-water periods depend on groundwater reserves, are also expected to face issues. It is likely that medium and small streams may dry out during summer. Additionally, surface water reservoirs, due to rising temperatures, will become warmer, potentially leading to the growth of cyanobacteria, which could compromise water quality regardless of the intended use of lake water. Belgrade, a city whose water supply system relies heavily on groundwater extraction, will face drinking water shortages if the trend of rising air temperatures continues. In the end, the year 2050 will provide an opportunity for future researchers to reassess trends and determine whether air temperature projections remain on their current strongly positive trajectory.

REFERENCES

Anđelić, M. (2012). Climate change and its impacts: an overview. In Stevanović, Ristić Vakanjac and Milanović (Eds), *CCWaterS Climate Chancegs and Impacts on Water Supply* (pp. 19-50), University of Belgrade, Faculty of Mining and Geology

Museo Galileo. (n.d.). Rete meteorologica medicea. Retrieved 24.7.2025, from https://catalogo.museogalileo.it/multimedia/ReteMeteorologicaMediceaBis.html

Bilak, V.(2013). Meteorološka merenja i osmatranja Vladimira Jakšića u Beogradu (Senjak) od 1848. do 1899. godine. METEOLOGOS. Retrieved 24.7.2025, from https://www.meteologos.rs/meteoroloska-merenja-i-osmatranja-vladimira-jaksica-u-beogradu-senjak-od-1848-do-1899-godine/

Pépin, M.-H. (2025). Meteorological observations over the past centuries. Encyclopedia of the Environment. Retrieved 24.7.2025, from https://www.encyclopedie-environnement.org/en/air-en/meteorological-observations-over-past-centuries-2/

WMO Global Annual to Decadal Climate Update (2025-2029). (2025). World Meteorological Organization. Retrieved 24.7.2025, from https://wmo.int/files/wmo-global-annual-decadal-climate-update-2025-2029

Hakdar, S., Choudlury, M., Choudlury, S., Samanta, P., (2023). Trend analysis of long-term meteorological data of a growing metropolitan city in the era of global climate change, Total Environmental Research Themens, Vol. 7, Elsevier, https://doi.org/10.1016/j.totert.2023.100056

Rohman, R.U.A. Wiyono, (2022). Trend analysis and forecasting of long-term temperature data to anticipate local warming: a case study of Jawa Timur, Indonesia. IOP Conf. Series: Earth Environ. Sci. **1084**(1), 012040 (2022). https://doi.org/10.1088/1755-1315/1084/1/012040

Mann, R., Gupta, A., Dhorde, A., Sharma S. (2023). Observed trends and coherent changes in daily rainfall extremes over greater Mumbai, 1985–2020. Theoret. Appl. Climatol. (2023). https://doi.org/10.1007/s00704-022-04354-4

Mainuddin, M., Peña-Arancibia, J. L., Karim, F., Hasan, M., Mojid, M. A., Kirby, J. M. (2022). Long-term spatio-temporal variability and trends in rainfall and temperature extremes and their potential risk to rice production in Bangladesh. Plos Climate (2022). https://doi.org/10.1371/journal.pclm.0000009

Islam Bhuyan, D., Islam, M., Bhuiyan, E. H., (2018) A Trend Analysis of Temperature and Rainfall to Predict Climate Change for Norhwestt region of Bangladesh, American Journal of Climate Change, Vol.7 No.2, 2018 DOI: 10.4236/ajcc.2018.72009

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

ECOLOGICAL RESPONSIBILITY AND ENVIRONMENTAL ACTIVISM IN KRUŠEVAC: RESIDENTS' ATTITUDES

MILAN ZAJIĆ[1], DUŠAN GARABINOVIĆ[2]

[1] University of Belgrade, Faculty of Organizational Sciences, Belgrade, milan.zajic@fon.bg.ac.rs, 0009-0005-4106-6318
[2] Higher Business School of Vocational Studies "Prof. dr Radomir Bojković", Kruševac, dusan.garabinovic.032@gmail.com, 0000-0002-6247-3060

Abstract: *This paper explores the attitudes of residents of Kruševac towards environmental responsibility, their environmental awareness and education, as well as their behaviour and environmental activism. The aim of the study is to determine the relationship of Kruševac residents to environmental protection as an indispensable element of sustainable development. The results of a survey conducted among 169 residents of Kruševac show that citizens recognize the importance of environmental responsibility and are interested in environmental topics, but there is a lack of confidence in their own awareness and space for improvement in adopting environmentally responsible behaviours. The paper emphasizes the need for further improvement of environmental education and citizen engagement in the local community.*

Keywords: *Environmental responsibility, Environmental awareness and education, Environmental activism, Residents' attitudes.*

DRAFT VERSION

1. INTRODUCTION

Although the term sustainable development is frequently used, its meaning remains insufficiently defined. Ruggerio (2021) argues that despite its significant importance, the definition of sustainable development, and sustainability itself, is imprecise, with a tendency for certain alternatives to the concept to emerge. There's a noticeable increase in the scientific literature regarding the importance of researching sustainable development goals (Mishra et al., 2024). This field is still not consolidated, as many goals are currently insufficiently researched (Yamaguchi et al., 2023).

There is also significant interest in the relationship between ecological services and sustainable development (Ndou & Aigbavboa, 2020). Furthermore, interest in the ecological footprint is growing (Anjum et al., 2024; Petrović et al., 2011).

The involvement of the local community in sustainable development is an increasingly important element in contemporary sustainable development research. Studies have shown that the local community, i.e., its involvement and participation, plays a crucial role in the development of sustainable tourism (Hasana et al., 2022; Iqbal et al., 2022; Narzary & Deb, 2024), and that such involvement leads to a higher level of empowerment, sustainable tourism development, and quality of life (Desalegn & Javed, 2024).

The social component plays an important role in the overall concept of sustainable development. For this reason, there is a significant increase in researchers' interest in the topic of education for sustainable development (Yang & Xiu, 2023). Adequate education for sustainable development forms the basis for the awareness, attitudes, and behaviour of the local community (Zajić et al., 2023).

Contemporary environmental problems necessitate that members of society possess appropriate skills and knowledge to contribute to their solution. For this reason, environmental education plays a vital role in preserving and improving the state of the environment. Formal, non-formal, and informal (experiential) environmental education constitute the complete concept of modern environmental education. Efforts to include environmental education in school curricula exist, but there are also obstacles and challenges that hinder such initiatives. It is crucial to connect these two elements (knowledge and practice/skills) to motivate young people for sustainability-based environmental activism. Awareness, education, and active participation in environmental preservation create the foundation for long-term sustainability achieved through so-called pro-environmental behaviour. One of the indispensable ways of informing today is through promotional activities via the internet, which can impact the sustainable development of protected areas (Ćuruvija et al., 2023).

The subject of this paper is the relationship of residents of the city of Kruševac to environmental protection as an indispensable element of sustainable development. The aim of this paper is to establish the attitudes of residents of Kruševac regarding environmental responsibility, their awareness and education about the environment, as well as their behaviour (habits) and environmental activism.

2. RESEARCH METHODOLOGY

This paper is a continuation of the research, parts of which have been previously presented at the SPIN'23 conference (Zajić et al., 2023). The research was conducted at the beginning of the fourth quarter of 2023, on a sample of 169 respondents. Data were collected through an online survey. The questions were closed-ended and related to the attitudes of Kruševac residents concerning environmental responsibility, residents' awareness and education about the environment, as well as their behaviour and activism regarding the environment. The questions asked in this survey were adapted or used from previous similar studies conducted in Bor (Društvo mladih istraživača Bor, 2020a, 2020b), Minnesota (Murphy, 2004; Murphy & Olson, 2008), and the United States (Coyle, 2005). During the research, all ethical principles were respected, including informed consent of the respondents and the anonymity of their answers.

Females were more numerous (55.0%) than males (45.0%). Most respondents had secondary education (69.8%), followed by higher education (25.4%) and incomplete primary and primary education (4.7%). Most respondents were under 20 years old (54.4%), followed by 20-40 years (24.9%), 40-60 years (16.0%), and over 60 years (4.7%). In terms of place of residence, rural areas (51.5%) were more represented compared to urban areas (48.5%).

3. RESEARCH RESULTS AND DUSCUSSION

Most of Kruševac residents (68.6%) believe that business and industry should bear responsibility for solving environmental problems in their city (Table 1). This high percentage indicates significant awareness among citizens about the potential impact of industrial activities on the environment and their expectation that these sectors actively participate in finding solutions.

Table 1: Responsible for solving environmental problems in the city of Kruševac should be…

	Business and Industry		Kruševac City Administration		Individual Citizens	
	No.	%	No.	%	No.	%
Fully agree	44	26.0	57	33.7	39	23.1
Agree	72	42.6	74	43.8	75	44.4
Disagree	15	8.9	7	4.1	19	11.2
Strongly disagree	1	0.6	3	1.8	7	4.1
Don't know	37	21.9	28	16.6	29	17.2

A smaller percentage of respondents disagree (8.9%) or strongly disagree (0.6%) with the stated claim. This can be interpreted in various ways: (1) They may believe that other actors should play a greater role in solving environmental problems; (2) It is possible that they do not perceive business and industry as the main polluters or believe that business entities are already fulfilling their obligations; (3) There may be scepticism regarding the ability or willingness of businesses to effectively solve environmental problems.

There are both similarities and differences in the attitudes of residents of Kruševac and Minnesota (Murphy, 2004). In both areas, the majority agree that business and industry should be responsible. However, in Kruševac, there is a significantly higher percentage of undecided respondents, indicating a potential lack of information or a greater complexity in perceiving the issue. In Minnesota (Murphy, 2004), on the other hand, there is a higher percentage of those who disagree with the statement. These differences may reflect different socio-economic contexts, regulatory frameworks, historical factors, or levels of awareness about environmental issues.

Most residents of Kruševac believe that the Kruševac city administration should be responsible for solving environmental problems – 77.5% (Table 1). This result was to be expected, given that the city administration is directly responsible for implementing local environmental policies, issuing permits, supervision, and inspection.

A low percentage of respondents with a negative attitude (5.9%) indicates very little resistance to the idea that the city administration should be responsible. Those who disagree may believe that responsibility should be shared with other actors or that the city has other priorities.

DRAFT VERSION

Unlike the results of the study conducted in Minnesota (Murphy, 2004), there is a higher percentage of undecided respondents in Kruševac, indicating a need for better communication between the city administration and citizens. In Minnesota (Murphy, 2004), there is a higher percentage of those who disagree, which may indicate different political views or perceptions of the government's role. Differences in levels of government also play a role in shaping citizens' expectations.

Most residents of Kruševac believe that individual citizens should be responsible for solving environmental problems – 67.5% (Table 1). This result indicates a fairly strong awareness of individual responsibility in environmental matters. This is important because many environmental problems cannot be solved solely at the state, local government, or industry level, but also require a change in individual behaviour and habits.

Comparing the attitudes regarding the responsibility of the city administration, business and industry, and individual citizens for solving environmental problems, it can be concluded that Kruševac citizens believe that responsibility should be shared among different actors, but that the local self-government should bear somewhat greater responsibility than the others. This may be because citizens believe that the city administration has greater power and legitimacy to act in the public interest and to regulate the behaviour of other actors, including businesses and individuals.

Comparing the results in Kruševac and Minnesota (Murphy, 2004), it can be concluded that although both studies show awareness of individual responsibility for environmental problems, there are some key differences: (1) In Minnesota (Murphy, 2004), there is a stronger consensus on this responsibility; (2) In Kruševac, there is a significantly higher percentage of respondents who are undecided on this issue, indicating a potential need for education and awareness-raising.

A high percentage of respondents (Table 2) indicated that they do not have enough information about the state of the environment (43.2%), suggesting: (1) a lack of accessible, clear, and understandable information on environmental issues (or perceived lack thereof) among residents of Kruševac; (2) Insufficient transparency in the work of local authorities regarding environmental problems and their resolution; (3) Insufficient media coverage of environmental topics at the local level; (4) Possible lack of interest among citizens due to insufficient awareness of the importance of environmental problems or a feeling of powerlessness to change anything.

Table 2: Do you feel you are sufficiently informed about the state of the environment in the city of Kruševac and beyond to successfully participate in decision-making?

	No.	%
Excellently informed	35	20.7
Well informed	61	36.1
Not enough information	73	43.2

The analysis of citizens' awareness of the environmental situation shows variations between surveys, with the highest percentage of those who feel insufficiently informed recorded in Bor at the end of 2020 (49.4%) (Društvo mladih istraživača Bor, 2020a), and the lowest in Bor in the second quarter of 2020 (39%) (Društvo mladih istraživača Bor, 2020b). The results for Kruševac (43.2%) place this city between the two survey periods in Bor in terms of the perceived lack of information among citizens.

The largest percentage of respondents get their information through internet websites and social networks – 41.4% (Table 3). Television and radio are traditional media that still have a large reach, but their popularity is declining compared to digital media. A particularly concerning result is that 16% of respondents do not get any information about the state of the environment, which may indicate a significant lack of interest or awareness of environmental issues among that part of the population.

Table 3: How do you usually get information about the state of the environment (circle one answer)?

	No.	%
Through television and radio	35	20.7
Through print media	7	4.1
Through internet websites and social networks	70	41.4
By visiting public forums of NGOs	4	2.4
From friends and neighbours	7	4.1
In another way	19	11.2
Not at all	27	16.0

DRAFT VERSION

Comparing the results in Kruševac and Bor (Društvo mladih istraživača Bor, 2020a, 2020b), it is concluded that the internet is the primary source of environmental information in all studies, but with variations in popularity. Television and radio are still significant, but less so than the internet. Participation in public forums of NGOs is more significant in Bor than in Kruševac. There are various "other" ways in which citizens get information, indicating the complexity of information sources.

The majority of respondents strongly supported the idea that environmental education should be part of the school system – 67.5% (Table 4). This result reflects a growing awareness of the importance of environmental education for creating generations that are aware of the importance of environmental preservation and capable of making responsible decisions.

Table 4: Do you believe that schools should provide environmental education (related to the environment) in elementary and high school?

	No.	%
Yes	114	67.5
No	14	8.3
Don't know	41	24.3

In Kruševac, there is support for environmental education in schools, but it is less than in Minnesota (Murphy, 2004; Murphy & Olson 2008). These results may indicate differences in awareness of environmental issues, educational priorities, or cultural values between Kruševac and Minnesota.

A large majority of respondents (69.8%) stated that they do not know how to name protected natural areas in Kruševac (Table 5). These results indicate a relatively low level of awareness among Kruševac citizens about the existence and location of protected natural areas in Kruševac. This lack of awareness could be a consequence of insufficient promotion of protected areas by competent institutions, a lack of educational programs, or a low level of citizen interest in this topic. All of the above can have negative consequences for the preservation of these areas, as citizens may not be aware of their value and importance, nor of the rules of conduct that must be observed within them.

Table 5: Do you know how to name protected natural areas in the city of Kruševac?

	No.	%
Yes	51	30.2
No	118	69.8

The majority of Kruševac residents (71%) save water to some extent, but there is space for improvement in habits (Table 6). A significant percentage of respondents (15.4%) state that they never save water, which is concerning, as it indicates a lack of awareness or disinterest in water conservation.

Table 6: Do you save water in your daily life?

	No.	%
Never	26	15.4
Sometimes	50	29.6
Often	70	41.4
Don't know	23	13.6

The majority of Kruševac residents (72.2%) use alternative modes of transport at least occasionally (Table 7). The neutral answer (I don't have a car – 16%) is separated because their choice of transport is not necessarily a matter of personal decision, but also of objective circumstances. This percentage indicates a significant portion of the population that is forced to use alternative modes of transport. The negative answer (never – 11.8%) indicates the existence of a group of people who have strong resistance to changing their mode of travel.

DRAFT VERSION

Table 7: In your daily life, do you use other types of transport such as walking, cycling, taking the bus, or carpooling instead of driving alone?

	No.	%
Never	20	11.8
Sometimes	53	31.4
Often	69	40.8
I don't have a care	27	16.0

Positive responses regarding willingness to actively participate in programs and actions aimed at improving the environment ("yes" and "depending on the program") are present in the majority of respondents – 69.3% (Table 8). This result indicates a significant interest among Kruševac citizens in participating in activities aimed at improving the environment. Conditional willingness (depending on the program) means that it is important for programs to be well-designed, in line with their interests and capabilities, and for there to be a clear perception of the benefits of participation. The negative attitude towards participation in environmental actions is relatively low, but it still indicates the existence of a group of citizens who are not interested or may have had negative experiences with similar actions in the past.

Table 8: Are you willing to actively participate in programs and actions to improve the state of the environment?

	No.	%
Yes	54	32.0
No	19	11.2
Depending on the program	63	37.3
Don't know	33	19.5

A comparison of the results from studies (Društvo mladih istraživača Bor, 2020a, 2020b) concerning citizens' willingness to participate in activities to improve the state of the environment shows that in Bor, there is a very high expressed willingness (over 90% combined "Yes" and "Depending on the program"). In Kruševac, this percentage of expressed willingness is lower (around 69% combined), but there is a significant proportion of undecided respondents (around 19.5%), which was not recorded in the Bor studies. This indicates a generally positive attitude towards participation in environmental actions in both areas, but also potential differences in information, the nature of environmental problems, or the programs offered that influence citizens' final decisions.

4. CONCLUSION

The research results indicate a positive attitude towards environmental protection among the residents of Kruševac, but also a need for further improvement of information, education, and citizen engagement.

Future research could focus on a more detailed examination of factors influencing citizens' environmental behaviour, such as socio-demographic characteristics, personal values and attitudes, and perceived barriers to participation in environmental actions. It would be beneficial to conduct similar research in other cities and municipalities across Serbia to gain a more comprehensive picture of the analysed topics at the national level and identify potential regional differences. Research on the effectiveness of various communication and education strategies in raising awareness about environmental issues and encouraging proactive citizen engagement is recommended.

LITERATURE

Anjum, W., Ghufran, B., & Abbas, M. (2024). Ecological footprint: a bibliometric analysis. *Environment, Development and Sustainability*. https://doi.org/10.1007/s10668-024-05463-y.

Coyle, K. (2005). *Environmental literacy in America: What ten years of NEETF/Roper research and related studies say about environmental literacy in the US*. National Environmental Education & Training Foundation, Washington, D.C. https://files.eric.ed.gov/fulltext/ED522820.pdf.

Ćuruvija, M., Ćuruvija, I., & Garabinović, D. (2023). Uticaj internet promotivnih aktivnosti hotela na održivi razvoj zaštićenih područja. U A. Skorup i G. Lapčević (Ur.), *Zbornik radova sa okruglog stola Inovativna Srbija, u Kruševcu 2023. god. Organizatori Visoka poslovna škola strukovnih studija „Prof.*

DRAFT VERSION

dr Radomir Bojković" Kruševac, Biznis inkubator d.o.o. Kruševac i Visoka škola za poslovnu ekonomiju i preduzetništvo Beograd. (str. 69-79). Visoka poslovna škola strukovnih studija „Prof. dr Radomir Bojković", Kruševac.

Desalegn, S. F., & Javed, M. (2024). A Literature Review of Residents' Empowerment, Sustainable Tourism Development, and Quality of Life. In *Proceedings of the 18th International Scientific Conference INPROFORUM. Entrepreneurship for Sustainability. November 7-8, 2024, České Budějovice* (pp. 69-76) University of South Bohemia in České Budějovice, Faculty of Economics, Czech Republic. https://doi.org/10.32725/978-80-7694-102-1.10.

Društvo mladih istraživača Bor (2020a). *Izveštaj o anketi građana o odlučivanju o zaštiti životne sredine*. https://mibor.rs/wp-content/uploads/2020/12/Izvestaj-o-anketi-gradana-o-odlucivanju-o-zastiti-zivotne-sredine.pdf.

Društvo mladih istraživača Bor (2020b). *Izveštaj o rezultatima ankete o stavovima građana Bora o zaštiti životne sredine*. https://mibor.rs/wp-content/uploads/2020/08/Izve%C5%A1taj-o-anketi-BOR-za-za%C5%A1titu-%C5%BEivotne-sredine-rezultati.pdf.

Hasana, U., Swain, S. K., George, B., & Silva, A. (2022). Management of ecological resources for sustainable tourism: A systematic review on community participation in ecotourism literature. Hasana, U., Swain, SK, & George, B.(2022). Management of Ecological Resources for Sustainable Tourism: A Systematic Review on Community Participation in Ecotourism Literature. *International Journal of Professional Business Review*, *7*(1), e0269. https://ssrn.com/abstract=3948162.

Iqbal, A., Ramachandran, S., Siow, M. L., Subramaniam, T., & Afandi, S. H. M. (2022). Meaningful community participation for effective development of sustainable tourism: Bibliometric analysis towards a quintuple helix model. *Journal of Outdoor Recreation and Tourism*, *39*, 100523. https://doi.org/10.1016/j.jort.2022.100523.

Mishra, M., Desul, S., Santos, C. A. G., Mishra, S. K., Kamal, A. H. M., Goswami, S., Kalumba, A. M., Biswal, R., da Silva, R. M., dos Santos, C. A. C.., & Baral, K. (2024). A bibliometric analysis of sustainable development goals (SDGs): a review of progress, challenges, and opportunities. *Environment, Development and Sustainability*, *26*, 11101–11143. https://doi.org/10.1007/s10668-023-03225-w.

Murphy, T. (2004). *The second Minnesota report card on environmental literacy*. Hamline University, Center for Global Environmental Education; Minnesota Office of Environmental Assistance. https://www.lrl.mn.gov/docs/2009/other/090330.pdf.

Murphy, T. P., & Olson, A. M. (2008). *The third Minnesota report card on environmental literacy: A survey of adult environmental knowledge, attitudes and behavior*. The College of St. Catherine; Minnesota Pollution Control Agency. https://www.lrl.mn.gov/docs/2009/other/090329.pdf.

Narzary, R., & Deb, R. (2024). Sustainable Ecotourism Development and Community Involvement: A Bibliometric Analysis on Research Trends (2003–2024). *Journal of Tourism, Hospitality & Culinary Arts*, *16*(3), 56-70.

Ndou, M., & Aigbavboa, C. (2020). Sustainable Development of Mankind Through Ecological Services: A Literature Review. In L. Scott, M. Dastbaz, C. Gorse (Eds) *Sustainable Ecological Engineering Design* (pp. 437-450). Springer, Cham. https://doi.org/10.1007/978-3-030-44381-8_33.

Petrović, N., Išljamović, S., Jeremić, V., Vuk, D., & Senegačnik, M. (2011). Ekološki otisak kao indikator nivoa ekološke svesti studenata Fakulteta organizacionih nauka Univerziteta u Beogradu i Mariboru. *Management - časopis za teoriju i praksu menadžmenta*, *16*(58), 15-21.

Ruggerio, C. A. (2021). Sustainability and sustainable development: A review of principles and definitions. *Science of the Total Environment*, *786*, 147481. https://doi.org/10.1016/j.scitotenv.2021.147481.

Yamaguchi, N. U., Bernardino, E. G., Ferreira, M. E. C., de Lima, B. P., Pascotini, M. R., & Yamaguchi, M. U. (2023). Sustainable development goals: a bibliometric analysis of literature reviews. *Environmental Science and Pollution Research*, *30*, 5502-5515. https://doi.org/10.1007/s11356-022-24379-6.

Yang, C., & Xiu, Q. (2023). A Bibliometric Review of Education for Sustainable Development, 1992–2022. *Sustainability*, *15*(14), 10823. https://doi.org/10.3390/su151410823.

Zajić, M., Višnjić, A., Jovanović, T., & Bogdanović, D. (2023). Ispitivanje nivoa ekološke svesti kod stanovnika Kruševca. In N. Petrović, & M. Ćirović (Eds), *Zbornik radova, Digitalni i zeleni razvoj privrede, XIV Skup privrednika i naučnika SPIN'23, 6.-7. novembar 2023., Beograd, Srbija* (pp. 276-285). Univerzitet u Beogradu, Fakultet organizacionih nauka, Centar za operacioni menadžment.

DRAFT VERSION

FINANCE AND BANKING

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Review article

ECONOMICS OF HAPPINESS: THE RELATIONSHIP BETWEEN GDP AND SUBJECTIVE WELL-BEING

DRAGANA DRAGANAC[1], IVANA IVKOVIĆ[2],

[1] University of Belgrade, Faculty of Economics and Business, Belgrade, dragana.draganac@ekof.bg.ac.rs, 0000-0002-8605-8881
[2] University of Belgrade, Faculty of Economics and Business, Belgrade, ivana.ivkovic@ekof.bg.ac.rs, 0000-0002-6804-5200

Apstrakt: *U ovom radu bavimo se ekonomijom sreće, sa posebnim fokusom na odnos između sreće i materijalnog bogatstva, sa ciljem da ispitamo postojanje Eastelinovog paradoksa. Analiza se zasniva na podacima iz Svetskih izveštaja o sreći, pri čemu se posmatra korelacija između sreće i bruto domaćeg proizvoda (GDP) po glavi stanovnika na međudržavnom nivou. Dodatno, posmatramo kretanje GDP per capita i skora sreće longitudinalno u okviru jedne zemlje. Dobijeni rezultati samo u određenoj meri idu u prilog Easterlinovom paradoksu. Naime, na međudržavnom nivou, sa rastom bogatstva do određenog nivoa dolazi do porasta sreće, nakon čega se javlja zasićenje. U slučaju Srbije, sa porastom GDP per capita, najpre dolazi do rasta sreće, nakon čega se uočava stagnacija. Ovi rezultati mogu poslužiti kao osnova za kreiranje mera ekonomske i javnih politika usmerenih ka unapređenju individualnog i društvenog blagostanja.*

Ključne reči: *ekonomija sreće, sreća, blagostanje, Easterlinov paradoks, GDP per capita.*

Abstract: *This paper explores the field of economics of happiness, with a particular focus on the relationship between happiness and material wealth, aiming to examine the validity of the Easterlin Paradox. The analysis is based on data from the World Happiness Reports, estimating the correlation between happiness and gross domestic product (GDP) per capita at the cross-country level. Additionally, we are examining the movement of GDP per capita and happiness scores longitudinally within a single country. To some extent, the findings support the Easterlin Paradox: namely, at the international level, as wealth increases, happiness initially rises to a certain threshold, after which saturation occurs. In the case of Serbia, with the rise in GDP per capita, firstly there is an increase in happiness, after which stagnation is observed. These results may serve as a foundation for designing economic and public policy measures aimed at improving individual and societal well-being.*

Keywords: *Economics of Happiness, Happiness, Well-being, Easterlin Paradox, GDP per capita.*

DRAFT VERSION

1. INTRODUCTION

Welfare economics is a branch of economic theory that studies how the allocation of resources affects individual and societal well-being and identifies changes in allocation that may lead to improvements in overall welfare. Neoclassical welfare economics assumes that individuals are rational agents who make choices according to their preferences. Well-being is measured indirectly, through the choices people make, under the assumption that *Choice = Preferences = Welfare*. In other words, individuals make choices based on their preferences, and those choices generate a certain level of well-being. Public policies are formulated with minimal intervention, following the *laissez-faire* principle.

In the aftermath of the Global Financial Crisis of 2007, behavioral welfare economics began to develop rapidly, incorporating insights from psychological research into actual human behavior. Behavioral welfare economics accounts for emotions, psychological factors, and cognitive limitations of economic agents. Instead of the traditional notion of preferences, the concept of *informed preferences* began to be used (Chetty, 2006). This implies that welfare is not measured solely based on the choices individuals make, but also on what they would choose if they were informed, attentive, and consistent. The concept of informed preferences acknowledges that people make systematic errors in decision-making.

Bernheim and Rangel (2007) developed a theoretical foundation for welfare analysis grounded in the choice theory, adapted to cases in which decision-makers are not consistent and make errors in judgment. They

introduced the concept of a *normative standard* of behavior, a criterion that allows us to evaluate what is genuinely in an individual's best interest and, accordingly, design public policies—even when individual choices deviate from the rational model.

Within the framework of *libertarian paternalism*, Thaler and Sunstein (2003), and later Sunstein and Thaler (2008) argue that for any given choice situation, there exists a better choice for the individual—one that leads to a higher level of actual well-being. If the individual is unable or unmotivated to make that choice, a *choice architect* can assist through subtle interventions known as *nudges*, encouraging better decisions without restricting freedom of choice. Thus, public policies are based on *soft paternalism*.

Economics of happiness is a relatively new scientific field that emerged within behavioral welfare economics. It proposes that economic policy measures should be designed to maximize the happiness of individuals and society as a whole. Defining and measuring happiness remains a significant challenge, given the abstract nature of the phenomenon. In this paper, we focus on the happiness score, a measurable and annually published indicator. We aim to assess whether greater wealth necessarily translates into greater happiness.

This paper is structured in the following way. In the second section, we present neoclassical and behavioral indicators of well-being. The third section offers a literature review. In the fourth section, the research results and discussion are provided. At the end, we conclude the paper, providing an overview of the key implications and insights.

2. INDICATORS OF WELL-BEING

Indicators of well-being are metrics used to measure the level of well-being. Well-being is not measured solely by economic wealth, but also includes various life dimensions — education, health, environmental quality, safety, freedom, and the subjective experience of happiness.

The main indicators of well-being within neoclassical welfare economics are Gross Domestic Product (GDP) and the Human Development Index (HDI). The Gini coefficient is used as an indicator of inequality and may be considered a partial indicator of well-being.

Gross Domestic Product (GDP) represents the total market value of all goods and services produced within a national economy over a specific period, typically one year. As an indicator of well-being, GDP has several advantages. It is a simple and widely accepted measure that is easy to calculate. Using GDP, we can easily make comparisons between countries and track economic growth. Economic policies are often formulated and revised based on GDP trends.

However, GDP has its limitations. It provides only a rough estimate of living standards and does not account for income distribution within a country. Thus, it fails to reflect inequality and poverty. It also excludes the informal and shadow economy (e.g., undeclared labor and household work). GDP measures national wealth only quantitatively, without capturing quality of life or happiness. Furthermore, it neither accounts for the sustainability of economic growth nor incorporates negative externalities.

The Human Development Index (HDI) was developed to address the shortcomings of GDP. Unlike GDP, HDI measures how healthy, educated, and long-living people are. It was introduced by the United Nations Development Programme (UNDP) in 1990 and is calculated as the geometric mean of three indices: the health index, the education index, and the income index:

$$HDI = \sqrt[3]{Health\ Index\ *Education\ Index\ *Income\ Index} \quad (1)$$

The health index is based on life expectancy at birth and reflects the general health, longevity, and access to healthcare within a society. The education index considers the average number of years of formal education among adults and the expected years of schooling for children entering the education system. The income index is based on gross national income per capita (GNI per capita), expressed in purchasing power parity (PPP) dollars, which enables cross-country comparisons regardless of price differences.

The main indicators of well-being within behavioral welfare economics are the Happiness Score published in the World Happiness Report (WHR), and Gross National Happiness measure (GNH). Subjective well-being is measured directly through surveys in which respondents answer questions about life satisfaction. These surveys serve as data sources for the Happiness Score. Gross National Happiness (GNH) is a concept first introduced in Bhutan, encompassing not only economic development but also cultural values, environmental sustainability, and spiritual well-being.

This paper focuses on the Happiness Score and the field of Economics of happiness. Richard Easterlin is considered the founder of economics of happiness, with his seminal paper from 1974 laying the foundation for this field (Easterlin, 1974). Other influential scholars include Kahneman, Deaton, Frey, Stutzer, and Oswald (Kahneman & Deaton, 2010; Frey & Stutzer, 2002; Oswald, 1998).

DRAFT VERSION

The World Happiness Report is an annual survey published by the United Nations Sustainable Development Solutions Network (UN SDSN) in collaboration with experts in economics, psychology, and statistics. First published in 2012, it has since served as a key resource for comparing levels of happiness and quality of life across countries worldwide.

A major advantage of the happiness score and the World Happiness Report is that they do not consider only economic growth and material wealth as sources of happiness, but also incorporate a wide range of other factors, such as healthy life expectancy, social support, and freedom. As such, they enable the formulation of economic and public policies oriented toward human well-being rather than solely GDP. The main limitation of the happiness score is that categories such as social support, freedom, and perceived corruption are subjective and difficult to measure. Surveys serve as the primary method for collecting such data and come with inherent limitations, including socially desirable responses, reliance on respondents' honesty, limited information depth, sample bias, and low response rates, as well as high variability in emotional states.

In addition to surveys, experimental method is also used—especially to test whether a proposed economic policy would have the desired effect in practice. The experimental method is particularly useful for establishing causal relationships. It is crucial to ensure high internal and external validity of experiments to confirm that true causal relationships have been identified and that the findings can be generalized from the sample to the population, and from specific conditions to broader contexts. In randomized controlled trials, a randomly selected group (treatment group) is exposed to the behavioral intervention, while a control group is not. The difference in happiness between the two groups is then assessed.

3. LITERATURE REVIEW

Empirical research has shown that within a country, increases in income from poverty to a decent standard of living do raise people's happiness. However, once income exceeds the threshold of basic comfort and enters the realm of luxury and abundance, additional money no longer leads to increased happiness, and in some cases, it may even reduce life satisfaction. This pattern is known as the Easterlin Paradox (Easterlin, 1974). Specifically, within a country, higher income is associated with greater happiness up to a certain income level, but over time, increases in average income do not lead to increases in average happiness. When comparing countries, wealthier nations are not necessarily significantly happier than poorer ones. Over time, as countries grow economically, happiness levels stagnate due to adaptation, social comparison (relative income), and changing values and expectations. It can thus be concluded that economic growth does not guarantee long-term increases in subjective well-being—contrary to the traditional economic assumption that higher GDP automatically implies greater welfare.

Clark et al. (2008) find that relative income, i.e. income compared to others, significantly affects happiness. People care more about their income rank than the absolute amount. Oswald (1997) and Clark (2001) show that unemployment has a profound and lasting negative impact on happiness, greater than the effect of income level. Alesina et al. (2004) finds that higher income inequality reduces happiness, especially among the poorer population. These findings have important policy implications: economic and public policies should consider not just the level of GDP, but also other key factors affecting well-being and happiness—such as freedom, inequality, corruption, and social support.

Easterlin et al. (2010) re-examine the Easterlin Paradox and concluded that it holds in both developed and developing countries. Stevenson et al. (2008) find a positive relationship between GDP per capita and happiness both across countries and over time within a single country. In contrast, Deaton et al. (2008) observe that the income-happiness link exists globally but weakens within wealthier countries. Kahneman & Deaton (2010) distinguish between evaluative and emotional happiness: evaluative happiness refers to life satisfaction, while emotional happiness includes day-to-day experiences. They found that in the United States, income above $75,000 no longer increases emotional happiness, although it continues to improve evaluative well-being.

4. RESULTS AND DISCUSSION

For the purposes of this analysis, we used data from the World Happiness Report spanning the years 2016 to 2023 (https://data.worldhappiness.report/map, accessed on May 26, 2025). The data, collected via the Gallup World Poll, cover approximately 150 countries, representing around 98% of the global adult population. The survey includes several categories of questions: overall life satisfaction and happiness (a subjective assessment of life quality), social support, GDP per capita, healthy life expectancy, inequality, freedom of choice, generosity, perceptions of corruption, positive and negative emotions, and altruism—including

DRAFT VERSION

donations, volunteering, and helping strangers. Some of the questions include: "Did you experience enjoyment/worry/sadness yesterday?" and "Have you donated money in the past month?"

Based on these factors, a composite score is calculated, and countries are ranked accordingly. The Happiness Score ranges from 0 to 10. For the year 2023, data were available for 139 countries. The countries with the highest Happiness Score include Finland, Denmark, Iceland, Sweden, and the Netherlands—nations characterized by high living standards, strong social support, low corruption, and a favorable work-life balance. Conversely, the lowest-ranked countries include Afghanistan, Sierra Leone, Lebanon, Malawi, and Zimbabwe. It is noteworthy that Finland has held the top position for eight consecutive years.

To examine the presence of the Easterlin Paradox at the cross-country level, we calculated the Spearman correlation coefficient between GDP per capita and the Happiness Score for all countries in 2023. This coefficient amounts to 0.86, indicating a strong positive correlation between national wealth and subjective well-being. The corresponding scatterplot is presented in Figure 1.

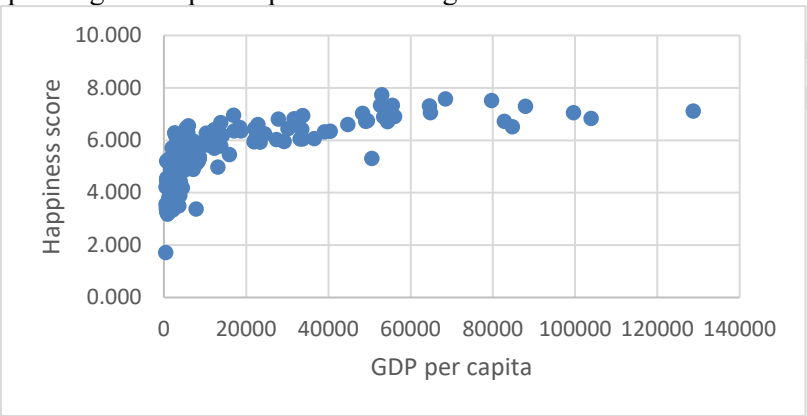


Figure 1: Scatterplot of Happiness Score vs. GDP per Capita

As GDP per capita increases, the happiness score initially rises. However, beyond a certain threshold—where comfortable living transitions into luxury—further increases in wealth is no longer followed by the increase in happiness. This plateau suggests the presence of a saturation effect. While this may indicate the existence of the Easterlin Paradox, further analysis is required.

To explore whether the correlation coefficient changes at different income levels, we divided all countries into quintiles based on GDP per capita, and calculated the correlation coefficient for each quintile. The results are shown in Table 1.

Table 1: Correlation Coefficients by GDP Quintiles

GDP quintiles	Correlation Coefficients
1	0,408
2	0,103
3	-0,082
4	0,362
5	0,519

Correlation coefficients decline from the first to the third quintile, with the third quintile even exhibiting a negative correlation, implying that in these countries, higher GDP per capita is associated with lower happiness scores. This provides support for the Easterlin Paradox. In the fourth and fifth quintiles, however, the correlation becomes positive and strengthens, which contradicts Easterlin paradox.

To check whether Easterlin Paradox exists within a single country, it is necessary to assess whether long-term increases in GDP per capita correspond to significant rises in average happiness. We examined this relationship in the case of Serbia, analyzing data from 2016 to 2023. The charts of GDP per capita and happiness score for the period 2016 to 2023 are given in Figure 2.

DRAFT VERSION

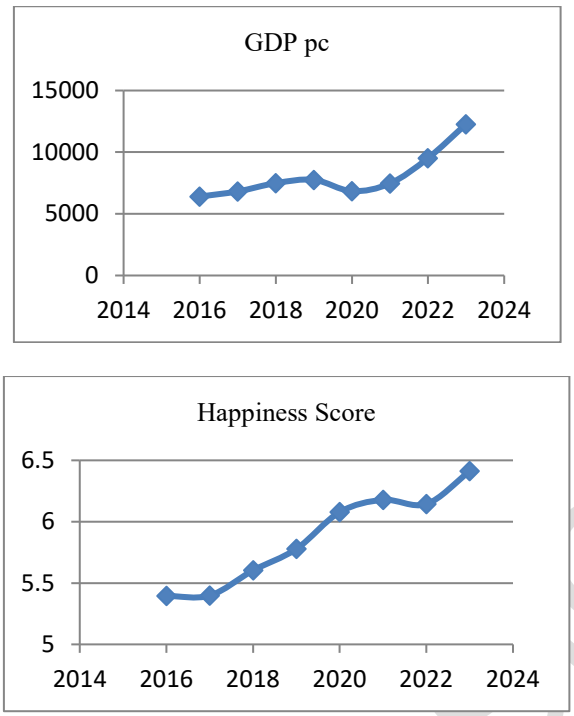


Figure 2: GDP per Capita and Happiness Score for Serbia, 2016-2023

It can be observed that GDP per capita rises in the period 2016-2019. In 2020, there is a decline and then, in the following years, it continues to rise. On the other hand, Happiness Score increase in all years but 2022. However, its increase in years 2020-2022 is not that intense. Further analysis is necessary to check if Easterlin paradox is present.

5. CONCLUSION

Although GDP and other neoclassical indicators of welfare remain relevant for shaping economic policies, there is a growing recognition of the need to incorporate subjective well-being indicators and measures based on informed preferences. The fields of behavioral welfare economics and happiness economics offer both theoretical and empirical frameworks for understanding the complex determinants of human welfare that go beyond material indicators—such as social support and freedom of personal choice.

High happiness levels in Nordic countries are not solely attributable to high GDP levels, but rather to a synergistic combination of economic, social, psychological, and institutional factors.

At the global cross-country level, we find a strong positive correlation between GDP per capita and subjective happiness, indicating that higher wealth of countries is associated with higher happiness. However, detailed analysis across GDP quintiles reveals that this relationship is not linear. In lower and middle quintiles, the correlation weakens and, in the third quintile, even becomes negative, indicating a saturation effect in the income-happiness relationship. These findings lend support to the Easterlin Paradox and related concepts such as relative income, adaptation, and shifting expectations. Nevertheless, in the fourth and fifth quintiles, the correlation rises again, which runs counter to the Easterlin hypothesis. An analysis of long-term trends in Serbia further explores the relationship between wealth and happiness at the national level.

The findings of this study carry significant policy implications. Economic and public policies should not focus solely on promoting economic growth but should also aim to improve quality of life, strengthen social cohesion, enhance access to healthcare and education, and foster environments that empower individuals to make choices conducive to long-term well-being. Policies grounded in such a holistic approach are more likely to support sustainable development and bring about genuine improvements in individual and collective welfare.

ACKNOWLEDGMENT

Dragana Draganac would like to acknowledge the financial support of the Ministry of Education, Science and Technological Development of the Republic of Serbia, project No. MPNTR41030. Ivana Ivković would like to acknowledge the financial support of the Ministry of Education, Science and Technological Development of the Republic of Serbia, project No. 179050.

DRAFT VERSION

LITERATURE

Alesina, A., Di Tella, R., & MacCulloch, R. (2004). Inequality and happiness: Are Europeans and Americans different? *Journal of Public Economics*, 88(9–10), 2009–2042. https://doi.org/10.1016/j.jpubeco.2003.07.006

Bernheim, B. D., & Rangel, A. (2007). *Behavioral public economics: Welfare and policy analysis with non-standard decision-makers*. In P. Diamond & H. Vartiainen (Eds.), *Behavioral Economics and Its Applications* (pp. 7–77). Princeton University Press.

Chetty, R. (2006). A new method of estimating risk aversion. *American Economic Review*, 96(5), 1821–1834. https://doi.org/10.1257/aer.96.5.1821

Clark, A. E. (2001). Unhappiness and unemployment: A panel data analysis for Europe. *Economic Journal*, 111(473), 641–658. https://doi.org/10.1111/1468-0297.00648

Clark, A. E., Frijters, P., & Shields, M. A. (2008). Relative income, happiness, and utility: An explanation for the Easterlin paradox and other puzzles. *Journal of Economic Literature*, 46(1), 95–144. https://doi.org/10.1257/jel.46.1.95

Deaton, A. (2008). Income, health, and well-being around the world: Evidence from the Gallup World Poll. *Journal of Economic Perspectives*, 22(2), 53–72. https://doi.org/10.1257/jep.22.2.53

Easterlin, R. A. (1974). *Does economic growth improve the human lot? Some empirical evidence*. In P. A. David & M. W. Reder (Eds.), *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz* (pp. 89–125). Academic Press.

Easterlin, R. A., McVey, L. A., Switek, M., Sawangfa, O., & Zweig, J. S. (2010). The happiness–income paradox revisited. *Proceedings of the National Academy of Sciences*, 107(52), 22463–22468. https://doi.org/10.1073/pnas.1015962107

Frey, B. S., & Stutzer, A. (2002). *Happiness and economics: How the economy and institutions affect human well-being*. Princeton University Press.

Kahneman, D., & Deaton, A. (2010). High income improves evaluation of life but not emotional well-being. *Proceedings of the National Academy of Sciences*, 107(38), 16489–16493. https://doi.org/10.1073/pnas.1011492107

Oswald, A. J. (1997). Happiness and economic performance. *Economic Journal*, 107(445), 1815–1831. https://doi.org/10.1111/j.1468-0297.1997.tb00085.x

Oswald, A. J. (1998). Commentary on the Easterlin Paradox. *Journal of Economic Behavior & Organization*, 37(1), 211–217.

Stevenson, B., & Wolfers, J. (2008). Economic growth and subjective well-being: Reassessing the Easterlin paradox. *Brookings Papers on Economic Activity*, 2008(1), 1–87. https://doi.org/10.2139/ssrn.1143221

Sunstein, C. R., & Thaler, R. H. (2003). *Libertarian paternalism is not an oxymoron*. University of Chicago Law Review, 70(4), 1159–1202.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

United Nations Development Programme (UNDP). (1990). Human Development Report 1990. Oxford University Press.

World Happiness Report. (2012–2024). World Happiness Report. UN Sustainable Development Solutions Network. https://data.worldhappiness.report/map, accessed May 26, 2025.

DRAFT VERSION

GAME THEORY

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

MORAL HAZARD IN THE 2008 FINANCIAL CRISIS

ĐORĐE MIRKOVIĆ[1], OGNJEN NIKOLIĆ[2], MARIJA KUZMANOVIĆ[3]

[1]University of Belgrade, Faculty of Organizational Sciences, Belgrade, djordjemirkovic001@gmail.com, 0009-0007-2232-8428
[2]University of Belgrade, Faculty of Organizational Sciences, Belgrade, ognjen.nikolic@fon.bg.ac.rs, 0000-0003-0756-4290
3University of Belgrade, Faculty of Organizational Sciences, Belgrade, marija.kuzmanovic@fon.bg.ac.rs, 0000-0001-5857-6932

Abstract: *Moral hazard refers to a situation where an actor takes excessive risks without bearing the consequences. It is particularly harmful in principal-agent relationships, where the agent exposes the principal to greater risk. Such patterns were evident in the financial system during the emergence and peak of the 2008 crisis. The aim of this paper is to model these interactions and assess the extent to which moral hazard contributed to the crisis, with emphasis on securitization, banks that are "too big to fail", and conflicts of interest of credit rating agencies. The methodology involves three dynamic Bayesian game models: the relationship between lenders and borrowers, between investment banks and the state, and among banks, rating agencies, and investors. The conclusion is that, although moral hazard played a significant role, it was not the sole cause of the crisis. Combined with weak regulation and the complexity of the financial system, it highlights the need for deeper examination to prevent similar issues in the future.*

Keywords:*, moral hazard, 2008 financial crisis, game theory, banking, loans.*

1. INTRODUCTION

In mid-2007, the economy of the United States and other major world economies stood on the brink of the worst financial crisis since the Great Depression of the 1930s. As in that earlier period, this crisis led to the collapse of major financial institutions, a deep global recession, and drastic measures by governments around the world to mitigate the economic consequences of the financial system's collapse (Rose & Marquis, 2011).

The key causes of this crisis included the process of mortgage loan securitization, the deregulation of the financial market, the creation of extremely complex financial derivatives, and the belief that the real estate market would never decline. However, among the most important causes, moral hazard stands out in particular. Moral hazard is a situation in which actors take greater risks because they do not bear the consequences of their decisions. This problem is the result of information asymmetry, a scenario in which one participant in the game possesses hidden information, creating room for suboptimal and potentially harmful outcomes. A common example of moral hazard can be seen in the health insurance market, where insured individuals are more prone to riskier behavior that increases the costs of the healthcare system. Moral hazard manifested itself on multiple levels during the 2008 financial crisis, producing outcomes harmful to society (Murray et al., 2017). During the financial crisis, moral hazard appeared on several levels: lenders securitized mortgage loans, packaging them into securities and selling them to investment banks. Lenders no longer bore the risk of mortgage default and thus approved loans to risky borrowers. Investment banks believed that the state would bail them out in the event of collapse, which encouraged excessive risk-taking. The bankruptcy of Lehman Brothers, for example, triggered a domino effect. Credit rating agencies, due to their dependence on banks, assigned excessively high ratings to risky products, further reinforcing risky behavior by banks.

The aim of this paper is to determine to what extent moral hazard contributed to the financial crisis. This will be done by modeling the relationships between actors during the crisis using game theory, with a focus on three dynamic Bayesian games: the lender–borrower model, examining how lenders extended credit to risky clients, the investment bank–state model, explaining why banks were prone to excessive risk, and the investment bank–credit rating agency–investor model, analyzing how conflicts of interest in rating agencies encouraged risk. The next section will briefly explain the basic concepts of game theory that will serve as the theoretical foundation for the rest of the paper.

DRAFT VERSION

2. MORAL HAZARD ANALYSIS

The first step in analyzing moral hazard is to identify the actors and their interactions within the financial system. This paper will use Ferguson's model (**Figure 1**) of player interactions in the financial system, which he calls the securitization food chain. Ferguson describes this model as "an industrial supply chain for generating mortgage loans and selling them to investment banks. The investment banks then package these mortgages into structured investments that they sell to various investors, including pension funds, hedge funds, and other institutional investors. This supply chain integrates virtually every segment of the financial system" (Ferguson, 2012).

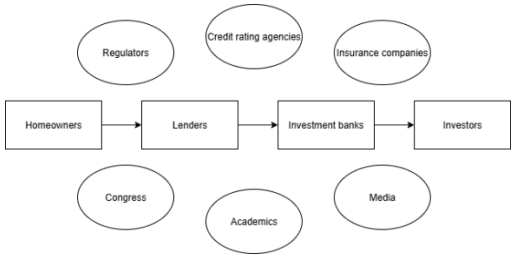


Figure 1: Ferguson securitization food chain (Ferguson, 2012)

Since the aim of this paper is to identify moral hazard, the interactions described above will be used to construct games through which the impact of moral hazard will be analyzed. The proposed models are as follows: the borrower–lender interaction model, the state–investment bank interaction model; and the investment bank–credit rating agency–investor interaction model.

The reason these interactions are not presented within a single model lies in the complexity of such a model. Moreover, while the payouts of one actor's decisions are influenced by the actions of all others, the players in these models exert the greatest influence on each other. Now, the players and their relationships with the actors in the real system will be briefly explained, and then the individual models will be presented.

Homeowners are individuals or families who take out mortgage loans from lenders. Although in practice there are different types of lenders, in this paper they are viewed as a homogeneous unit whose only role is to grant loans. Investment banks buy these loans, package them into securities and sell them to investors, while credit agencies assess the risk of these securities. In the model, the state can bail out the bank to prevent its bankruptcy, and investors (pension funds, individuals, banks) generally choose low-risk instruments

3. STATE AND INVESTMENT BANK INTERACTION MODEL

The model in **Figure 2** shows the interaction between banks (player B) and the state (player D) through a staged dynamic Bayesian game lasting an unlimited number of periods, and **Table 1** symbols used in the model. At the beginning of the game, "nature" determines the type of state with probability p: whether the state prefers to save the bank (Type 1 – Save) or not (Type 2 – Let Fail). Then the bank chooses the risky strategy (R), which increases the chances of higher profit but also of failure, or safe (S), which reduces risk. It is important to note that the probability of bank failure also depends on the risk of other banks due to their interconnectedness. After the bank chooses a strategy, "nature" decides whether the bank fails or not. If the bank survives, the game continues to the next period T+1, and the game information is updated. If the bank fails, the state decides whether to save it (strategy S) or let it fail (strategy P).

The game ends in the event of a crisis, regardless of the state's decision, because the bank either disappears or is under significant state control. The reason for introducing the "nature" player is that this is a signaling game in which the state wants to signal that it will not always save banks to discourage them from excessively risky behavior. Through this game, banks form beliefs about whether they will be saved, which is based on the long-term reputation of the state. Bank owners aim for higher profits through dividends but must ensure the bank's survival, because in the event of failure without state intervention they lose their invested capital. When deciding on a bailout, the state weighs the economic consequences, bailout costs, political costs, and impact on its reputation. Frequent bailouts encourage banks to take greater risk, which is a key aspect of this dynamic interaction.

DRAFT VERSION

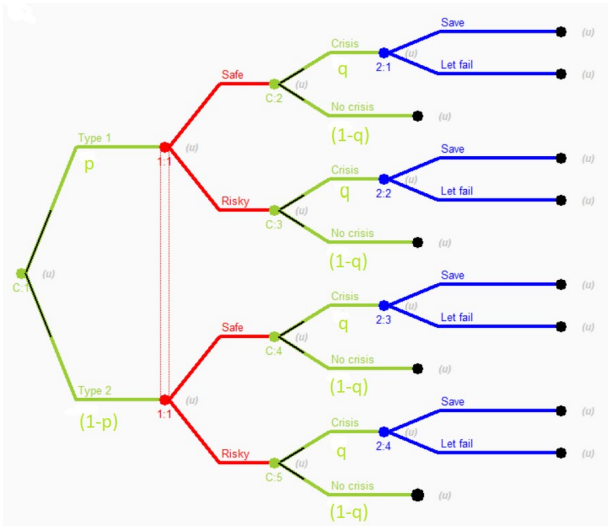


Figure 2: Bank-state interaction model

Table 1: Symbols for Bank-state interaction model

R	Risky strategy
S	Safe strategy
p	Probability State is type 1
q	Probability of crisis
A	Bank payout if fail
B	Bank payout if it doesn't fail
U	Total Bank payout
V	Total State payout
A1	Bank payout if it's saved
A2	Bank payout if State let's it fail
B(R)	B if Bank played R
B(S)	B if Bank played S

Moral hazard in this model represents a situation where banks behave risky because they know the state will bail them out. The equilibria leading to this situation are (Risky, Save) and (Risky, Let Fail) because regardless of the state's decision, there will be negative consequences. The question is how high the probability p must be to avoid this equilibrium. The general formula for the bank's total payoff in one stage is:

$$D = q * A + (1 - q) * B \quad (1)$$

The bank's payoff is calculated by multiplying the probability of bank failure by the respective payoffs. $B(R)$ is greater than $B(S)$, but q also increases when playing R. The higher the q, the greater the share of A in the final payoff. It is important to note that B is always greater than A. This is because B represents the bank's payoff in normal operations, while A represents the scenario where the bank fails. Even if the bank is bailed out, the bank owners will not profit from it but will simply avoid losing money. From this, it can be concluded that the bank can play the risky strategy if the value of q is not high enough for A to start reducing the total payoff. The formula for the bank's payoff in case of failure:

$$A = p * A1 + (1 - p) * A2 \quad (2)$$

The bank's payoff in case of failure depends on the type of state and is calculated by multiplying payoffs with the corresponding probabilities. If the state is Type 1, it prefers to save the bank, so the $A1$ is higher. Conversely, if the state is Type 2, it does not prefer to save, which results in a large loss of $A2$, which may be a negative value. The value of $A1$ is lower than any value of B (payoff in case of survival), but $A2$ is a huge loss. Therefore, for the expected value of A to be acceptable, the probability that the state is Type 1 must be very high. The conclusion is that banks' belief that they will be saved leads to a worse equilibrium because it encourages risky behavior. However, the model also considers the probability that the bank will not fail. If that probability is sufficiently low, the bank may play risky even without the state's guarantee of a bailout.

4. INVESTMENT BANK, CREDIT RATING AGENCY AND INVESTOR INTERACTION MODEL

Proposed model of interaction among investors, credit rating agencies, and investment banks as a dynamic Bayesian game is given in **Figure 3**. The game takes place over a period T. During this period, we observe the interaction of three rational players in a game with incomplete information. Within period T, the bank purchases a certain quantity of mortgage loans of variable quality from lenders. Nature assigns an aggregate value to this mortgage package as safe or risky. Based on the chosen strategy, the bank creates securities whose quality is assessed by the credit rating agency. Investors decide whether to invest based on the rating (safe or risky). In each period, the bank receives a new amount of mortgage loans whose risk changes over time.

DRAFT VERSION

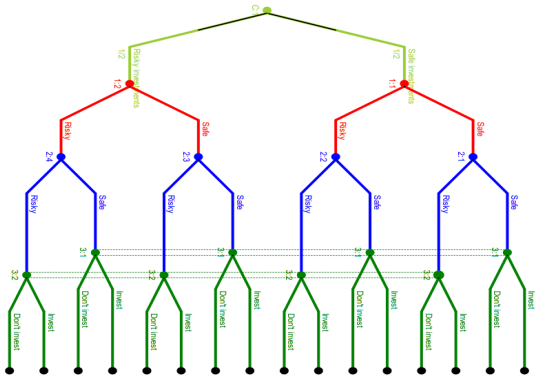


Figure 3: Bank, credit agency, investor interaction model

The game begins with nature determining, based on probability p, the quality of the mortgages forming the basis of the securities the bank offers and sells to investors. In the early stages of the game, mortgages represent safe and stable investments because loans were granted to people capable of repayment. As time passes and the volume of approved loans increases (also influenced by the bank's choice of a risky strategy), loans are given to people who cannot repay them, and thus mortgages become increasingly risky. The proposed model then branches into two subgames, depending on whether nature selected a risky mortgage outcome or a safe mortgage outcome. The model assumes investment bankers have insight into the true risk level of these financial instruments. After nature, the bank chooses a risky (r) or safe (S) strategy. The safe strategy means creating instruments based on safe mortgages, while risky means creating riskier and more complex instruments.

Next is the credit rating agency player, which can either rate the investment as risky, or to rate is as safe. The model assumes credit rating agencies have perfect information regarding the risk of the security. The credit rating agency must rate the instrument according to its true risk. If the bank plays risky, the instrument will be risky and should be rated as such. Conversely, if the bank plays safe, the instrument should be rated as safe. The last step in this game is the investor's choice to invest or not in the security offered by the bank. The only information the investor has about the risk of the financial instrument is the credit rating agency's rating, on which they base their decision.

The bank's payoff depends on three factors: the price of selling a single security, the quantity of securities sold, and non-monetary factors, such as moral beliefs, reputation, and others. The payoff formula is:

$$D = c * Q - \mathrm{B} \tag{3}$$

where D is the bank's profit, c is the price of a single security, Q is the quantity sold, and B are the non-monetary factors.

There are additional factors to consider. First, if the security is riskier, it carries a higher yield and its selling price is higher, but investors are less likely to buy such securities. On the other hand, some non-monetary factors, like moral belief, may lead the bank not to play risky. Second, there is a collective action problem. If competitors sell risky instruments while the bank sells safe ones, there is a risk of losing market share, further reducing payoff. Third, if the majority of issued mortgages are risky, payoff from the safe strategy is lower because the bank must issue fewer securities, reducing total payoff.

Credit rating agencies are funded by investment banks that create the financial products, not by investors. This creates a potential conflict of interest because banks that hire credit rating agencies have an interest in their securities receiving high safety ratings to attract investors. The payoff function also includes non-monetary factors such as moral beliefs, reputation loss, and similar aspects. There is potential for a collective action problem, where if one credit rating agency refuses to rate a bank's securities as safe, that bank may go to another credit rating agency.

Because the investor cannot know whether a security is truly safe or risky, the model assumes the dominant strategy of the investor is to buy if safe, or not buy if risky.

Moral hazard in this model represents a situation where the bank behaves risky. Its risky behavior leads investors to buy securities that have a risk of default. There are two potential reasons for this: first, the bank believes the credit rating agency will always rate its security as safe, second is that the bank believes all other banks play risky and that it would lose too much if it does not also play risky, resulting in a large loss of market share.

DRAFT VERSION

The game will be analyzed through two scenarios depending on the credit rating agency's strategy. The first scenario assumes the credit rating agency always plays "Safe," while the second assumes it gives an accurate rating. In reality, there is a possibility, which this model ignores, that the agency intends to give a correct rating but there is some probability of error.

Initially, the assumption is that the credit rating agency will always rate the security issued by the bank as safe. If the bank believes this, by backward induction it can realize that the investor's dominant strategy is to buy the security. Therefore, the bank can play risky and increase its payoff without fear that the investor will not buy. This conclusion relies on two key assumptions. The first is that the profit from sales ($c * Q$) exceeds the loss from non-monetary factors (B). The second is that the profit from selling securities under the risky strategy exceeds the profit from selling securities under the safe strategy. This is because the quantity of securities sold (Q) remains the same while the price per security (c) increases. Without the assumption about the credit rating agency's actions, the bank would balance increased profit per security against a reduction in the number sold. This situation represents moral hazard because the bank increases the risk for investors and the entire financial system without bearing the full risk itself. An additional factor in this scenario is that even if the bank wants to play safe (high value of non-monetary factors), payoff can decrease further due to loss of market share. Assuming non-monetary factors can be neglected, the bank's dominant strategy will be to play risky.

In the second scenario, assuming the credit rating agency will rate the security correctly, the bank's strategy changes and depends on whether it pays off to sell fewer securities at a higher price.

The conclusion is that if a conflict of interest existed in credit rating agencies, it contributed to the financial crisis. Also, if it did not exist, the collective action problem of investment banks likely created undesirable effects.

5. BORROWER AND LENDER INTERACTION MODEL

A proposed model of a dynamic Bayesian game involving a borrower (the actor taking the loan) and a lender (the actor granting the loan) is given in **Figure 4**, with the payout symbols used in the model in **Table 2**. The model incorporates certain elements from the works of Medak (2009) and Luo & Zhang (2010). The game begins with nature's decision on the type of borrower: a safe borrower has a high credit score and a high probability of repayment, while a risky borrower has a low score and a higher risk of default. The lender identifies the borrower's type based on the submitted documentation and decides whether to grant the loan. The interest rate for risky borrowers is higher and often variable, while safe borrowers receive a lower, more stable rate. The borrower always accepts the offered loan, and their strategy is to either repay it or default.

The model introduces two global variables: the rate of property price growth (A) and the interest rate (B). These change during the crisis so that property prices grow for a long period and then suddenly drop, while interest rates rise after the central bank raises the discount rate. Borrower payoffs depend on the situation: if the loan is not granted, the payoff is zero; if it is granted and repaid, the payoff depends on the subjective cost of interest (SK) and the relative property value (VN) according to the formula:

$$C = SK - VN \tag{4}$$

A high SK reduces the payoff, while a high VN increases it, making repayment possible even through selling the property. These values are influenced by the borrower's type, the interest rate (B), and property price movements (A). If the borrower does not repay the loan, their payoff is D, which includes reputational, psychological, and moral loss.

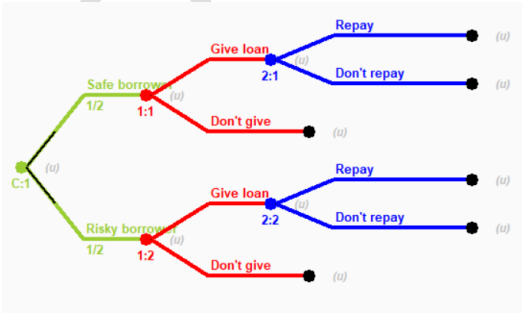


Figure 4: Borrower-lender interaction model

Table 2: Payout symbols for borrower-lender model

Before securitization		
	Repay	Don't repay
Z	B	F
D	C	D
After securitization		
	Repay	Don't repay
Z	G	G
D	C	D

Moral hazard in this situation occurs when the lender grants loans to risky borrowers because they do not bear the risk of their default. In the pre-securitization model, the bank's payoff can be the resale of the

DRAFT VERSION

property (F) or interest income (B), depending on the borrower's strategy. Since interest income is a long-term stable return, B is a better outcome compared to F, as resale of the property is a more difficult process that is more likely to result in a smaller payoff. However, assuming that property growth is high enough, the bank may find both outcomes acceptable. Therefore, it can be concluded that the bank's action depends on two factors. First, whether property price growth is high enough so that even if the borrower does not repay the interest, the bank will not lose money ($F \geq B$) and second, whether the borrower will choose to repay or not. If $F \geq B$, then the bank can grant the loan in any case. If $F < B$, then the bank considers the second factor—whether the borrower will repay. As noted, the borrower's strategy depends on the cost of interest payment, which in turn depends on three variables: borrower type, interest rate (B), and property price movements (A). Analysis through scenarios shows that at one point, property price growth, and consequently the profit from resale, was so high that lenders found it profitable to lend even to risky borrowers. There was also the presence of a "strategic default" strategy when property value loss outweighed the benefit of repayment. With securitization, lender payoffs become independent of borrower behavior, as loans are sold to financial institutions. The dominant strategy is to grant loans ($G > 0$), even to risky borrowers, leading to a non–Pareto optimal equilibrium. In conclusion to this model, securitization fostered moral hazard, but rising property prices also played a major role by enabling riskier behavior and short-term gains. Moral hazard also arises among borrowers who choose to default when penalties are low.

6. CONCLUSION

The aim of this paper was to model the interactions of financial market participants relevant during the financial crisis and analyze the occurrence of moral hazard. Using concepts from game theory, particularly Bayesian games and information asymmetry, the study examines the securitization chain, from lenders granting mortgages to borrowers, to investment banks repacking them into securities sold to investors, with credit rating agencies signaling their safety and the state acting as a potential rescuer.

Three models were developed: borrower–lender, investment bank–state, and investment bank–credit rating agency–investor. The borrower–lender model shows that, besides securitization, the housing market boom made lending to risky borrowers a dominant strategy, while some borrowers exhibited moral hazard by choosing not to repay despite having funds. The bank–state model indicates that bailout expectations encouraged risky behavior, especially when the bank's probability of failure was low.

While the paper focuses on the most evident moral hazards, others exist also existed in managerial bonuses, political corruption, academic conflicts of interest, and more. Future research could integrate additional crisis elements, merge and refine the models, and design mechanisms to prevent moral hazard and conflicts of interest in financial market interactions.

LITERATURE

Ferguson, C. H. (2012). *Predator nation: Corporate criminals, political corruption, and the hijacking of America.* Crown Currency.

Luo, Q., & Zhang, S. (2010). The Analysis of Mortgage under the Subprime Crisis Based on the Evolutionary Game. *In 2010 International Conference on E-Business and E-Government*, 4712-4715.

Medak, D. (2009). A Case for Government Enforcement: A Game Theory Analysis. *Undergraduate Economic Review*.

Murray, N., Manrai, A., & Manrai, L. (2017). The financial services industry and society: The role of incentives/punishments, moral hazard, and conflicts of interests in the 2008 financial crisis. *Journal of Economics, Finance and Administrative Science*, 168-190.

Rose, P. S., & Marquis, M. H. (2011). *Financial institutions and markets.* New York: McGraw-Hill.

DRAFT VERSION

GEOINFORMATION SYSTEM

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

ESTABLISHMENT OF A BURIAL SITE REGISTRY

DRAGAN ĐURĐEVIĆ[1], BILJANA KURIDŽA[2], DRAGANA SKORUP[3], MIROSLAV VUJASINOVIĆ[4], GORAN MARINKOVIĆ[5]

[1] Vojnogeografski institut, Beograd, dragan.djurdjevic.1982@gmail.com
[2]Vermessungsbüro Michalke, Eningen unter Achalm, biljanakuridza@ymail.com
[3] Faculty of Architecture, Civil Engineering and Geodesy, University of Banja Luka, dragana.skorup@aggf.unibl.org, 0009-0003-7067-5046
[4] Faculty of Architecture, Civil Engineering and Geodesy, University of Banja Luka, miroslav.vujasinovic@aggf.unibl.org, 0009-0008-0398-5614
[5] Faculty of Technical Sciences, University of Novi Sad, goranmarinkovic@uns.ac.rs, 0000-0002-1923-1082

Abstract: *In order to improve service provision, better resource management, and utilization in communal activities of particular public interest, such as funeral services, records are necessary. In the Republic of Srpska, there are legislative provisions regarding records of funeral services that have been analyzed for the purposes of this study. By employing unmanned aerial vehicle surveys, it is possible to obtain accurate data in a relatively short time frame, and by linking this with databases, up-to-date records are obtained. Through research and the practical part of the study, it has been demonstrated that maintaining records in databases and linking them with graphical representations provide a much clearer and more comprehensive picture of the situation on the ground, as well as offer various possibilities for planned development and expansion of cemeteries.*

Keywords: *Burial site, Cemetery, Unmanned aerial vehicles, Orthophoto, Database.*

1. INTRODUCTION

Records of land date back to the Roman Empire when tax assessment registers existed, continuing to the present day where cadastral records represent a collection of data about land and the objects on it (Miladinović Manojlo, 2013). Land data is utilized for statistical, economic, and various other purposes, as well as for maintaining diverse records. With the advancement of science and technology, there emerged a need for utilizing databases in cadastral systems, aiming to refine record-keeping across different domains. The demand for various land records is increasing, leading to the development of cadastres in fields such as geology, ecology, transportation, agriculture, forestry, mining, construction, and others. In order to enhance services, and improve resource management, and utilization, records are required in communal activities of particular public interest, such as funeral services.

In the Republic of Srpska, legislative provisions exist pertaining to records in funeral services (Official Gazette of the City of Banja Luka, 2013; Law on Cemeteries and Funeral Services, 2014), and with the advancement of technology, the utilization of databases and software packages for graphical representation enables the enhancement of such records. The demand in this field and urban planning indicates the necessity to analyze, assess, and establish unified methodologies for cemetery documentation, as well as the need for organized management, coordination, and direction of activities in this sector. Through new surveying methods, it is possible to obtain accurate data in a relatively short time frame, and by linking this with databases, up-to-date records are obtained (Barrile et al., 2020; BUZNA et al., 2021; Rahardjo et al., 2020).

The cemetery cadastre contains data on graves, burial sites for urns, owners of burial sites and graves, individuals entitled to burial rights, as well as all changes and causes of death. An integral part of this record is the positional plan of burial sites and graves (Law on Cemeteries, 2017). This type of cadastral record exists in countries within the region, such as Croatia (*Cemeteries*, 2022), as well as in other countries like Austria and Germany (*Friedhöfe Wien*, 2022)

Researchers have long recognized the need to integrate IT and geodetic methods to create a record that allows for real-time monitoring and regulation of funeral services. Works (Coruhlu et al., 2021; IEVSIUKOV et al., 2022; Mohamad et al., 2023; Schmidt, 2018) describe the use of geographic information

DRAFT VERSION

systems in terms of their application for management, site selection, and intelligent management and manipulation of funeral service systems.

Conducting a survey of a cemetery and recording burial sites from the ground requires significant time and money, especially for large areas. An alternative to this is aerial digitization of the cemetery. Using a high-resolution camera-equipped drone, cemeteries can be captured in detail from a high altitude. Thanks to preserved, precise geospatial data for each image, a digital cemetery cadastre can be created. Data captured in a relatively short time frame, which is highly accurate, can be imported into cemetery management software, used for managing graves, and retrieved from the database at any time (Unger et al., 2017).

The creation of a record and visualization that provides users with all necessary information is the starting point of this study. Emphasis is placed on using the QGIS software package to display data, with data being stored in a PostgreSQL database. Various queries in the database and the QGIS software package should display attributes related to burial sites such as site number, owner, area, or the individual buried. Practical examples will be conducted on orthophotos of the city cemetery at Pobrđe and the cemetery at Vrbanja.

2. MATERIALS AND METHODS

2.1. Area of interest

The burial site cadastre was prepared for the city cemetery at Pobrđe and the cemetery at Vrbanja, both located in the city of Banja Luka, as depicted in Figure 1. The city cemetery at Pobrđe covers an area of approximately 83500 m^2, while the city cemetery at Vrbanja occupies an area of around 31088 m^2. The cemetery situated at Pobrđe location encompasses multiple cadastral parcels, namely parcels 32, 35, 36, 38, 39, 40, and 41, whereas the cemetery located at Vrbanja is situated on a single parcel, parcel number 677/1.

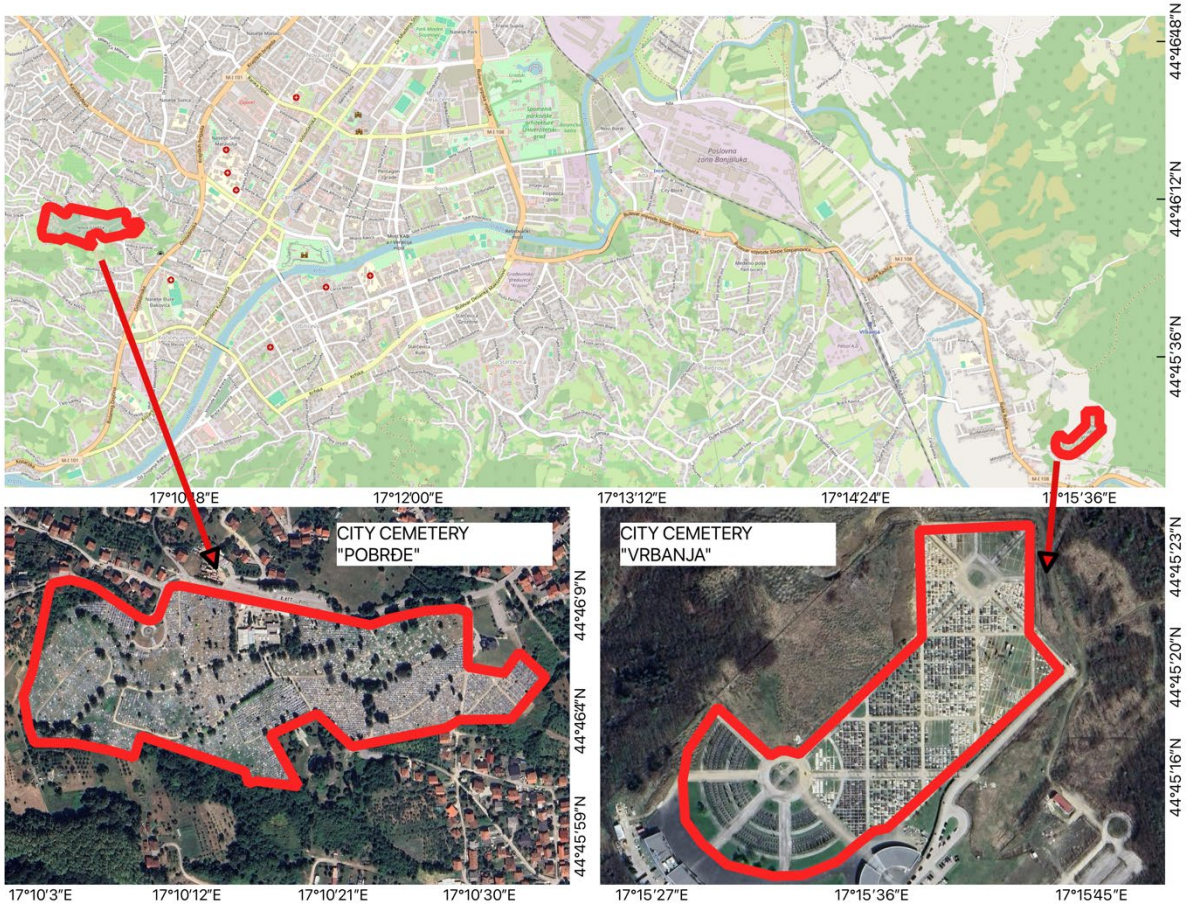


Figure 1: Area of interest.

2.2. Input data

The images of the surveyed areas were captured by an unmanned aerial vehicle – the Ruide Drone-eco Pro, by a surveying company. The survey was conducted in April 2022. The Ruide Drone-eco Pro model features a body with four foldable propellers arranged in an X-shaped configuration. The drone is fully integrated and ready to use without the need for assembly, with the capability of vertical takeoff and landing. The maximum length of the device diagonally is 716 mm, and the dimensions of the device are 564 x 564 x

360 mm (length, width, height). The weight of the device with the battery is 5.15 kg, and without the battery, it is 2.35 kg. The maximum payload capacity of the drone is 1.4 kg, making the maximum takeoff weight 6.55 kg. The positioning accuracy of the device is 1 cm + 1 ppm horizontally and 1.5 cm + 1 ppm vertically. With a single flight, an area of up to 6 km² can be surveyed.

For the processing needs, 188 terrain images of the central city cemetery at Vrbanja and 125 terrain images of the cemetery at Pobrđe are utilized.

The Law on Cemeteries and Funeral Services (Law on Cemeteries and Funeral Services, 2014) defines the fundamental concepts relevant to the development of the database, and a schema has been devised based on which tables and data types are created in the database.

2.3. Methods

The cemetery cadastre database creation is performed using the pgAdmin 4 tool, which serves to manage the open-source PostgreSQL database worldwide (Drake & Worsley, 2002).

Processing of images obtained by the unmanned aerial vehicle was conducted in the Agisoft Metashape software, which performs photogrammetric processing of digital images, captured by RGB or multispectral cameras, including multi-camera systems. It generates 3D spatial data and enables image processing into spatial information in the form of dense point clouds, polygon models with textures, georeferenced orthomosaics, digital surface models, and digital elevation models (*Agisoft Metashape User Manual Professional Edition, Version 1.7*, 2021). Further post-processing allows for shadow and texture removal from the model, vegetation index calculation, extraction of information for interactive maps, automatic classification of dense point clouds, and more.

After processing the terrain images, orthomosaics for both cemeteries are loaded into QGIS as raster layers. Additionally, the PostgreSQL database created is linked to QGIS. Upon integrating the database and orthomosaics, geometric elements of the cemetery are delineated, and database tables are populated. Real data were not used for populating the database tables for the study due to the fact that the enterprise "City Cemetery Banja Luka" possesses such data and, for privacy reasons, did not consent to their usage.

2.3.1. Database creation

The schema comprises seven tables and five data types. The tables are intended to contain basic information about the cemetery, burial site, grave, individuals buried or entitled to burial, and ownership. During table population, the type of cemetery can be selected based on surface area and burial method, cemetery type according to purpose, cemetery administration, as well as the legal basis for acquiring ownership of the burial site. The schema was developed in accordance with the Law on Cemeteries and Funeral Services (Law on Cemeteries and Funeral Services, 2014), as depicted in Figure2.

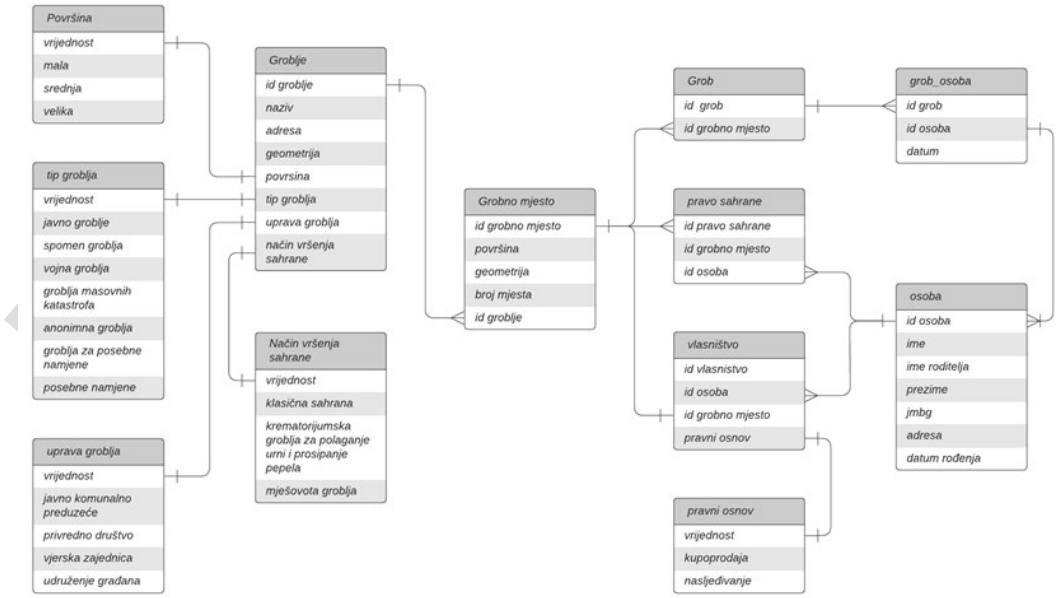


Figure 2: Schema of database creation in Serbian language.

2.3.2. Image processing obtained by an unmanned aerial vehicle

Image processing obtained by an unmanned aerial vehicle was conducted using Agisoft Metashape software. The first stage of image processing is carried out by the command Workflow - Align Photos, which

PREVIEW VERSION

improves the camera position for each photograph. In the second phase, aerotriangulation is performed with greater accuracy. If the Reference option is enabled, the overlap between photographs is predicted based on the geolocation of the photographs. Aerotriangulation is performed only with overlapping photographs. Agisoft Metashape enables the creation and visualization of dense point clouds. The point cloud generation is based on the estimated camera locations, and the generated point cloud has nearly the same density as the point cloud generated using LiDAR technology. The process is initiated by Workflow - Build Dense Cloud.

A Digital Elevation Model (DEM) is a digital representation of the terrain surface, i.e., it is a 2.5D surface model represented in the form of a regular grid, with height values for each grid point (Balasubramanian, 2017). Metashape allows the creation of this model via the command Workflow - Build DEM. With the interpolation mode enabled, Metashape will compute DEM for all areas visible in at least one image, and the setting is recommended for generating DEM.

An orthomosaic is a combined image created by merging original images projected onto the object surface and transforming into the selected projection, while an orthophoto projection is an individual image projected onto the object surface and transformed into the selected projection. Thus, an orthomosaic is created by merging orthophotos (*Agisoft Metashape User Manual Professional Edition, Version 1.7*, 2021).

Creating a high-resolution orthomosaic based on existing photographs and reconstructed terrain: Workflow - Build Orthomosaic. For the Surface parameter, it is recommended to use DEM, while for more demanding applications, DSM is used, and the Blend mode is set to Mosaic.

The orthomosaic of a part of the Vrbanja cemetery, obtained by processing photographs, was exported in .tiff format, and its appearance can be seen in Figure 3.

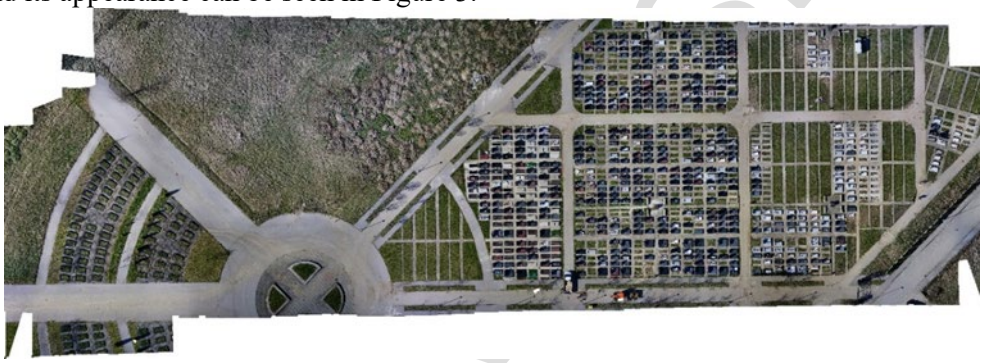


Figure 3: The orthomosaic of a part of the cemetery at Vrbanja.

2.3.3. Integration into QGIS

After the terrain image processing, orthomosaics for both cemeteries are loaded into QGIS as raster layers. After the raster layers are loaded, they are connected to the database. After the connection is established, layers with geometry in the database are created.

The tables are populated in pgAdmin 4 either directly or through code:

```
INSERT INTO table_name (column_1, column_2, column_3 ...)
VALUES (value 1, value 2, value 3...);
```

3. RESULTS

As a result of database creation, image processing, and polygon drawing, a comprehensive record of burial sites with graphical representation is obtained. This allows for the search and display of available and occupied burial plots, associated individuals, and other related data stored in tables. The integration of data and graphics provides a clearer understanding of the current situation on the ground, serving as a basis for needs assessment, planning for cemetery expansion, and landscape development.

Various tools are available in pgAdmin 4 for viewing existing data. One method involves creating a view using SQL code, which consolidates relevant data from multiple tables into a single virtual table.

To display occupied and vacant burial sites, the tables burial_site2 (grobno_mjesto2) and grave2 (grob2) need to be joined. If the primary key from burial_site2 exists in the grave2 table, the site is considered occupied, and the corresponding column in the view is assigned the value 1. If the primary key is not found in grave2, the site is vacant, and the value is set to 0. Other attributes from both tables can also be selected and included in the view.

If there is no need to retain the displayed data permanently, a temporary table can be used instead.

In addition to viewing data in pgAdmin 4, the same data can be visualized in QGIS. A layer is added using the created view grobna_mjesta_zauzeto2. By right-clicking on the layer and selecting Layer

DRAFT VERSION

Properties, then navigating to the Symbology tab, the Categorized option can be chosen from the dropdown menu. The classification is based on the occupied column from the view, which contains two values (0 and 1), as defined in pgAdmin 4. This classification is now displayed visually in the QGIS interface, as shown in Figure 4.

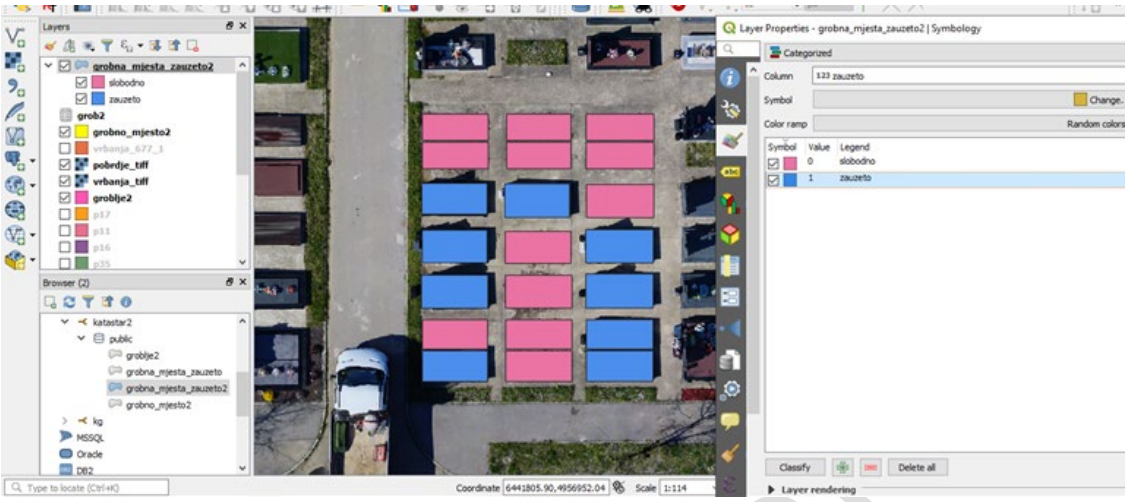


Figure 4. The depiction of available and occupied burial sites in QGIS.

QGIS also enables table joins, allowing the display of all necessary attributes in a single view. Using the Join option, the tables grob2, grob_osoba2, and osoba2 can be associated with grobno_mjesto2. All columns from the joined tables are then available in the attribute table. During the join process, QGIS allows you to choose which columns to include. Once the joins are complete and attributes are added, the information can be visualized graphically.

4. DISCUSSION AND CONCLUSION

From the time when land records were used for tax collection to the present day, where cadastral surveys represent a comprehensive record containing data about land and objects on it, significant advancements in technology, data collection methods, and record-keeping can be observed. Once, cemeteries were merely marked on maps with specific topographic symbols, but today, with appropriate computer software, electronic records can be maintained, and data can be linked with graphics.

Following the analysis of legal regulations for the practical part of the study, deficiencies were identified regarding the definition of the term "grave chamber" and the condominium of burial sites. The concept of a grave chamber is not defined either in the Law on Cemeteries and Funeral Services or in the Decision of the City of Banja Luka, thus, it is not clear what is meant by this term. Therefore, it would be necessary to eliminate ambiguities regarding this concept in terms of its meaning and determining the number of individuals who need to and can be buried, so that the land area can be called a grave chamber. Subsequently, this term could be added to the database itself, enabling an even more advanced review of cemetery records.

In practice, it happens that multiple individuals are buried in one burial site. The Law on Cemeteries and Funeral Services does not define the condominium of burial sites, while Article 43 of the Decision of the City of Banja Luka defines that a maximum of two individuals can be buried in one grave if a grave chamber has not been previously built. Since the concept of a grave chamber is unclear, and the number of individuals that can be buried is unknown, this Decision also does not clearly define the number of individuals that can be buried in one burial site or grave chamber. Therefore, the allowed number of condominiums when creating the database could not be defined.

The Regulation on the Content and Manner of Maintaining Cemetery Records provided for by the Law on Cemeteries and Funeral Services has not been adopted, so the Cemetery Administration maintains records in accordance with this Law, which was considered for the practical part of the study. To improve record-keeping, it would be necessary to develop the Regulation on the Content and Manner of Maintaining Cemetery Records.

With the introduction of the Regulation on the Content and Manner of Maintaining Cemetery Records, all records would be aligned, avoiding individual interpretation of the Law by each Cemetery Administration regarding the data kept in the records, data presentation, and search methods, and the appearance of the

DRAFT VERSION

records themselves, which should be clear and concise, enabling necessary information checks for buyers or visitors.

Survey methods for data collection demonstrate significant progress in terms of equipment and terrain recording methods. Today, it is possible to quickly and efficiently produce survey maps with satisfactory accuracy. This study shows that the method of recording data with unmanned aerial vehicles is a suitable solution for these purposes. After processing the images, an orthophoto is obtained, which can be used as a basis for a graphical representation of cemeteries and burial sites.

By using specific software packages, images obtained from unmanned aerial vehicle surveys can be processed, and an orthophoto can be obtained. Further use of software packages digitizes this base and links it with a database. In this way, various queries can be created, making data search faster and simpler.

The application of new terrain survey methods yields results of satisfactory accuracy, and data is collected relatively quickly. Maintaining records in databases and linking them with graphics provides a much clearer and more transparent picture of the situation on the ground, offering various possibilities for planned development and expansion of cemeteries. Buyers can more easily access information about available sites, while visitors can more easily locate the graves of family members.

LITERATURE

Agisoft Metashape User Manual Professional Edition, Version 1.7. (2021).

Balasubramanian, A. (2017). Digital elevation model (DEM) in GIS. University of Mysore.

Barrile, V., Fotia, A., Bernardo, E., & Bilotta, G. (2020). Road cadastre an innovative system to update information, from big data elaboration. Computational Science and Its Applications--ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1--4, 2020, Proceedings, Part IV 20, 709–720.

BUZNA, C. C., TAB\uAR\uA, M., \cSANDOR, C. I., & Petru, R. (2021). PREPARATION OF THE AGRICULTURAL REGISTER BASED ON UAV TECHNOLOGY AND GIS TECHNIQUES. CASE STUDY. LIFE SCIENCE AND SUSTAINABLE DEVELOPMENT, 2(1), 139–144.

Cemeteries. (2022). https://www.gradskagroblja.hr/trazilica-pokojnika/15

Coruhlu, Y. E., Baser, V., & Yildiz, O. (2021). Object-based geographical data model for determination of the cemetery sites using SWOT and AHP integration. Survey Review, 53(377), 108–121.

Drake, J. D., & Worsley, J. C. (2002). Practical PostgreSQL. " O'Reilly Media, Inc."

Friedhöfe Wien. (2022). https://www.friedhoefewien.at/

IEVSIUKOV, T., CHETVERIKOV, B., OPENKO, I., KOVALCHUK, I., SHEVCHENKO, O., STEPCHUK, Y., TYKHENKO, R., & MAKAROV, O. (2022). TOPOGRAPHIC AND GEODETIC SUPPORT FOR THE DEVELOPMENT OF THE GIS REGISTER OF POLISH BURIALS-CASE STUDY ON BAIKOVE CEMETERY IN KYIV. Scientific Papers. Series E. Land Reclamation, Earth Observation \& Surveying, Environmental Engineering, 11.

Miladinović Manojlo. (2013). Katastar nepokretnosti,. Arhitektonsko-građevinsko-geodetski fakultet.

Mohamad, F., Maulud, K. N. A., & Taib, A. M. (2023). Modelling of geospatial data through multi criteria analysis for cemetery land management. IOP Conference Series: Earth and Environmental Science, 1167(1), 12051.

Law on Cemeteries, (2017).

Official Gazette of the City of Banja Luka, N. 18/13. (2013). Decision of the City of Banja Luka.

Law on Cemeteries and Funeral Services, (2014).

Rahardjo, N., Santosa, D. H., & Marhaento, H. (2020). Drone Application for Generating a High Precision Orthophoto to Support Village Boundary and Land Use Mapping in Indonesia. International Journal of Geoinformatics, 16(2).

Schmidt, M. (2018). Cemetery Management in a GIS Framework: A Case Study of Woodland Hills Memorial Park.

Unger, L., Knopf, P., & Monetti, D. (2017). Praxis der Datenerfassung mit Drohnen für GIS und Vermessung. REAL CORP 2017--PANTA RHEI--A World in Constant Motion. Proceedings of 22nd International Conference on Urban Planning, Regional Development and Information Society, 717–722.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

GEOSPATIAL ANALYSIS OF RADIOACTIVITY DISTRIBUTION IN ALLUVIAL SOILS OF THE KOLUBARA RIVER BASIN AT 20 CM DEPTH

VUK GAJIĆ[1], BORIS VAKANJAC[2], NIKOLA MILIĆ[3]

[1] Singidunum University, Environment and Sustainable Development, Belgrade, vgajic@singidunum.ac.rs, 0000-0001-6464-8226
[2] Military Geographical Institute "General Stevan Bošković", Belgrade, International Research Academy of Science and Art IRASA, Belgrade, borivac@gmail.com, 0000-0002-9429-6943
[3] Ministry of Agriculture, Forestry and Water Management of the Republic of Serbia, Directorate for Agrarian Payments, Belgrade, nikola.milic@minpolj.gov.rs,

Abstract: *This paper presents a study of the spatial distribution of radioactivity in alluvial soils across the Kolubara river basin. A total of 136 alluvial soil samples were collected at a depth of 20 cm along the main course of the Kolubara and its tributaries and subsequently analysed for gamma radiation under laboratory conditions. The measured gamma dose rates (μSv/h) ranged from approximately 0.12 to 0.41 μSv/h, mainly exhibiting values around 0.20-0.25 μSv/h. Most recorded values correspond to the expected natural background radiation, while about 5% of locations demonstrated moderately elevated levels (above ~0.30 μSv/h). The highest gamma dose rates were observed locally at several sites within the catchments of the Ljig, Tamnava, Ub, and Obnica rivers. In contrast, the lowest levels were registered in the downstream section of the Kolubara and the Jablanica sub-basin. Spatial GIS analysis (using quantile classification) revealed the absence of a simple longitudinal gradient-zones of higher values appear sporadically, likely reflecting heterogeneous geological substrates and/or localized radionuclide sources. However, certain "hot spots" warrant further investigation regarding the origins of elevated radioactivity to preserve environmental quality.*

Keywords*: radioactivity; alluvial soil, natural radionuclides, anthropogenic sources, Kolubara.*

1. INTRODUCTION

This material describes the spatial distribution of gamma radiation levels in alluvial soils of the Kolubara River Basin based on laboratory measurements of soil samples, with special attention to identifying areas with potentially increased levels of radioactivity and their origin (natural or anthropogenic). Natural background radioactivity in soil originates from radionuclides that naturally occur in the mineral composition of rocks and soils (primarily radionuclides from the uranium-238 and thorium-232 series, as well as potassium-40). Globally, average concentrations of these radionuclides in surface soil are about 400 Bq/kg for ^{40}K, 35 Bq/kg for 238U (or ^{226}Ra), and 30 Bq/kg for ^{232}Th (UNSCEAR, 2000). Also, the average natural background radiation typically ranges from 0.05 to 0.10 μSv/h; in residential areas, a level up to 0.30 μSv/h is considered normal. These radionuclides emit gamma radiation, contributing to the natural external radiation level outdoors. However, the content of natural radionuclides can vary significantly depending on the geological settings dominated by magmatic and metamorphic rocks (such as granite), which often have higher levels of gamma radiation from the soil than areas with mostly sedimentary rocks.

In addition to natural sources, radionuclides enter the environment due to human activity. One of the most significant global anthropogenic sources of radioactivity was atmospheric nuclear testing during the 1950s and 1960s, as well as nuclear accidents (most notably the Chornobyl disaster in 1986). The long-lived fission product cesium-137, produced by nuclear explosions and accidents, remains in the environment for decades; even nearly forty years after Chornobyl, it can still be detected in trace amounts in soil and sediments (Drndarski & Križman, 1994; Steinhauser et al., 2014). Specific industries can lead to so-called technologically enhanced naturally occurring radioactive materials (TENORM), where the concentration of natural radionuclides in a material is increased by human activity. An example of TENORM is the exploitation and burning of coal. Coal contains natural radionuclides (^{226}Ra, ^{232}Th, ^{40}K), which are concentrated in ash during combustion.

Alluvial plains of rivers are specific areas when it comes to radionuclide distribution. Due to periodic flooding and sediment deposition, natural and residual artificial radionuclides distribution in alluvial soils can

DRAFT VERSION

be heterogeneous. The Kolubara River and its tributaries form a broad alluvial valley in western and central Serbia. Kolubara originates from the merging of the Obnica and Jablanica rivers in Valjevo. Stretching approximately 123 km, it flows into the Sava River near Obrenovac. Its basin includes numerous smaller and larger tributaries, such as Gradac, Peštan, Ljig, Tamnava, Turija, Beljanica, Ub, Lukavica, Rabas, Kladnica, Toplica, and others. The valley of Kolubara and its tributaries is mainly composed of Quaternary alluvial sediments (sand, gravel, clay) that make up fertile agricultural land. Serbia's largest coal mining basin, the Kolubara coal basin near Lazarevac, with several open-pit mines (Field D, Tamnava Field, etc.) and accompanying coal-fired thermal power plants, is also located in this region. This combination makes the Kolubara basin interesting for exploration of radioactivity distribution: there are typical natural radionuclides in the alluvial ecosystem; also, there are potential anthropogenic influences from industry (mining, energy production) and residual radioactivity originating from global nuclear events. Previous studies have addressed the content of radionuclides in Kolubara lignite and the local environment (Djurašević et al., 2014; Krneta Nikolić et al., 2016), as well as the mineral composition of Kolubara sediments and the natural radioactivity within them (Šaponjić et al., 2022).

2. METHODOLOGY

Sampling was carried out at 136 locations distributed along alluvial terraces and the banks of the main course and tributaries at intervals of approximately 2 km (Figure 1). At each site, a sample of the surface soil layer was collected from about 20 cm depth (digging a small pit). Each sample (~1 kg) was packed and transported to the laboratory for measurement. All measurements were performed in laboratory conditions using a calibrated portable digital gamma-radiometer with a Geiger-Müller detector (Gamma-scout), designed for measuring the ambient dose equivalent rate of gamma radiation (μSv/h). A compact handheld device with a digital display and measurement stabilization function was used, with prior accuracy checks performed at reference points.

In the laboratory, the gamma spectrometer was positioned centrally within each soil sample under controlled conditions, simulating the natural environment of the detector surrounded by 20 cm of soil. The device was protected with a thin transparent plastic film to allow direct contact with the sample while preventing damage or contamination of the probe. Measurement for each sample lasted several minutes until the dose readings stabilized. Both the minimum and maximum registered dose rates were recorded during the measurement interval. The maximum stable dose value (μSv/h) was used as the representative value for each sample. In this approach, the maximum rather than the mean value estimates potential external exposure. It enables the identification of local peaks in radioactivity, which may indicate the presence of environmental anomalies. In parallel with sampling, UTM coordinates of each site were recorded using a GPS device, enabling spatial analysis and visualization of results in a GIS environment.

The measured dose values (μSv/h) for the 136 sites were analysed statistically to determine basic descriptive parameters-minimum, maximum, median, quartiles, variance, and so on (using Python/Pandas software). The resulting value distribution was classified by quantiles (e.g., defined thresholds for the lower quartile Q1, median Q2, upper quartile Q3, as well as the 90th and 95th percentiles) to distinguish relatively "low," "medium," and "high" categories of radiation levels within the dataset.

3. RESULTS

Based on laboratory measurements of gamma dose rates for 136 alluvial soil samples from the Kolubara River basin (20 cm depth), the recorded range of values was from 0.121 μSv/h (the lowest) to 0.405 μSv/h (the highest). The minimum registered dose was about 0.12 μSv/h, and the maximum was about 0.41 μSv/h. The mean value for the entire dataset was not used in the analysis (due to the nonlinear distribution of data), while the median was 0.215 μSv/h.

A histogram of the frequency of measured doses shows that most locations (over 60% of samples) have gamma doses in the range of approximately 0.20-0.25 μSv/h. A smaller portion of the sites (about 18%) have lower values, between approximately 0.16 and 0.19 μSv/h, while about 16% of the samples exceed 0.25 μSv/h. Only a small number of isolated cases (around 7% of all samples) show higher values above 0.30 μSv/h. In other words, most measurement sites display stable and moderate dose values, while elevated levels appear rarely and at specific points (Figure 2). The highest recorded dose rates were identified in ten locations with the highest measured dose values. These locations and the corresponding maximum doses are listed in Table 1.

DRAFT VERSION

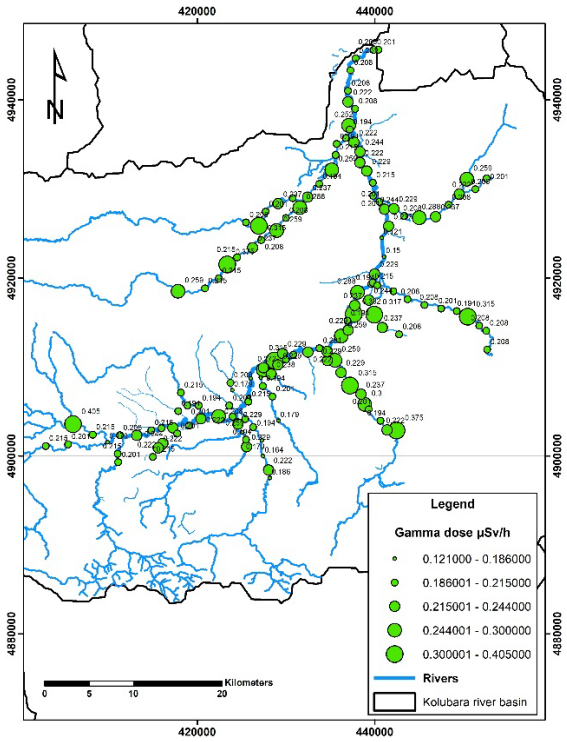


Figure 1: Spatial distribution of maximum gamma dose rates in alluvial soils of the Kolubara river basin (µSv/h) at 20 cm depth

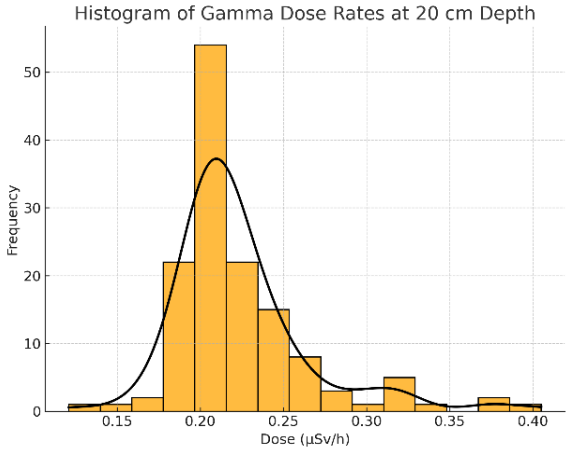


Figure 2: Distribution of measured gamma dose rates (µSv/h) at 20 cm depth in Kolubara alluvial soil samples

Table 1: Highest measured values

Sample ID/location	Max dose (µSv/h)	Lithological unit	Possible cause
Obnica 2	0.405	Paleozoic schists (metamorphic formations, Gradac basin)	Natural: high Th/U content in metamorphic rocks; minor industrial influence (distant mining sites)
Ub 7	0.375	Tertiary clay-schist formations (Kolubara basin)	Combined: local geology with increased K/U content; possible minor influence of power plant ash
Ljig 10	0.375	Palaeozoic limestones and schists (Ljig mountain slopes)	Natural: granite fragments
Kolubara 23	0.332	Quaternary alluvium (deposits from various sources)	Natural: sediments of mixed origin
Lukavica 3	0.317	Tertiary limestones (tributary of Ub, near a mine)	Natural: sediments with high concentrations of natural radionuclides
Kolubara 33	0.315	Quaternary alluvium	Natural: deposits of geologically diverse origin; possible mixing of materials with elevated K and Th content
Tamnava 10	0.315	Tertiary lacustrine sediments (Kolubara coal basin)	Anthropogenic: proximity to the Tamnava open pit
Peštan 7	0.315	Limestone formations (Valjevo region)	Natural: carbonate sediments rich in Ra/U
Ljig 4	0.315	Paleozoic schists (Ljig source area)	Natural: sediment material originated from rocks with high radionuclide content
Ljig 6	0.300	Paleozoic schists (Ljig source area)	Natural: sediment material originated from dominant metamorphic rocks and flysch

DRAFT VERSION

4. DISCUSSION

Natural sources and geological influences: The obtained results are generally consistent with the expected levels of natural radioactivity in soils for this region. The recorded values (~0.2 μSv/h, typically) correspond to background gamma radiation originating from the geological substrate, characteristic for areas with a moderate content of natural radionuclides. Globally, according to UNSCEAR reports, the average annual effective dose from natural radionuclides in soil is about 0.5–0.6 mSv/year outdoors (UNSCEAR, 2000). If our measured values of about 0.2 μSv/h are recalculated to an annual dose, assuming 20% outdoor occupancy, the effective dose is approximately 0.35 mSv/year, comparable to the global average (UNSCEAR, 2000). Differences between locations within the basin can be attributed to variations in the mineral composition of sediments and the basin's geology. Kolubara collects sediments from geologically diverse regions, from limestone (e.g., the Gradac catchment) and flysch zones to tertiary basins with coal-bearing layers (Lazarevac area). Elevated dose values recorded in the Ljig River basin (see Table 1) could be associated with the fact that this area partly drains regions richer in natural radionuclides (possibly including fragments of granitic or other igneous rocks in the alluvium). Conversely, the lowest values in the downstream part of Kolubara are likely the result of large amounts of sediment deposited from various sources. Material may be poorer in natural radioactive elements and "dilutes" the overall radionuclide content in the soil at the confluence.

Anthropogenic influences: Elevated values at some locations may also be due to human activities that have led to localised enrichment in radionuclides. The impact of coal mining on the Kolubara basin and the operation of thermal power plants were also considered. It is known that coal ash contains higher concentrations of natural radionuclides (especially from the uranium and thorium series) compared to unburned coal (Habib et al., 2018). At open-pit coal mines, these radionuclides may enter the surrounding soil through dust and ash deposition or surface runoff. Previous studies around the Kolubara open-pit mines have found somewhat higher levels of natural radionuclides in soil; for example, Mitrović et al. (2013) reported that the radium equivalent in uncultivated soil near one of the mines was about 235 Bq/kg, the highest value recorded in their study (Mitrović et al., 2013). Such activity corresponds to an external hazard index of about 0.64, below the prescribed limit (Hazard Index = 1). Therefore, even at that site, no unacceptable risk to the public was found (Mitrović et al., 2013). The influence of fly ash from more distant thermal power plants should also considered. The Kolubara A power plant (Veliki Crljeni) and the TENT A/B plants in Obrenovac (which use Kolubara coal) have operated for decades so that ash emissions could have contributed to the radionuclide content (such as ^{137}Cs, ^{40}K, ^{226}Ra) in nearby soils. However, systematic monitoring by the relevant institutes has shown no accumulation significant enough to alter the radiation background over time (Krneta Nikolić et al., 2016). Accordingly, our data do not show a clear pattern of elevated values exclusively downwind of the plants or pits- some of the highest measured sites (e.g., Ljig 10) are located outside the direct influence of power facilities. Mentioned suggests that the contribution of current anthropogenic sources, if any, is mainly localised and relatively minor compared to natural variations. Besides industry, anthropogenic radionuclides from other sources may also be present in alluvial soils. The fallout following nuclear accidents (primarily ^{137}Cs after Chornobyl) contaminated surface soils throughout Europe in 1986. Some of that cesium was redistributed on alluvial terrain through erosion and flooding. Radioactivity measurements in Kolubara sediments in the 1990s showed trace ^{137}Cs (Drndarski & Križman, 1994). Today, almost forty years later, the contribution of residual ^{137}Cs to external exposure is minimal compared to natural radionuclides-probably only a few percentages of the total dose (Steinhauser et al., 2014).

5. CONCLUSION

The results indicate that the current level of radioactivity in the alluvial zone of the Kolubara River primarily reflects natural sources, with only sporadic and limited influence from anthropogenic factors. Compared to similar studies in other river basins of western Serbia (Dugalic et al., 2010) and studies from neighbouring countries (Gajić, 2022), the measured dose rates in the Kolubara basin do not significantly deviate-they fall within the range typical for natural conditions, with occasional local variations. The information mentioned is reassuring from a radiological protection perspective: the population of the Kolubara region is not exposed to elevated external radiation from the soil (Dugalic et al., 2010). The average gamma dose rates from soils in Nigeria, Ghana, Turkey, India, and China range from about 0.05 to 0.15 μSv/h. At the same time, in this study, slightly higher values were measured around 0.2 μSv/h-primarily due to the specific measurement methodology (detection inside the sample at a depth of 20 cm, as opposed to standard surface measurements). Nevertheless, these values remain far lower than those recorded in rare areas of exceptionally high natural radioactivity, such as Ramsar in Iran, where dose rates may reach 1-5 μSv/h. The results suggest that the

DRAFT VERSION

observed differences in radioactivity within Kolubara's alluvial soils are primarily due to natural geological composition, without clear evidence of significant anthropogenic impact.

LITERATURE

Djurašević, M., Todorović, D., Savić, P., Forkapić, S., Bikit, I., & Mrdja, D. (2014). Natural radioactivity in lignite samples from open pit mines "Kolubara", Serbia - risk assessment. *Applied Radiation and Isotopes*, 87, 73-77. https://doi.org/10.1016/j.apradiso.2013.11.096

Drndarski, N., & Križman, M. (1994). Radioisotopes in periphyton and sediments from the Kolubara River and its tributaries. *Water Research*, 28(6), 1471-1474. https://doi.org/10.1016/0043-1354(94)90315-8

Dugalic, G., Krstić, D., Jelić, M., Nikezić, D., Milenković, B., & Pucarević, M. (2010). Heavy metals, organics and radioactivity in soil of western Serbia. *Journal of Hazardous Materials*, 177(1-3), 697-702. https://doi.org/10.1016/j.jhazmat.2009.12.087

Gajić, V. (2022). *Prostorna analiza zagađenja u Kolubarskom slivu*. Doktorska disertacija, Univerzitet Singidunum, Beograd. https://singipedia.singidunum.ac.rs/izdanje/44266-mikrobioloska-dekontaminacija-priobalnog-zemljista-u-slivu-reke-kolubare-primenom-visokoenergetskog-jonizujuceg-zracenja

Habib, M. A., Basuki, T., Miyashita, S., et al. (2018). Distribution of naturally occurring radionuclides in soil around a coal-based power plant and their potential radiological risk assessment. *Radiochimica Acta,* 107(3), 201-210. https://doi.org/10.1515/ract-2018-3044

Krneta Nikolić, J., Rajačić, M., Janković, M., Sarap, N., Anđelković, T., & Krstić, M. (2016). Study of radioactivity in environment around power plants TENT A and Kolubara due to coal burning for 2015. *RAD Conference Proceedings*, 1, 84-89. https://doi.org/10.21175/RadProc.2016.20

Mitrović, B. M., Vitorović, G., Andrić, V., Stojanović, M., Vitorović, D., Grdović, S., & Vićentijević, M. (2013). Prirodni radionuklidi u mineralnim đubrivima i obradivom zemljištu. *Veterinarski glasnik*, 67(5-6), 359-367. https://doi.org/10.2298/VETGL1306359M

Steinhauser, G., Brandl, A., & Johnson, T. E. (2014). Twenty-five years after Chernobyl: situation and perspectives of Fukushima. *Journal of Environmental Radioactivity*, 136, 155-170. https://doi.org/10.1016/j.scitotenv.2013.10.029

United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR). (2000). *Sources and Effects of Ionizing Radiation*. UNSCEAR 2000 Report to the General Assembly, United Nations, New York. https://www.unscear.org/unscear/en/publications/2000_1.html

Šaponjić, A., Gyoshev, S., Baščarević, Z., Janković-Mandić, Lj., Ljubenov, G., & Kokunešoski, M. (2022). Characterization of sedimentary minerals from Kolubara mining basin, Serbia, with the determination of natural radioactivity. *Science of Sintering*, 54(1), 39-50. https://doi.org/10.2298/SOS2201039S

DRAFT VERSION

SYM-OP-IS 2025
52. Simpozijum o operacionim istraživanjima
Palić, 7 - 10.9.2025.

UNIVERZITET U BEOGRADU
FAKULTET ORGANIZACIONIH NAUKA

Review article

DRONE SURVEYING FOR THE RESTORATION OF HISTORIC BUILDINGS

NENAD GALJAK[1], JELENA KOLOROGIĆ[2], SLAĐANA STANIŠIĆ[3], MIROSLAV VUJASINOVIĆ[4], MIODRAG REGODIĆ[5]

[1] University of Defence, Military Academy, Belgrade; galjaknenad@gmail.com, 0000-0002-2493-9202
[2] University of Banja Luka, Faculty of Architecture, Civil Engineering and Geodesy, Banja Luka; jelena.kolorogić@student.aggf.unibl.org
[3] University of Banja Luka, Faculty of Architecture, Civil Engineering and Geodesy, Banja Luka; sladjana.stanisić@aggf.unibl.org
[4] University of Banja Luka, Faculty of Architecture, Civil Engineering and Geodesy, Banja Luka; miroslav.vujasinović@aggf.unibl.org, 0009-0008-0398-5614
[5] University of Banja Luka, Faculty of Architecture, Civil Engineering and Geodesy, Banja Luka; miodrag.regodic@aggf.unibl.org , 0000-0003-4675-4150

Abstract:*Unmanned aerial vehicles (UAVs), thanks to the rapid development of sensors, software, and communication and navigation technologies, have become economically very accessible and easy to use, which has opened up great possibilities for their application. UAVs have become an excellent platform for collecting spatial data, which can be used to create 3D models of terrain or objects. With the help of appropriate technology, it is possible to achieve the preservation, protection, and valorization of architectural, archaeological, and all other forms of cultural heritage. Photogrammetry is a scientific field that plays a very important role in documenting cultural and historical heritage. This paper presents the method of data collection using a UAV, as well as the process of data processing and modeling. The imagery obtained by using the UAV was processed, a digital model of the building was created, and the modeling procedure was described.*

Keywords:*Historical structures, Unmanned aerial vehicle, Terrestrial photogrammetry, Digital building model.*

DRAFT VERSION

1. INTRODUCTION

The continuous advancement of computer technology and its increasing availability have led to a growing demand for high-quality geospatial data. Due to their complexity and diversity, primarily in geometric terms, an increasing amount of measured data is required for accurate and detailed representation. One of the most efficient methods currently used for collecting data in the field includes photogrammetry and laser scanning techniques.

Unmanned aerial vehicles (UAVs) have become excellent platforms for collecting spatial data, which can be used to create 3D models of terrain or objects. UAVs can be remotely controlled, automated, or autonomous systems capable of gathering large amounts of data in a short period of time. Thanks to the rapid development of sensors, software, and communication and navigation technologies, they have become economically accessible and easy to use. This has resulted in an increasing number of potential applications.

The use of photogrammetry in the preservation of historical structures has a very long tradition. Photogrammetric measurements have been used for decades to create documentation of historical buildings and assist in their restoration.

2. APPLICATION OF CLOSE-RANGE PHOTOGRAMMETRY FOR THE PURPOSE OF CREATING GEODETIC-TECHNICAL DOCUMENTATION OF OBJECTS

The purpose of creating geodetic-technical documentation is to appropriately document the existing geometric condition of the object. The goal can be inspection of the object, preparation of bases for reconstruction and rehabilitation, harmonization of new objects with the existing environment, or the creation of technical documentation for archival purposes.

The emergence of digital photogrammetry has opened up entirely new possibilities in the creation of geodetic-technical documentation for building structures. The changes are very significant and relate not only to the existing work phases, such as:

- surveying;
- preparation of graphic documentation;
- archiving;
- creation of digital orthophotos and 3D models of objects.

The process of creating geodetic-technical documentation is divided into two logical phases. The first consists of photogrammetric surveying, necessary field measurements, and preparation of archival documentation. The second phase involves the creation of detailed geometric documentation, such as precise digital orthophotos, 2D or 3D models of the object, and similar, all in digital form.

2.1. Digital Building Model

Digital Building Model - DBM is created by collecting data about constructed structures and can be described as a set of individual 3D objects that are not interconnected. (Cetl, Tomić & Lisjak, 2013).

Figure 1: Digitalbuilding model(DJI, 2025)

Digital models (3D models) are created based on spatial data obtained using various geodetic methods. In addition to data collection using unmanned aerial vehicles (UAVs), data for the creation of 3D models can also be obtained using the following methods:

- field geodetic surveys;
- laser scanning;
- radar data acquisition;
- vectorization method using existing topographic bases (Lukić, Terzić, Galjak, Gigović, Stojanović & Kremić, 2024).

2.2. Restoration of historical buildings

With the emergence of photogrammetric surveying, this discipline has assumed a crucial role in documenting cultural and historical heritage.

Monuments and historical buildings are the most reliable records of history. They are of immeasurable value, but every structure has its lifespan and ages along with the materials from which it was built (even under ideal conditions).

For culturally and historically significant structures, there are various types of documentation, which also represent one of the first steps in the preservation of cultural heritage. These usually include documents related to geometry, landscape, archaeology, construction materials, and the historical and cultural significance of the site, among others. During the process of documentation and recording, knowledge about the structures can be expanded based on old photographs, drawings, and plans, state archives, archaeological data, and notes.(Mulahusić, Tuno, Topoljak, Balić, Hadžiosmanović, Stanić, & Hajdar, 2013).

DRAFT VERSION

Photogrammetric surveys have been used for the documentation of historical structures and their restoration over many years.

3. AERIAL SURVEYING OF HERITAGE STRUCTURES USING UNMANNED AERIAL VEHICLES FOR RESTORATION APPLICATIONS

The practical part of the work consists of surveying the structure using the photogrammetric method (close-range photogrammetry) and subsequent modeling aimed at the restoration of the building's facade.

The surveying and modeling were carried out on a section of the City Administration building of the City of Banja Luka (Figure 3), which represents one of the most significant structures in terms of cultural development, as well as in the study of architectural heritage in Yugoslavia. The construction of the Ban's Administration building lasted from August 27, 1930, until the autumn of 1931. The greatest damage to the building and its interior occurred during World War II and the 1969 earthquake. The restoration of the building was carried out in 1972, based on the original project documentation. (Vidaković, 2006).

3.1. Unmanned Aerial Vehicle DJI Phantom 4 Pro

European Aviation Safety Agency - EASA defines an unmanned aerial vehicle as an aircraft without a pilot, whose flight is controlled either autonomously or remotely by a pilot on the ground or in another vehicle.

Figure2: Unmanned Aerial Vehicle DJI Phantom 4 Pro (DJI, 2025)

The DJI Phantom 4 Pro quadcopter can reach a maximum speed of up to 72 km/h in Sport mode, with continuous operation lasting up to 30 minutes, and a maximum altitude of 6,000 meters (DJI, 2025).

The Phantom 4 Pro quadcopter is equipped with a one-inch 20 MP sensor, offering a dynamic range of nearly 12 stops. The video capability is also 4K, utilizing the H.265 codec, allowing for fast and efficient processing – 4096 x 2160 at 50 fps and 3840 x 2160 at 60 fps, both at 100 Mbps (with the H.264 codec, 4K video can reach up to 60 fps). (DJI, 2025).

3.2. Data Collection and Processing

The structure was photographed from approximately equal distances (due to the manual operation of the drone) to ensure minimal lateral and longitudinal overlap (60% longitudinal and 30% lateral) between two images. The captured images are in JPG format, and the coordinates of the control points are in a .txt file. Data processing was carried out using the 3D modeling software AgiSoft Metashape Professional.

The final result of the processing is a photorealistic 3D model that can be further used for various purposes: GIS, cultural heritage documentation, visual effects production, or indirect measurements of structures.

First, it is necessary to load and then align the photographs. The software searches for common points in the images, overlaps them, determines the camera position for each photo, and calculates the camera calibration parameters. As a result, a sparse point cloud is generated, which in this case consists of a total of 220,955 points. The sparse point cloud represents the outcome of the image alignment process and is not directly used in the subsequent steps of 3D model creation.

Based on the sparse point cloud, a dense point cloud is generated. Using the estimated camera positions and images, the software automatically creates the dense point cloud (Figure 3). The dense point cloud consists of 28,541,718 points. After that, the point cloud is georeferenced using ground control points.

DRAFT VERSION

Figure3: Dense point cloud

A dense point cloud was used to create a digital elevation model, which was then used to generate an orthophoto or orthomosaic. The surveying and modeling were performed on a section of the City Administration building in Banja Luka, which is considered one of the most significant structures for the study of architectural heritage in Yugoslavia. (Figures 4 and 5).

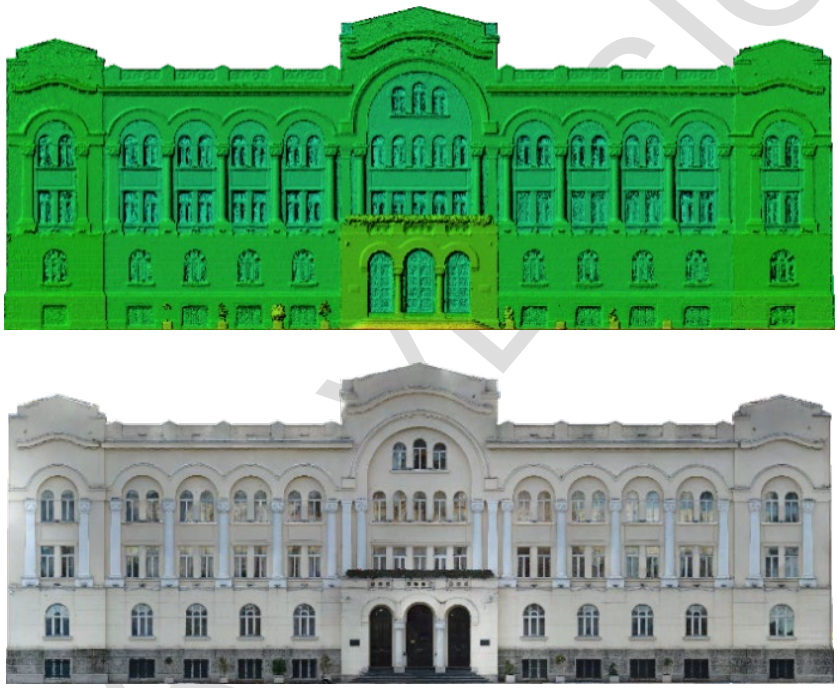
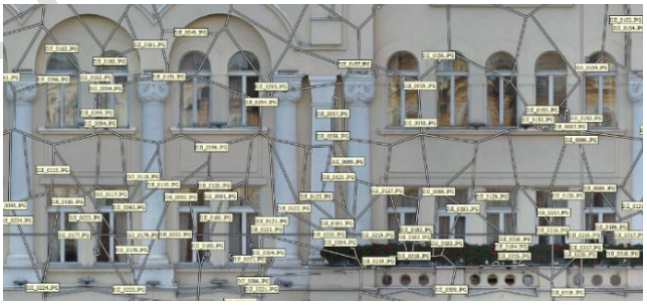


Figure4: Digital elevation model and digital orthophoto



DRAFT VERSION

Figure5: Detail of the building facade shown as an orthomosaic

Based on the generated and processed dense point cloud in Agisoft Metashape Professional, a 3D polygon mesh is reconstructed to represent the surface of the object. Generally, two algorithmic methods can be applied for mesh generation: Height Field – for flat surfaces, and Arbitrary – for all types of objects. Both methods use the dense point cloud as the basis, while the sparse point cloud is used for faster generation. The resulting geometric reconstruction can be textured or used to generate textures or orthophotos in further processing. The software SketchUp was used to create the 3D model of the building facade.

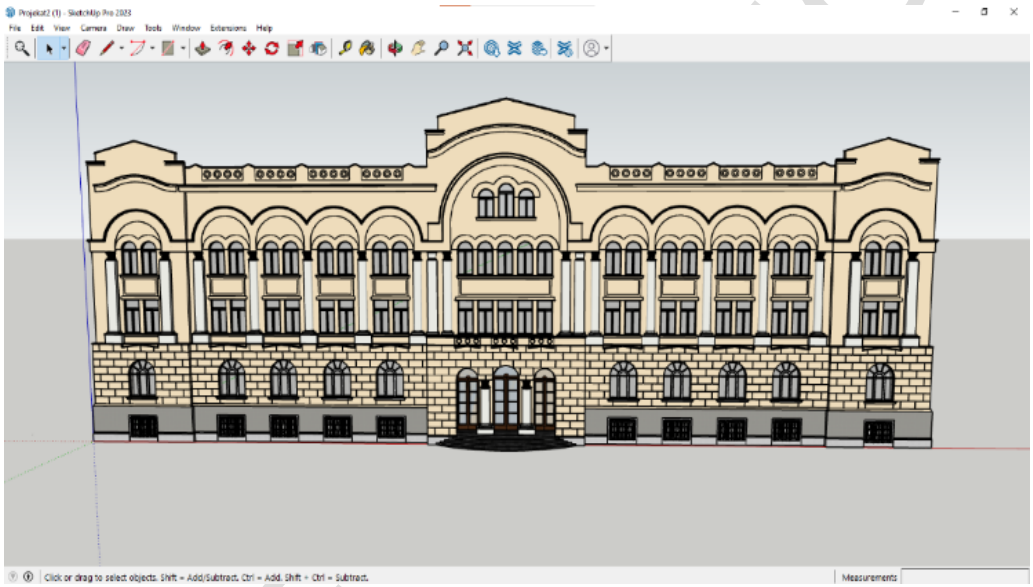


Figure6: Modeling the Facade of the Ban's Administration Building in SketchUp

In order to improve the visualization, the facade of the City Administration building in Banja Luka was rendered using the Enscape software solution. Rendering, as a final process, represents a consistent conversion of three-dimensional models into two-dimensional images. The following images show the final result of the restoration of the building. (Figures 7 and 8).

Figure7: Rendered 3D model of the facade

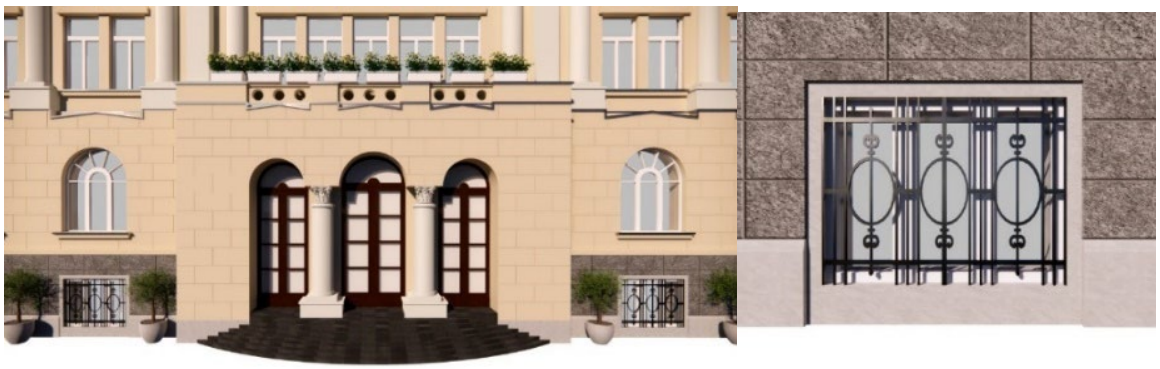


Figure8: Details of the 3D model of the facade

4. CONCLUSION

Thanks to the ease of data collection using unmanned aerial vehicles, it is now possible to survey areas of interest at any time, as well as measure hard-to-reach building elements, which is a significant advantage over traditional surveying methods. Additionally, the application of this technology leads to a considerable reduction in project costs.

One of the most efficient methods of data collection in the field used today are photogrammetry and laser scanning techniques.

The data processing workflow is user-friendly and fully automated. The software allows for editing and removal of unwanted objects, as well as the creation of additional products such as orthophoto maps, orthophoto facades, vectorization of characteristic lines, point clouds, digital models, and more.

Based on the point cloud generated from drone images, detailed and up-to-date bases for modeling the current state of structures can be obtained. The modeling and visualization of the facade for the purpose of its restoration are successfully carried out using the obtained digital building model, point cloud, and orthophoto. Additionally, the resulting model is suitable for inclusion in any existing 3D registry (3D cadastre, BIM, virtual 3D city model) and can be used in virtual and augmented reality applications.

In the field of large-scale data collection, drone surveying technology is undoubtedly becoming the technology of the future. This technology enables geodesy to expand the scope of its activities and successfully collaborate with other disciplines.

LITERATURE

Cetl, V., Tomić, H., & Lisjak, J. (2013). *Application of 3D City Model in Spatial Planning of the City of Zagreb.* Zagreb.

DJI. (2025). *DJI Phantom 4 Pro*. Retrieved from https://pcfoto.biz/dji-phantom-4-pro.html

Inc.,*3D model – SketchUp 3D Warehouse*. Retrieved from2025 sahttps://3dwarehouse.sketchup.com/warehouse/v1.0/content/public/1da022b0-856f-4b06-8563-f5c7fd344165.

Lukić, D., Terzić, M., Galjak, N., Gigović, Lj., Stojanović, M., & Kremić, Ž. (2024). *GIS multicriteria model of paths planning for potential mini-UAV attacks,* 472-479, Zlatibor, doi: 10.5937/KonGef24052L.

Mulahusić, A., Tuno, N., Topoljak, J., Balić, D., Hadžiosmanović, E., Stanić, S., & Hajdar, A. (2013). *Application of Photogrammetry and Laser Scanning in Protecting the Cultural and Historical Heritage, 47*(44), 34–57. doi:10.58817/2233-1786.2013.47.44.34

Vidaković, S. (2006). Buildings of the Ban's Court and Administration.In*Architecture of Public Buildings in Banja Luka (1918–1941)* (str. 36–51). Banja Luka: Academy of Arts RS Banja Luka.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

OPTIMIZATION OF PRECISE EDM GEODETIC CALIBRATION BASELINE

SINIŠA DROBNJAKOVIĆ[1], SRĐAN STOJAKOVIĆ[2], VOJKAN STANOJEVIĆ[3], RADOJE BANKOVIĆ[4]

[1] Military Geographical Institute "General Stevan Bošković" Belgrade, sinisa.drobnjak@vs.rs, 0000-0001-6566-5538
[2] Military Geographical Institute "General Stevan Bošković", Belgrade, sstojakovic81@gmail.com
[3] Military Geographical Institute "General Stevan Bošković", Belgrade, vojkan.stanojevic@vs.rs
[4] Military Geographical Institute "General Stevan Bošković", Belgrade, radojebankovic@yahoo.com

Abstract: *EDM calibration baselines are critical for ensuring the accuracy of distance measurements in surveying, engineering, and geodetic applications. Optimizing these baselines involves careful design, high-precision instrumentation, rigorous measurement protocols, and error mitigation strategies. This paper's primary goal was to outline the best practice for EDM calibration using the Kovin baseline. The outcomes demonstrate that, in terms of accuracy and precision, the suggested optimal approach meets all requirements.*

Keywords: *Optimization, Electronic Distance Measurement (EDM), Calibration, Baseline.*

1. INTRODUCTION

Maintaining the precision and dependability of geodetic measurements requires optimizing Electronic Distance Measurement (EDM) calibration baselines. In order to provide high-precision results, this method include choosing and setting baselines in a way that minimizes errors and uncertainties. For surveying, engineering, and geodetic applications, EDM calibration baselines are essential to guaranteeing the precision of distance measurements. Careful planning, highly accurate equipment, exacting measuring procedures, and error-reduction techniques are all necessary to optimize these baselines (Bell, 1992).

Calibration represents a set of procedures that, under certain conditions, determine the relationship between the value of the quantity shown by the scale and the corresponding value achieved by the standard. Inspection and calibration of EDM on the baseline is very important and has advantages over calibration in laboratory conditions because the expected values of the results of length measurements in real conditions, with such a determined multiplication constant of the EDM, which is most often decisive for the accuracy of length measurements, will be closer to the real ones values of those lengths (Būga, et al., 2016). In order to calibrate EDM on the baseline, it is necessary to first determine the standard distance values of the baseline, and this procedure is called *baseline calibration*.

Optimization is a science whose goal is to determine the "best" solutions for certain mathematically defined problems, which are often physical realities. The word "optimum" is synonymous with the word "best", which can mean "maximum" or "minimum" depending on the specific case, and both of these meanings are mathematically defined. Optimizing geodetic calibration baselines involves a combination of advanced mathematical methods and precise calibration techniques (Grafarend, 1974).

In the process of optimizing geodetic networks, in addition to accuracy criteria, mathematical optimization models include criteria for the reliability of geodetic networks, sensitivity and cost-effectiveness criteria (Khoptar et al., 2024). The above-mentioned criteria for the quality of geodetic networks can be included in the integral optimization model. The second-order sequential optimization method design, which was used in the optimization of the calibration of the Kovin baseline, includes criteria for the quality of geodetic networks, when the interaction of criteria occurs when solving the problem and we obtain optimal standard deviations of planned measurements. In this approach, the functional and stochastic model is identical to the mathematical model of the least squares method, that is, the indirect adjustment model (Perović, 2005).

The baseline is a unique kind of geodetic network that is positioned on a flat, roughly horizontal terrain and has points in the same vertical plane that are made visible by a certain kind of high-stability pillars. Its purpose is to ensure a metrological guaranteed calibration for all kinds of EDM. The standard value of those distances must be established initially since its distance should function as a first-order standard (Zrinjski et

DRAFT VERSION

al., 2019). This is accomplished by measuring the baseline distances using appropriate EDM or devices in accordance with a predetermined (optimum) design and measuring technique. The standard values of the baseline distances are then obtained by modifying these measurements.

2. GENERAL INFORMATION ABOUT THE KOVIN BASELINE

2.1 Location

In 1989, the KOVIN baseline was established in the area of the Kovin airport, northeast of the runway. The goal of establishing this base is to provide the primary army standard of long lengths (Bozic et al., 2013). The first measurements at the Kovin baseline were carried out in 1990, so this year we celebrate 35 years of continuous measurements (Figure 1).

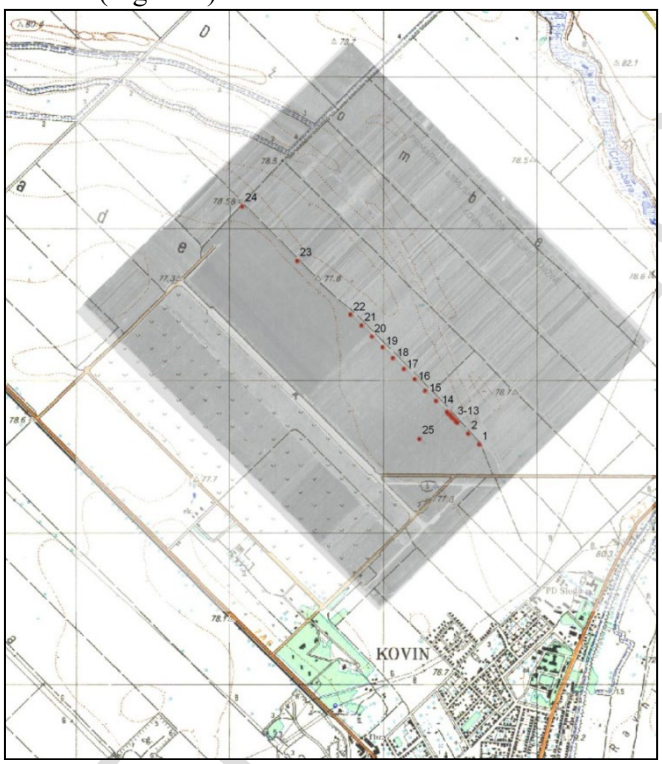


Figure 1: Location of primary army standard – KOVIN baseline

2.2 Stabilization

The points of the baseline are stabilized by reinforced concrete pillars with a foot and two Asbestos cement (AC) pipes. The external AC pipe serves to protect the pillars from meteorological influences and light mechanical shocks and rests on the foot of the pillars. Into the pillars devices for forced centering of the instrument and reflectors (stamps) are installed. The height of the pillars above the ground is 130 cm, and below it is 160 cm counted from the bottom surface rates. The foot is in the horizontal section of a square shape with a side 100 cm (Bozic et al., 2013).

2.3 Geological composition of the soil

The soil on which the KOVIN baseline was built is characterized by dusty clay of medium and low plasticity, with a fairly thick horizon. humunized brown clay soil. The geological profile of the terrain is uniform, with approximately the following layer thicknesses (Bozic et al., 2013):

- humus up to 0.6 m
- dark humunized clay up to 1.6 m
- dusty-sandy clay of medium and low plasticity up to 4.7 m
- gray-yellow highly plastic clay up to 6.3 m
- gray clayey sand (lower limit not determined)
- groundwater level from 5.2 m to 3.55 m .

The soil in the surface part is water-permeable, so that storm water quickly infiltrates into the deeper parts and does not stay on the surface.

DRAFT VERSION

2.4 Baseline geometry

The layout of the pillars of the KOVIN baseline with the numbers of that row and their mutual distances is given in Figure 2. The difference in height between the end points of the base is 0.52 m . The smallest distance between the pillars is 10 meters, and the largest is 2200 meters (Figure 2). This arrangement of baseline points ensures proper and even transfer of the unit and determination of other metrological characteristics of the instrument that provide the secondary standard (addition constant, eccentricities, etc.) (Bozic et al., 2013).

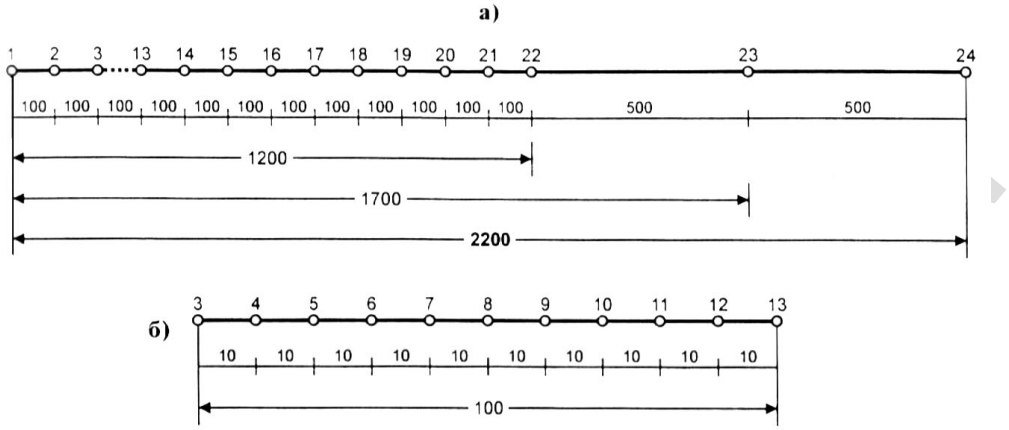


Figure 2: The layout of the pillars of the KOVIN baseline, with the numbers of the points and the distances between them: i) on the entire baseline, b) on the part between pillars 3 and 13.

2.5 Calibration EDM instrument

For the purposes of calibration measurements, we used an EDM Kern MEKOMETER ME 5000, serial number 357051. The Kern ME5000 is a high-precision electro-optical distance meter (EDM) capable of sub-millimeter accuracy (0.2 mm ± 0.2 ppm). To maximize its performance, a geodetic calibration baseline must be carefully designed, monitored, and maintained (see Table 1 and Figure 3).

Table 1: Uncertainty Budget for ME5000 (Bell, 1992)

Source	Uncertainty Contribution
ME5000 Instrumental	±(0.2 mm + 0.2 ppm)
Atmospheric Refraction	±(0.3 ppm)
Centering Error	±0.1 mm
Pillar Stability	±0.05 mm/year
Total Expanded Uncertainty (k=2)	±(0.3 mm + 0.5 ppm)

Figure 3: Kern ME 5000 laser EDM

DRAFT VERSION

3. DEFINING THE OPTIMIZATION PROBLEM

In order to research the optimal solution for calibrating the KOVIN baseline, with the EDM Kern ME 5000, the following is foreseen during work:

- On the basis of previous measurement data (with the same EDM) on Kovin baseline, investigate the accuracy of previous measurements; which should:
    - Determine the variance components within the series of measurements;
    - Determine the variance components of measurement from adjustment;
- Give a global answer about the stability of the baseline pillars;
- Based on the results of the research, make an optimal plan for observing the distance so that:
    - Reliability of observations, measured by coefficients r_{ii} matrices $R = Q_{\hat{V}} P$ be:
    - a) for a particular observation: greater than 0.6
    - b) global reliability ($\bar{r}$) : greater than 0.8;
    - When calibrating the EDM with standards measuring length *1 mm + 1 ppm* and larger, calibration errors bases ***should be negligible*** according to the principle:

$$\sigma_S = \frac{1}{3}\left(1\,\text{mm} + 1\,\text{mmkm}^{-1} D_{km}\right)$$

The true value of standard distances, in the real meaning of this word, are never known, but are always known with an error . However, if the total error of the reference distance (as measured by the standard error) is negligibly small compared to the measurement error of the EDM being calibrated, then such reference distance is considered accurate. As the principle of negligible errors of reference distances in applications, the following is use (Perović, 2005)

$$\left(\frac{\text{Mean square error of standard distance}}{\textit{Variance of measurement with calibrated EDM}}\right) \leq \frac{1}{9}$$

The components of the total error of the standard distance are (Perović, 2005):

- Determination error for standard values of baseline distance;
- Error of the standard distances due to pillars self displacement, and
- Eventual influence of movement of observers and helpers during measurement on pillars self displacement.

The displacement of the pillars-mark points of the baseline from the moment (epoch) of determining the reference values to the moment (epoch) of the EDM calibration, as a result of external forces, vehicle impact, etc., is not included in the components of the total error. Methods for calculating baseline geometry deformations by epochs are used to establish these displacements (Perović, 2005).

4. RESULTS

Second-order design optimization was employed to calibrate the Kovin baseline, resulting in the network's ideal weights or accuracy of intended observations. Finding the ideal distribution of observation tasks inside the geodetic network's chosen design is essential to achieving the best possible accuracy of the projected observations as a final outcome. Because various kinds of observations might take place in the network, these data are crucial for choosing the best instrumental equipment (Pandžić el al., 2012).

Determining the dispersion components within the series of measurements and from leveling indicates that the procedure's primary characteristic is the measurement in both stable and unstable atmospheric conditions, which led to a large dispersion of the atmospheric parameter measurement and, consequently, a large dispersion of the distance measurement. There were three successive measurements of the same distance in the series.

The length measurement variance components derived from adjucement by epochs are displayed in Table 2.

DRAFT VERSIC

Table 2: Variance components of lengths measured with EDM ME 5000 obtained from adjucement by epochs.

Epoch (Campaign)	Number of measured lengths			Variance components within epochs		
	In total	Rejected	The rest	$\sigma_1^2\left[mm^2\right]$	$\sigma_2^2\left[mm^2\right]$	Freedom degree
2022	532	39 (7%)	493	0.0930	0.2104	470
2023	182	17 (9%)	165	0.3926	0.1483	142
2024	210	19 (9%)	191	0.5832	0.3890	168
2025	135	4 (3%)	131	0.1435	0.2359	108
All epoch	1059	79 (7%)	980	0.2249	0.2405	888

According to Table 2, the definitive variance components within the epochs (for the epochs: 2022, 2023, 2024 and 2025) are:

$$\sigma_1^2 = 0{,}2249\ mm^2, \qquad \sigma_2^2 = 0{,}2405\ mm^2\ ,\ \text{with}\ \ f = 888\ \text{freedom degree}$$

so the standard *error of measuring* the distance of one kilometer is:

$$\sigma_D = 0{,}38\ mm$$

Additionally, declared accuracy amounts to $0{,}2\ mm + 0{,}2\ mmkm^{-1} \cdot 1\ km = 0{,}4\ mm$; therefore significant the difference which is inferred from *F* dispersion test:

$$F = \frac{0.38^2}{0.4^2} = 0{,}9025 < 1{,}187 = F_{0{,}999}(888, \infty)$$

- which indicates a high agreement between the two variances.

Based on the analysis of measurements on Kovin baseline carried out in previous epochs - campaigns (2022, 2023, 2024 and 2025) according to the proposed optimal observation plan and the proposed measurement method, the following *basic conclusions* are drawn:

- Since it is

$$0{,}11\ mm \le m_{\hat{D}} \le 0{,}23\ mm \qquad \text{with} \qquad \overline{m}_{\hat{D}} = 0{,}18\ mm$$

that EMD $\sigma_D = 0{,}38\ mm$ can be very reliably compared with the measurement standard on the Kovin baseline;

- The stability of the pillars on Kovin Baseline is very high;
- By using the optimal measurement plan and measurement method, the conditions defined by the optimization problem are met, ie. reference distance values of the Kovin baseline can be used for the calibration of all types of EDM, while at the same time reducing the number of measurements by more than two times compared to the previously used observation plan (epoch 2022);
- The pillars on Kovin baseline must be better secured by a protective fence.

5. CONCLUSION

The paper presents a relatively simple optimal mathematical model for determining variance components, and thus the accuracy of length measurements with EDM without the need to know the conditionally correct values of the measured lengths, and with the knowledge of true errors, they provide the opportunity to calibrate and investigate certain characteristics of EDM in an economical and efficient way.

DRAFT VERSION

According to the obtained results of the measurement analysis, it is concluded that there is a very high precision of reference distances of the KOVIN baseline, which was also contributed to by the very high stability of the pillars of the baseline.

In the conclusion, it is essential to clearly emphasize the key advantages and limitations of the research, as well as the potential practical applications of the results.

Optimizing a Kern MEKOMETER ME5000 calibration baseline requires:

- Strict pillar stability (bedrock, forced-centering).
- Precision atmospheric corrections (temp, pressure, humidity).
- Reciprocal measurements + least-squares adjustment.
- Regular validation (interferometry/GNSS).

By following this protocol, sub-millimeter accuracy can be maintained, ensuring the ME5000 operates at its full metrological potential. Optimizing EDM calibration baselines involves a combination of precise measurement techniques, robust calibration methods, and standardized baseline design. By implementing these strategies, geodesists can enhance the accuracy and reliability of EDM instruments, ensuring high-precision geodetic measurements.

ACKNOWLEDGMENT

This study was supported by the project of the Ministry of Defense 1.18.2025. Development of a primary standard for length and angle in laboratory conditions.

LITERATURE

Bell, B. (1992). *The use and calibration of the Kern ME5000 Mekometer: Proceedings* (No. SLAC-R-403; SLAC-403; CONF-9206270-). Stanford Linear Accelerator Center, Menlo Park, CA (United States).

Bozic, B., Fan, H., & Milosavljevic, Z. (2013). Establishment of the MGI EDM calibration baseline. *Survey Review*, 45(331), 263–268. https://doi.org/10.1179/1752270611Y.0000000030

Būga, A., Birvydienė, R., Kolosovskis, R., Krikštaponis, B., Obuchovski, R., Paršeliūnas, E., ... & Šlikas, D. (2016). Analysis of the calibration quality of the Kyviškės Calibration Baseline. *Acta Geodaetica et Geophysica*, 51, 505-514. https://doi.org/10.1007/s40328-015-0140-6

Grafarend, E. W. (1974). Optimization of geodetic networks. *The Canadian Surveyor*, *28*(5), 716-723. https://doi.org/10.1139/tcs-1974-0120

Khoptar, A., Lechner, J., & Vanchura, R. (2024). Prospects of Geodetic Calibration Baselines for Increasing Geodetic Measurement Accuracy. *In International Conference of Young Professionals* «GeoTerrace-2024» (Vol. 2024, No. 1, pp. 1-5). European Association of Geoscientists & Engineers.

Pandžić, J., Erić, V., & Aleksić, I. R. (2012). Second order design of geodetic network using different TK structured criterion matrices. *Tehnika*, 67(4), 530-536.

Perovic, G. (2005). Least Squares. Monograph, University of Belgrade, Faculty of Civil Engineering, Belgrade.

Zrinjski, M., Barković, Đ., & Baričević, S. (2019). Precise Determination of Calibration Baseline Distances. *Journal of surveying engineering*, 145(4), 05019005. https://doi.org/10.1061/(ASCE)SU.1943-5428.0000288

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

DETERMINATION OF THE POSITION SOLAR MODULES USING UAV PHOTOGRAMMETRY AND COMPUTER VISION

NIKOLA SANTRAČ[1], TATJANA BUDIMIROV[2], ĐURO KRNIĆ[3], MEHMED BATILOVIĆ[4], VLADIMIR BULATOVIĆ[5]

[1] University of Novi Sad, Faculty of Agriculture, Novi Sad, nikola.santrac@polj.edu.rs, 0000-0001-7183-8511
[2] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, tatjana.kuzmic@uns.ac.rs, 0000-0001-8586-4382
[3] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, djuro.geo@uns.ac.rs, 0000-0003-4481-3209
[4] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, mehmed@uns.ac.rs, 0000-0001-5605-1112
[5] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, vbulat2003@uns.ac.rs, 0000-0002-0772-5333

Abstract: *Solar modules are crucial in renewable energy production, especially in large-scale solar farms. Precise spatial location of individual modules is essential for monitoring, maintenance and detection of potential damage. This study combines Unmanned Aerial Vehicle (UAV) photogrammetry and computer vision techniques to determine the 3D coordinates of solar modules. An UAV equipped with RGB and thermal cameras captured the area of interest. The You Only Look Once (YOLO) algorithm was used to automatically detect modules in thermal images, and the same modules were tracked across multiple images by applying keypoint matching and homography estimation in order to determine coordinates.*

Keywords: *Photogrammetry, Computer vision, YOLO, UAV, Thermal images, Solar Farm.*

1. INTRODUCTION

Solar farms involve the installation of a large number of solar modules in one place, with the aim of capturing solar energy and converting it into electricity. The electricity produced in this way is fed directly into the power grid and serves the electrical needs of local buildings, homes and larger facilities, industries, companies, etc. Increasing support for the development of solar modules is expected, in line with the European Green Deal initiative. This initiative aims to make Europe a climate-neutral continent by 2050. One of the envisaged strategies is exactly the use of renewable energy sources, including solar farms, in order to stop using fossil fuels (European Commission, 2019).

An UAV is a part of UAS (Unmanned Aerial System) that usually consists of an aircraft platform mounted with one or more sensors combined with a ground-based control station from where it is operated. An aerial platform is a basic component that carries with it various sensors and devices that enable safe flight of the aircraft and data collection. It contains an engine that provides flight, a Global Navigation Satellite System (GNSS) antenna and an Inertial measurement unit (IMU) that can determine the position and orientation of the aircraft at any time, a battery that powers all electronic components during the aircraft flight and camera that collects high-resolution images and videos (Marković et al., 2021). In general, UAVs can be equipped with different types of sensors (RGB, thermal, multispectral, LiDAR) (E. J. Lee et al., 2016; Morales et al., 2020; Speth et al., 2022). Thermal camera uses the information at the emitted radiation in the thermal infrared range of the electromagnetic spectrum and converts this information into temperature. In such way it is possible to easily detect solar modules (Messina & Modica, 2020).

The concept of UAV imaging and photogrammetry involves firstly to define the area of interest, and based on that, to create a flight plan. During the flight, the UAV continuously collects images and records measured data from other mounted sensors such as GNSS, IMU and stores them in Exchangeable Image Format (EXIF). The key thing in the application of the photogrammetric method is to collect successive images with an appropriate overlap between them. It is this overlap that allows the processing of the collected images in order to create a three-dimensional model of the captured area.

Object detection is a kind of image segmentation based on object geometric and statistical features, also known as object extraction. One of the most commonly used object detection method, in case of UAV photogrammetry, is YOLO. This method completely abandons the detection mode of "region candidate + regression" in the two stages, and hence, this is what makes this method very fast and suitable for near real-

DRAFT VERSION

time object detection based on UAV images. YOLO scales the image to be measured into a uniform size and divides it into multiple grids, then predicts the target category according to the grid where the target center is located, and outputs the detection results on the last convolution layer. The core idea of YOLO is to transform target detection into a regression problem, using the whole image as the input of the network, and just through a neural network, the location of the bounding box and its category can be obtained (Cong et al., 2023; Krishna et al., 2021).

2. DATA AND METHODS

This chapter presents the entire start-to-end process, from collecting UAV imagery to formating a mathematical model for determine three-dimensional module coordinates based on two-dimensional measurements. For clarity, each element in this process, such as workflow overview, UAV configuration, camera orientation parameters, YOLO detection, and planar homography estimation, is covered in a separate subsection.

2.1. Workflow Overview

As illustrated in Figure 1, the workflow proceeds in the following stages. First, the UAV acquires thermal images and records an EXIF file containing GNSS and IMU-based position metadata. Second, the YOLO model is applied to the thermal frames to define a bounding box for each module. Third, the feature points extracted from the overlapping images provide the point pairs needed to estimate the planar homography. Finally, these two-dimensional observations from multiple views, along with internal and external orientation parameters, are input into a triangulation process that determines the three-dimensional coordinates of all modules.

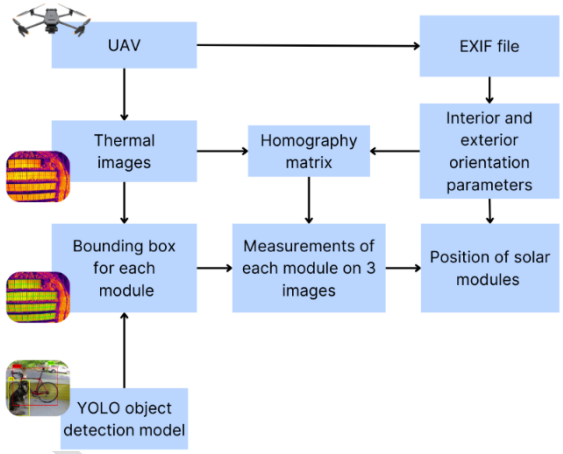


Figure 1: Workflow for determining the position of solar modules

2.2. UAV Platform and Flight Configuration

The research was conducted in Brazil on a 143 ha solar farm using the DJI Mavic 3 Thermal (Global Power Generation, 2018). The aircraft provides up to 45 minutes of flight, enough to cover solar farms of approximately 200 ha in one go at an altitude of 40 meters. The specifications of the UAV can be found in Table 1. The solar farm was surveyed with front overlap of 80% and 60%, and side overlap of 60% and 45% for RGB and thermal images, respectively, at an altitude of 40 meters. A total of 436 RGB and 436 thermal images were collected.

Table 1: Key specifications of the DJI M3T (dji, 2024.)

Parameter	Specification
Weight (with propellers)	920 g
Max flight time (no wind)	45 min
Wide camera	1/2-inch CMOS, Effective pixels: 48 MP
Tele camera	1/2-inch CMOS, Effective pixels: 12 MP
Thermal camera	Uncooled VOx microbolometer; 640 × 512 px @ 30 Hz
GNSS	GPS, Galileo, BeiDou (GLONASS supported with RTK)

DRAFT VERSION

2.3. EXIF Metadata and Camera Orientation Parameters

Metadata for each image are included in EXIF file using a UAV equipped with GNSS/IMU sensors. The file contains the parameters of internal orientation (the intrinsic matrix **K**) and external orientation (rotation matrix **R** and translation vector **t**), as well as additional information such as image dimensions, acquisition time, etc. (Verykokou & Ioannidis, 2016). The mathematical representation of the said matrices is as follows:

$$\boldsymbol{K} = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_x & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

where f_x and f_y represents focal length, c_x and c_y represent the principal point and s represents skew. The rotation matrix **R** is equivalent to the following formulations:

$$\boldsymbol{R} = \boldsymbol{R_\omega} * \boldsymbol{R_\varphi} * \boldsymbol{R_k} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} = \tag{2}$$

$$\begin{bmatrix} \cos(\varphi)\cos(\kappa) & -\cos(\varphi)\sin(\kappa) & \sin(\varphi) \\ \cos(\omega)\sin(\kappa) + \sin(\omega)\sin(\varphi)\cos(\kappa) & \cos(\omega)\cos(\kappa) - \sin(\omega)\sin(\varphi)\sin(\kappa) & -\sin(\omega)\cos(\varphi) \\ \sin(\omega)\sin(\kappa) - \cos(\omega)\sin(\varphi)\cos(\kappa) & \sin(\omega)\cos(\kappa) + \cos(\omega)\sin(\varphi)\sin(\kappa) & \cos(\omega)\cos(\varphi) \end{bmatrix},$$

where ω, φ, and κ are the Euler angles of rotation around X, Y and Z axis, and $\boldsymbol{R_\omega}$,$\boldsymbol{R_\varphi}$, $\boldsymbol{R_k}$ denote the corresponding axis aligned rotation matrices. The vector **t** represents the translational (positional) vector of the camera, i.e. the three-dimensional coordinate of the projection centre in the global coordinate system. Its components define how much the coordinate origin is shifted relative to the reference coordinate system and are expressed by the formula:

$$\boldsymbol{t} = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}, \tag{3}$$

where t_x, $t_y$and t_z denote the camera-centre coordinates along the X, Y and Z axes.

2.4. Module Detection with YOLOv8

One of the widely used artificial intelligence and computer vision methods for object detection in images is the YOLO (Ultralytics YOLO8, 2024). The original YOLO model processes the entire image in a single pass through a Convolutional Neural Network (CNN) and can achieve real-time detection, which has made it the standard for UAV applications (Joseph Redmon et al., 2016). By introducing an anchor-free detection head, YOLOv8 directly regresses box coordinates and objectless scores without requiring manually defined anchors, which reduces the number of hyperparameters and speeds up model configuration (Ali & Zhang, 2024). The modifications to the YOLOv8 method lead to significant improvements in detection reliability, making it ideal for real-time applications in various areas of computer vision. Figure 2 illustrates an example of the YOLO algorithm applied for object detection in an image.

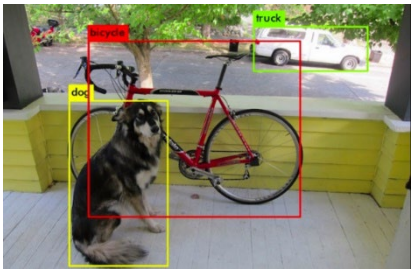


Figure 2: Example of operation of Darknet YOLO (Kanehisa & Neto, 2019)

2.5. Planar Homography Estimation

When two cameras observe the same physical plane from different viewpoints, their images are related by a unique 3 × 3 projective transformation called homography. This matrix compactly encodes the full

DRAFT VERSION

perspective warp (combining rotation, translation, scaling, shearing and foreshortening) and is the cornerstone of tasks such as image rectification, mosaicking and planar 3D reconstruction. The homography matrix **H** defines a projective transformation between two overlapping images. It establishes a geometric relationship that maps points from one image to the corresponding points in the other (2D to 2D point correspondences). Estimating of planar homography matrix **H** using the Direct Linear Transform (DLT) method (Hartley & Zisserman, 2004), is based on the keypoints detected on both images with scale-invariant feature transform (SIFT) algorithm (Lowe, 2004). The matrix **H** represents a wide range of geometric transformations, including rotation**,** translation**,** scaling**,** shearing, and perspective distortion**,** all combined within a single 3×3 matrix (Hartley & Zisserman, 2004). Homography describes the transformation using the following Equation 4 as in the example in Figure 3:

$$\tilde{x}' = \left[\widetilde{\boldsymbol{H}}\right]_{3x3}\tilde{x}, \tag{4}$$

where $\tilde{x}$ denotes a point in the first image, $\widetilde{\boldsymbol{H}} \in \boldsymbol{R}^{3x3}$ is the homography matrix, and $\tilde{x}'$ denotes the corresponding point in the second image.

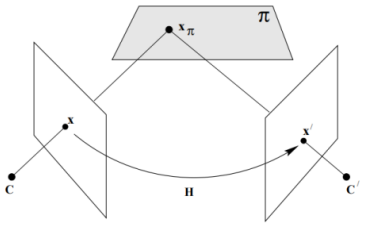


Figure 3: Illustration of the planar homography (Hartley & Zisserman, 2004)

In this way, two independent projections of the same object (in this case, a module) are obtained and a condition for further 3D reconstruction. The intersection of rays using the triangulation method yields a 3D coordinate, a standard procedure in multi-view photogrammetry (Hartley & Zisserman, 2004). The complete calculation of 3D coordinates in the global coordinate system using 2D measurements in the images is described by Equation 5:

$$\begin{bmatrix} X_S \\ Y_S \\ Z_S \end{bmatrix} = \boldsymbol{R}\begin{bmatrix} \frac{(u-c_x)*Z}{f_x} \\ \frac{(v-c_y)*Z}{f_x} \\ Z \end{bmatrix} + \boldsymbol{t}, \tag{5}$$

where u and v are the 2D coordinates of the point in images (measured values), Z is the point depth (distance from camera center to 3D point), f_x, c_x and c_y are the instrict camera parameters, **R** and $\boldsymbol{t}$ denote the rotation matrix and translation vector (camera center) in the global coordinate system, and $(X_S, Y_S, Z_S)^T$ are the final 3D coordinates of a point in the global coordinate system (unknown values).

From the Equation 5, the unknown values are the 3D coordinates, while the measured data are the corresponding 2D image coordinates u and v. In order to solve for the three unknowns the same point must be detected in at least two images taken from different viewpoints. This requirement arises because a single 2D observation constrains the 3D point to lie along a ray; the intersection of at least two such rays (from different camera positions) is necessary to compute the 3D location through triangulation.

3. RESULTS

Using a trained YOLO object detection model, individual solar module were detected in all thermal images, as illustrated in Figure 4. The output from the YOLO detection process was saved in a .json file, where the bounding boxes are encoded using normalised coordinates relative to the image dimensions (ranging from 0 to 1), with the top-left corner as the origin. This method provides only the 2D coordinates of the bounding box corners, but its format facilitates automated processing and seamless integration into Geographic information system (GIS) workflow systems (Ultralytics YOLO8, 2024).

After detecting the modules in the thermal images, keypoints were extracted from all images using the SIFT algorithm. Each image was matched with its two nearest neighbours (those with the highest overlap) and homography matrices were computed using feature matching techniques. Once established, these homography matrices can be used to correspond any pixel from between two images, effectively enabling the localisation of the same object in multiple views. By identifying the same object in three overlapping

DRAFT VERSION

images, a total of six 2D measurements (two per image) are obtained, which is sufficient to compute the 3D coordinates of the object using triangulation.

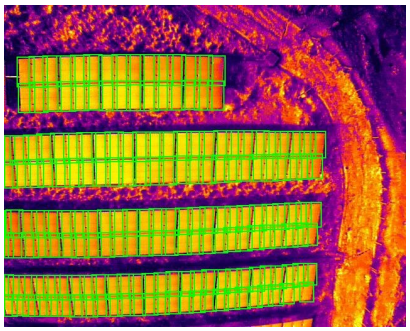


Figure 4: Thermal images with detected bounding box of solar farm

Using Equation 5, the 3D coordinates of each solar module were calculated and written in the .json file originally generated by the YOLO object detection algorithm. As a result, the final .json file contains the 2D coordinates of the four bounding box corners, corresponding latitude and longitude coordinates for each of those four points, and the coordinates of the module centre, as illustrated in Figure 5.

This approach enables the extraction of each module's coordinates with satisfactory accuracy without the need for time-consuming processes such as aligning all images, generating a digital surface model, and creating an orthophoto for the entire area (a processing pipeline that can be particularly slow when dealing with a large number of images). Based on the obtained coordinates, further analyses and measurements can be performed on the modules, such as detecting tracker deformation caused by damaged support structures, calculating the width and height of solar modules in the 3D coordinate system, and easily locating individual modules in the field by exporting their coordinates to a KMZ file or similar formats.


```
{
  "TagId": "77834cd9-66af-414c-b1f3-67beb950034b",
  "Name": "modulo",
  "Confidence": 0.99912983,
  "Coordinates": {
    "Center": "-18.828124703669985, -46.676876700640044",
    "TopLeft": "-18.828129170515663, -46.676885457448",
    "TopRight": "-18.82812047007784, -46.676885585755485",
    "BottomLeft": "-18.82812893064024, -46.67686783832317",
    "BottomRight": "-18.828120243841486, -46.67686796631655"
  },
  "BoundingBox": {
    "Top": 0.38172325,
    "Left": 0,
    "Width": 0.8378141,
    "Height": 0.19233504
  }
}
],
"ThermalImage": "AEROID250312100353T-DJI_20250312100353_0006_T.JPG"
```


Figure 5: Example of final JSON file

5. CONCLUSION

The paper presents the application of photogrammetry and computer vision for the purposes of detection and determination of the coordinates of each solar module. As mentioned, the computer vision method itself detects modules in a local coordinate system, in the form of two-dimensional images. By integrating with photogrammetry, georeferencing is enabled, i.e., the determination of the actual spatial coordinates of the modules in the desired coordinate system. The method presented in the paper has proven to be very practical in the case of large solar farms, where a large number of images is collected. Thermal cameras have a smaller field of view compared to RGB cameras, so the only prerequisite is that the thermal images are collected with a large overlap, so that the modules can be identified using the detection algorithms. Otherwise, in the case of a smaller overlap, the processing would be very long and difficult. In general, the use of thermal cameras provides much more information in this case and allows for the use of more object detection methods than RGB cameras. In addition, on thermal images, it is possible to detect damage on the modules using computer vision algorithms and photogrammetry. Hence, by using the presented method, it is possible to extract the coordinates of damaged modules, which greatly reduces the time in the office required to detect damaged modules, as well as the time until these modules are found in the field.

ACKNOWLEDGMENT

This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-137/2025-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad 2025" (No. 01-50/295).

DRAFT VERSION

This paper is supported by Ministry of Science, Technological Development and Innovation of the Republic of Serbia (451-03-137/2025-03/200117).

LITERATURE

Ali, M. L., & Zhang, Z. (2024). The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection. *Computers*, 13(12), 336. https://doi.org/10.3390/computers13120336.

Cong, X., Li, S., Chen, F., Liu, C., & Meng, Y. (2023). A Review of YOLO Object Detection Algorithms based on Deep Learning. *Frontiers in Computing and Intelligent Systems, 4*(2), 17-20. https://doi.org/10.54097/fcis.v4i2.9730.

dji. (2024). Specifications of the DJI Mavic 3T. Retrieved May 11, 2025, from https://enterprise.dji.com/mavic-3-enterprise/specs.

The European Green Deal (2019), https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en, accessed on 13th May 2025.

Global Power Generation. (2018). https://ippjournal.com/project/82-5-mw-guimarania-solar-project-in-minas-gerais, accessed on 03th August 2025.

Hartley, R., & Zisserman, A. (2004). Multiple View Geometry in Computer Vision (2nd ed.). *Cambridge University Press*. https://doi.org/10.1017/CBO9780511811685.

Kanehisa, R., & Neto, A. (2019). Firearm Detection using Convolutional Neural Networks. *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*, 707–714. https://doi.org/10.5220/0007397707070714.

Krishna, N.M., Reddy, R.Y., Reddy, M.S.C., Madhav, K.P. & Sudham, G. (2021). Object Detection and Tracking Using Yolo. *Third International Conference on Inventive Research in Computing Applications (ICIRCA),* Coimbatore, India, 2021, 1-7, doi: 10.1109/ICIRCA51532.2021.9544598.

Lee, E. J., Shin, S. Y., Ko, B. C., & Chang, C. (2016). Early sinkhole detection using a drone-based thermal camera and image processing. *Infrared Physics & Technology*, 78, 223–232. https://doi.org/10.1016/j.infrared.2016.08.009.

Lopez-Franco, C., Fink, G., Arana-Daniel, N., & Alanis, A. Y. (2011). A visual servo control based on geometric algebra. *2011 8th International Conference on Electrical Engineering, Computing Science and Automatic Control*, 1–6. https://doi.org/10.1109/ICEEE.2011.6106633.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94.

Marković, M., Budimirov, T., Batilović, M., Bulatović, V. & Sušić, Z. (2021). Influence analysis of the number and distribution of control points on the quality of data collected by UAV. *iNDiS 2021 – Planninh, design, construction and building renewal*, Novi Sad, Republic of Serbia, 24-26th November 2021, 420-427.

Messina, G., & Modica, G. (2020). Applications of UAV Thermal Imagery in Precision Agriculture: State of the Art and Future Research Outlook. *Remote Sensing, 12*(9), 1491. https://doi.org/10.3390/rs12091491.

Morales, A., Guerra, R., Horstrand, P., Diaz, M., Jimenez, A., Melian, J., Lopez, S., & Lopez, J. F. (2020). A Multispectral Camera Development: From the Prototype Assembly until Its Use in a UAV System. *Sensors*, 20(21), 6129. https://doi.org/10.3390/s20216129.

Redmon Joseph, Santosh Divvala, Ross Girshick, & Ali Farhadi. (2016, August 6). You Only Look Once: Unified, Real-Time Object Detection. *Computer Vision Foundation*.

Speth, S., Gonçalves, A., Rigault, B., Suzuki, S., Bouazizi, M., Matsuo, Y., & Prendinger, H. (2022). Deep learning with RGB and thermal images onboard a drone for monitoring operations. *Journal of Field Robotics*, 39(6), 840–868. https://doi.org/10.1002/rob.22082.

Ultralytics YOLO8. (2024). Ultralytics YOLO Docs. Retrieved May 11, 2025, from https://docs.ultralytics.com/models/yolo8/.

Verykokou, S., & Ioannidis, C. (2016). Exterior orientation estimation of oblique aerial imagery using vanishing point. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B3, 123–130. https://doi.org/10.5194/isprsarchives-XLI-B3-123-2016.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original paper

SPATIAL PLANNING IN THE REPUBLIC OF CROATIA

BOJAN LINARDIĆ[1], VLADIMIR BARIČEVIĆ[2], LUKA BABIĆ[3] , IVAN LANDEK[4], ROBERT PAJ[5]

[1] Ministry of Physical Planning, Construction and State Assets, Zagreb, bojan.linardić@mpgi.hr
[2] Ministry of Physical Planning, Construction and State Assets, Zagreb, vladimir.baricevic@mpgi.hr
[3] Ministry of Physical Planning, Construction and State Assets, Zagreb, luka.babic@mpgi.hr
[4] Ministry of Physical Planning, Construction and State Assets, Zagreb, ivan.landek@mpgi.hr
[5] Ministry of Physical Planning, Construction and State Assets, Zagreb, robert.paj@mpgi.hr

Abstract: *Spatial planning is one of the foundations for the development of an area, city or municipality, county or the entire state. It ensures the conditions for the use (management), protection, and governance of the territory of the Republic of Croatia, including its exclusive economic zone, as a particularly valuable and limited national asset. In doing so, it provides prerequisites for social and economic development, environmental and nature protection, quality construction, and rational use of natural and cultural resources. In Croatia, spatial planning is conducted based on legislation and regulations, in digital form via the Spatial Planning Information System (ISPU). This paper provides an overview of spatial plans and their transition to the new spatial planning system in the Republic of Croatia.*

Keywords: *Spatial planning, ISPU, new generation spatial plans*

DRAFT VERSION

1. INTRODUCTION

Spatial planning provides the conditions for the use, protection, and governance of the territory of the Republic of Croatia, recognized as a particularly valuable and limited national good. It supports the prerequisites for social and economic development, environmental and nature protection, high-quality construction, and the rational use of natural and cultural resources (URL 1).

The goals of spatial planning include: balanced spatial development aligned with economic, social, and environmental frameworks, spatial sustainability regarding rational use and conservation of spatial capacities on land, sea, and seabed for effective spatial protection (NN 153/2013). Cross-border influences are taken into account, linking Croatia with European spatial planning systems while fostering and developing regional spatial characteristics. Special attention is given to the harmonious distribution of human activities in space for functional and sustainable community development, the protection of integral spatial values, reasonable use and conservation of natural assets, nature protection, and pollution risk prevention. Cultural and other significant assets enjoy special protection ensured through spatial plans.

Successful spatial planning involves well-organized land use, high-quality and humane development of urban and rural settlements, and a safe, healthy, and socially functional living and working environment. It also encompasses the integrity of valuable coastal ecosystems, sea quality, adequate transport systems (especially public transport), accessibility to services, and inclusive urban design for children, elderly, and persons with disabilities. Design quality and spatial aesthetics, high-value built environments, and respect for natural and urban landscapes and heritage, especially tourism development areas, are core components. Planning also addresses economic development, national security, and protection from natural and other disasters. Spatial planning goals are achieved through applying planning principles during plan preparation, adoption, and implementation (NN 153/2013).

Spatial planning is based on principles of integrated approaches, scientifically and professionally verified facts, sustainable development, public and private interest protection, horizontal and vertical integration, and public access to data and documents relevant to spatial planning (NN 153/2013).

2. SPATIAL PLANNING

As an interdisciplinary activity, spatial planning is both an institutional and technical framework for managing the spatial dimension of sustainability. It is based on the assessment of: (1) development potentials respecting spatial identity, (2) space protection requirements, and (3) environmental and nature quality

preservation. It defines land uses, development conditions for activities and infrastructure, spatial allocation, urban renewal conditions, and procedures for implementing planned interventions (NN 153/2013). These elements align with the classical definition of effective spatial planning as an institutional function of development governance (Faludi, 2000).

In Croatia, spatial planning is implemented at several levels, producing various plans: the State Spatial Development Plan (DPPR), Special Area Plans (PPPPO), State-level Urban Development Plans (UPU), County Spatial Plans (PPŽ), the City of Zagreb Spatial Plan (PPGZ), County-level Urban Development Plans, and Municipal/City Spatial Plans (PPUO/PPUG), General Urban Development Plans (GUP), and Urban Development Plans (UPU) (NN 153/2013).

Plan preparation begins with the Spatial Development Strategy of the Republic of Croatia, adopted by the Ministry and confirmed by the Croatian Parliament. Lower-level plans must align with higher-level ones. Each plan includes implementation provisions (legal regulations), a graphic component (maps), and explanatory text (analysis, goals, and rationale).

2.1. "Old" Generation Plans

Prior to the new Regulation (NN 152/2023), spatial plans were prepared according to the 1998 Regulation (NN 106/1998). Plans consisted of a textual component in .doc format and graphic elements created in CAD software, printed, signed, and sealed. Approval required multiple signatories, including local mayors or county prefects. The process consumed significant energy, resources (paper and ink), and time. Data access was difficult, analyses cumbersome, and inconsistencies arose due to differing projections (Gauss-Krueger zones 5 and 6) and non-unified cartographic symbology. These issues prompted the transition to the ISPU system.

2.2. "New" Generation Plans – Spatial Planning Information System (ISPU)

With the adoption of the new Regulation on Spatial Plans (NN 152/2023) on January 1st, 2024, spatial planning entered a new phase. This digital transformation, developed over several years, is recognized as a major innovation in Croatian spatial governance (Janković & Tomić, 2019). ISPU enables plan preparation, adoption, implementation, oversight, spatial monitoring, and reporting in digital form. It integrates systems from various public bodies producing or maintaining spatial data.

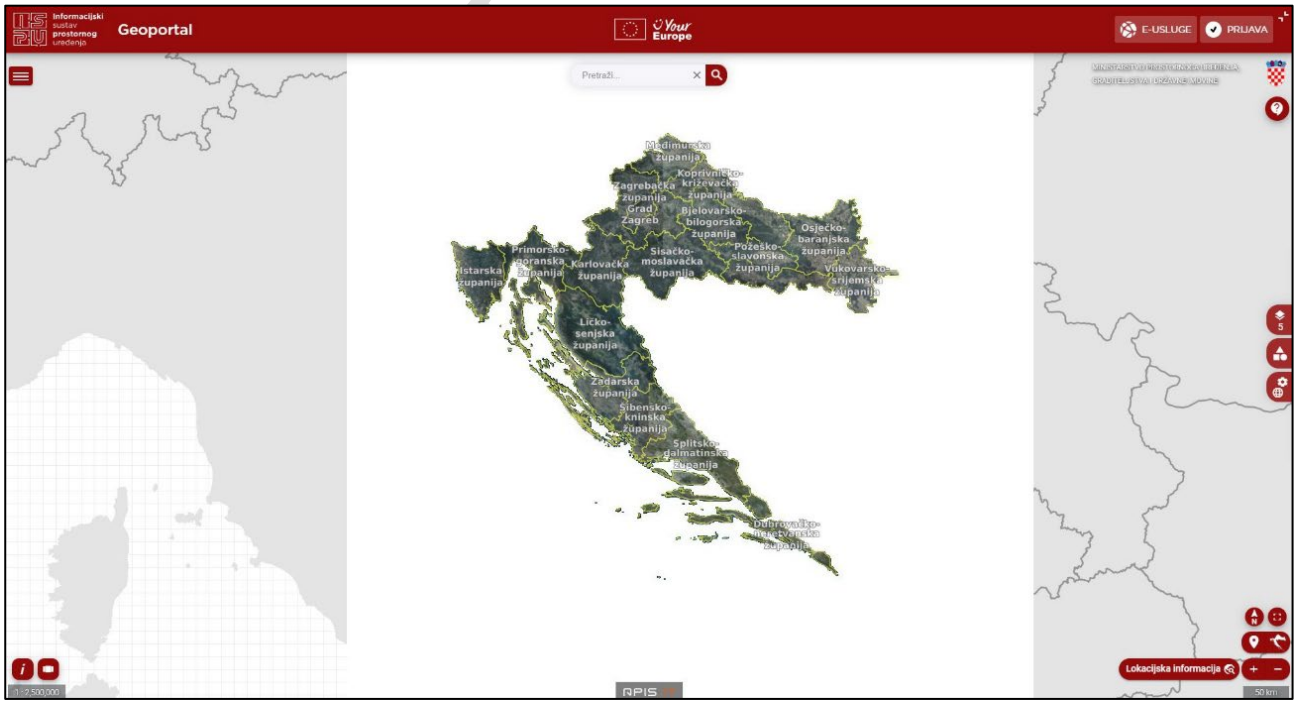


Image 1: ISPU

This architecture ensures compliance with the INSPIRE Directive principles on spatial data infrastructures in the EU (Nebert & Vretanos, 2007). ISPU is managed electronically and uses the official Croatian coordinate system. (Image 1).

VERSION

ISPU includes data on land use, property ownership, existing and draft spatial plans, land use designations, legal and physical planning conditions, construction permits, usage, demolition, and sectoral policies and documents.

Authorities must provide all relevant spatial data through ISPU, which is publicly accessible.

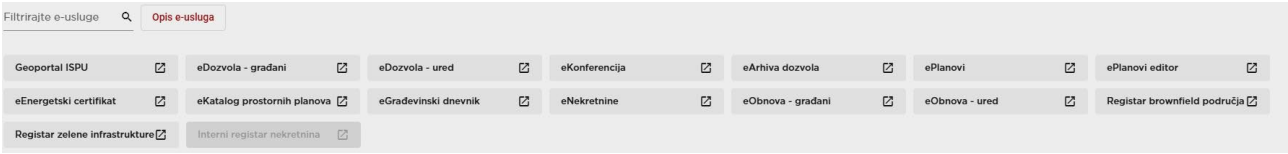


Image 2: ISPU modules

ISPU contains 15 modules; the most relevant for planning are ePlans and ePlans Editor (Image 2). These support both the development of new plans and the transformation of analogue plans. Planners define the scope, assign roles, manage public consultations, and adopt plans within the system. Editing occurs in ePlans Editor, linking vector graphics and text. Plans may also be prepared in compatible software ensuring spatial-textual linkage (NN 153/2013).

Plans are verified for topology and compliance, then transferred into ePlans for adoption. The adoption decision is generated from ePlans Editor data and templates. The legend is auto-generated, and metadata ensures clarity. A digital plan dossier (PDF) is created for publication (NN 153/2013).

2.3. Data Transformation

All analogue spatial plans in raster form have been imported into ISPU. These were mainly CAD-based with symbology defined by the 1998 regulation (Image 3). Due to inconsistencies, overlapping plans, and outdated projections (HDKS), transformation into vector format is required. This ensures better integration and alignment between neighboring plans. Transformation occurs within ePlans similarly to new plan creation but skips certain steps since plans are already in force.

After transformation, old generation data provides a reliable basis for new plans. Combined with statistical and registry data, transformed plans support enhanced spatial analysis and decision-making.

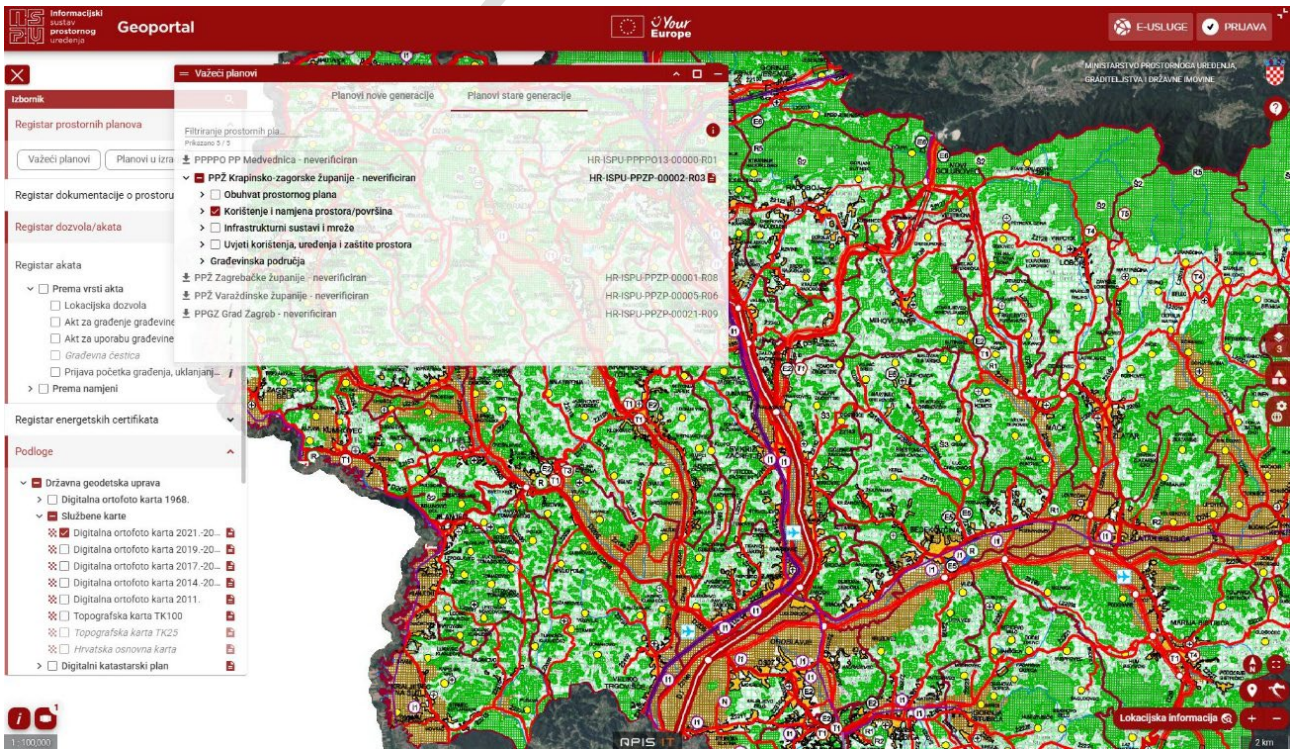


Image 3: ISPU – "Old" Generation Plan

PREVIEW VERSION

3. ISPU MODULES AND TOOL FOR THE PUBLIC

The registry of all spatial plans, those in-force and the outdated ones, old and new generation ones, can be found using the eCatalogue module of ISPU.

By searching for in-force plans made after January 1st, 2024,through the eCatalogue module,over 7.200 resultswere obtained in July 2025.Some of those results are related to translated plans while others to modified or completely new ones. As Croatia has allocated funds for the translation of old generation plans, this number is expected to grow rapidly in the coming years. There are numerous filtering options available, and a search bar, allowing users to easily navigate to their area of interest, and to inspect the plans. It is simply a matter of clicking on one of the search results to pan and zoom to the desired location (image 4).Clicking on the arrow button within the result allows users to access data on the plan itself, and its provisions (image 5). The module makes spatial plans easily accessible to the public from any place, at any time.

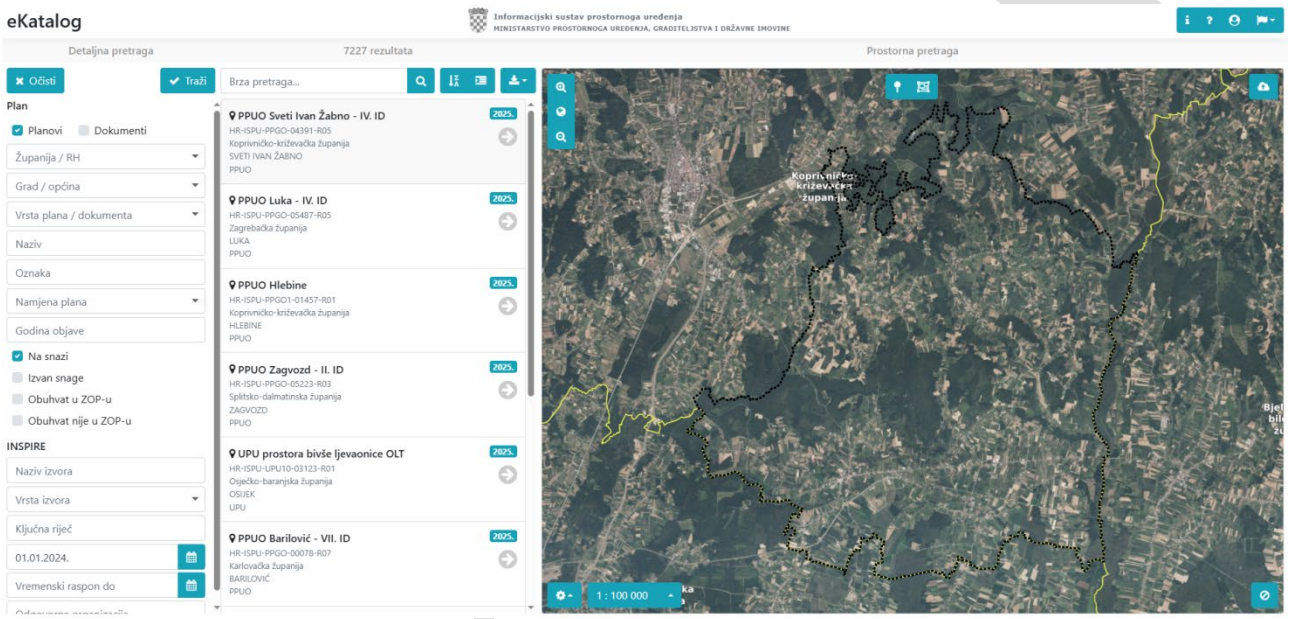


Image 4:eCatalogue module search for new generation spatial plans

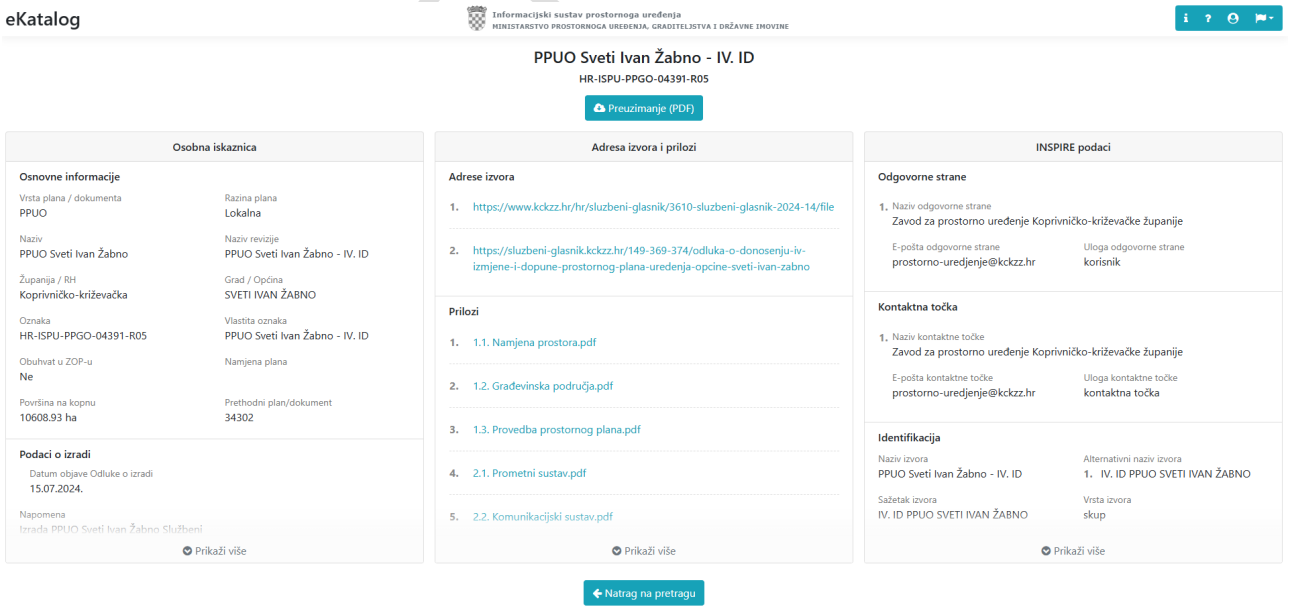


Image 5: accessing data and provisions of the spatial plan

Another tool useful to the public is the location information tool (Croatian “Lokacijska informacija”) that can be found within the main ISPU interface. By clicking on the button, a window pops up where different selection methods and types of information to be found can be chosen. Once the location has been confirmed,

the requested search data and easy access to it is provided in the pop-up window. The example of one such search is given in image 6 where the search entails acquisition of information of plans that are in-force for the selected point location.

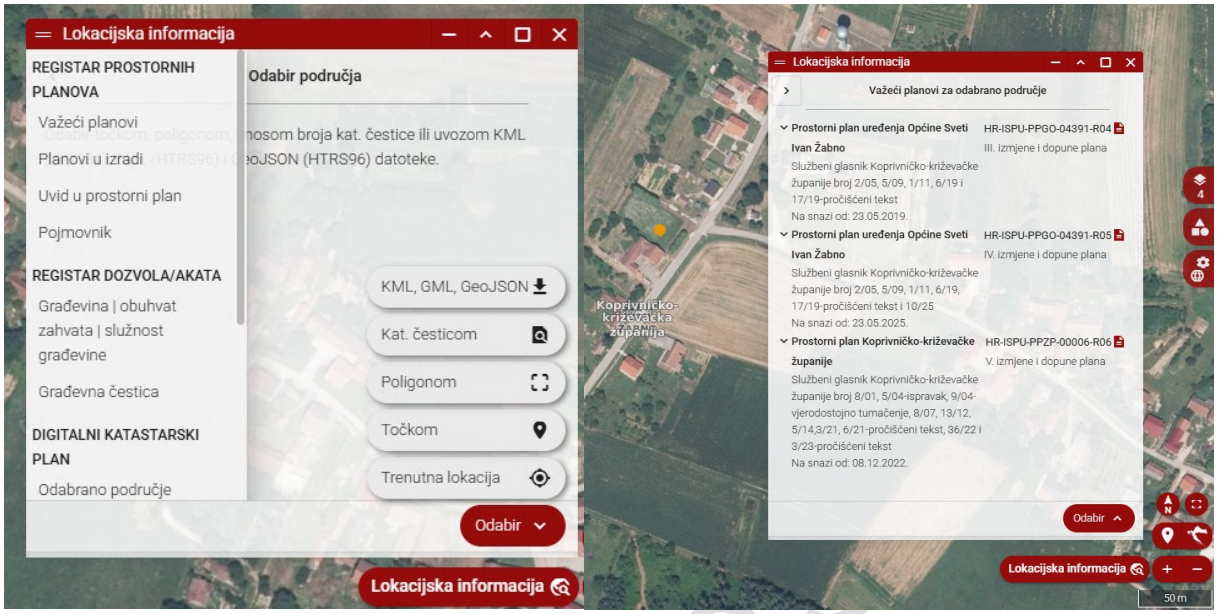


Image 6: Location tool application

A user also has an option tomanually inspect different areas of interest, use the main toolbar (on the left), find the plans in the Spatial plan registry, and display its different thematic information on the map. One can view the basic distribution of intended uses, infrastructure systems, and special provisions if there are any for the area. The user only needs to use the appropriate zoom level for each plan to display it as local, regional and national level plans are visible on different level scales.

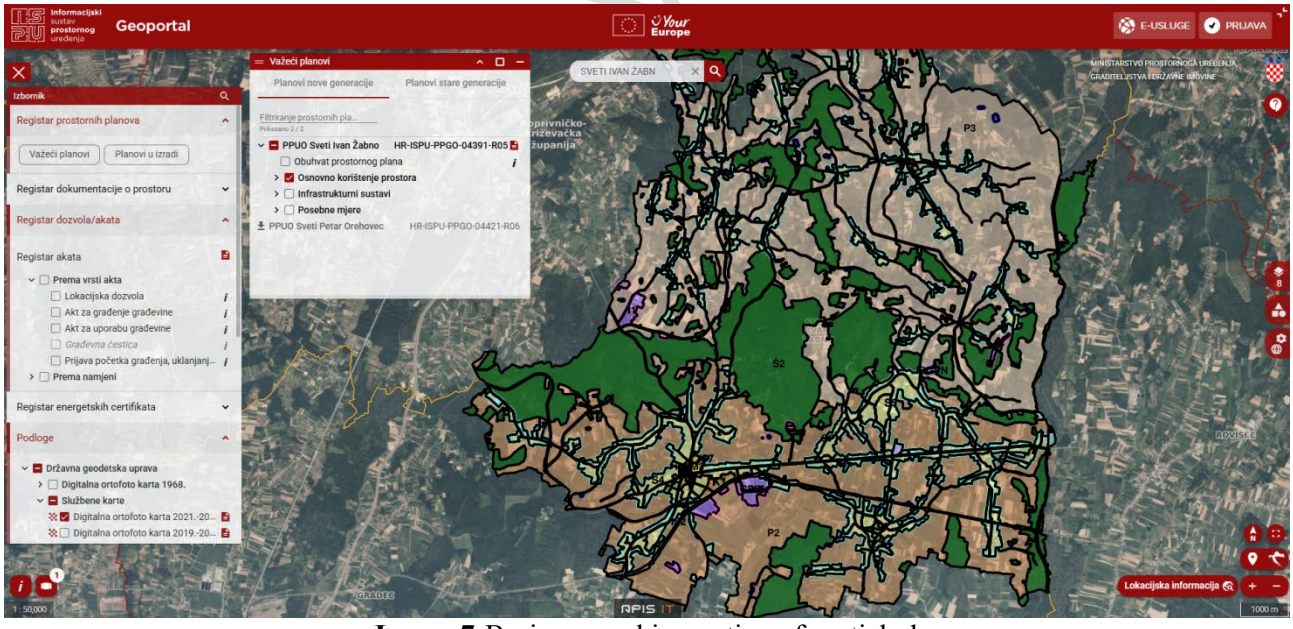


Image 7: Basic manual inspection of spatial plans

4. CONCLUSION

Spatial plans function as by-laws, requiring clarity and uniformity for legal certainty. Through ISPU and plan transformation, transparency and consistency have improved. All data processed via ePlans Editor follow Croatia's official coordinate system (HTRS96/TM), using standardized regulations and nomenclature. Data is publicly available.

ISPU simplifies access to spatial data and planning regulations, promotes interpretative consistency, and enhances legal certainty for all stakeholders.

Further system improvements are enabled by technological advances, better registry management, and widespread digital literacy. Planned upgrades include integration with eCitizens to promote public participation in planning, aligning with European practices of web-based public GIS participation (Kingston et al., 2000). The new eRegimes module will streamline planning and provide instant spatial information.

Future development includes 3D data (digital twins), artificial intelligence support, and continued user-driven module enhancements. Digital twins are becoming essential tools for dynamic urban system modeling and analysis (Batty, 2018).

REFERENCES

Batty, M. (2018). Digital twins. Environment and Planning B: *Urban Analytics and City Science, 45(5),* 817–820. https://doi.org/10.1177/2399808318796416

Faludi, A. (2000). *The Performance of Spatial Planning. Planning Practice and Research*, 15(4), 299–318. https://doi.org/10.1080/713691907

Janković, M., & Tomić, H. (2019). *Digitalna transformacija u sustavu prostornog uređenja Republike Hrvatske*. Geodetski list, 73(2), 109–124. https://hrcak.srce.hr/225830

Kingston, R., Carver, S., Evans, A., & Turton, I. (2000). *Web-based public participation geographical information systems: an aid to local environmental decision-making. Computers, Environment and Urban Systems,* 24(2), 109–125. https://doi.org/10.1016/S0198-9715(99)00049-6

Narodne novine, (2013): *Zakon o prostornom uređenju,* Narodne novine broj 153/13, Zagreb.

Narodne novine, (2023): *Pravilnik o prostornim planovima,* Narodne novine broj 152/2023, Zagreb

Nebert, D. D., & Vretanos, P. A. (2007). *Implementing the European INSPIRE Directive: the Spatial Data Infrastructure Approach. GSDI 10 Conference Proceedings.*

URL 1: https://mpgi.gov.hr/o-ministarstvu/djelokrug-50/prostorno-uredjenje-3335/3335

URL 2: https://www.nipp.hr/default.aspx?id=3114

URL 3: https://narodne-novine.nn.hr/clanci/sluzbeni/2004_08_110_2107.html

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

PYTHON-BASED AUTOMATION OF GEOPROCESSING IN DIGITAL TOPOGRAPHIC MAP PRODUCTION: CONTENT ALIGNMENT BETWEEN MAP SHEET BORDERS

MARKO SIMIĆ[1], MIHAJLO BOBAR[2], SAŠA BAKRAČ[3]

[1] Military Geographical Institute – "General Stevan Bošković", Belgrade, simic.marko.rs@gmail.com, 0000-0002-8762-2931
[2] Military Geographical Institute – "General Stevan Bošković", Belgrade, bobarmihajlo96@gmail.com, 0000-0001-9646-1730
[3] Military Geographical Institute – "General Stevan Bošković", Belgrade, sbakrac2017@gmail.com, 0000-0003-0211-3765

Abstract: *This paper presents a semi-automated geoprocessing workflow for the 1:25 000 digital topographic map (DTM25) production at the Military Geographic Institute (MGI). Traditional manual alignment of adjacent sheet boundaries is labor-intensive, error-prone, and demands detailed visual inspection of vector features. We developed a Python toolbox for ArcGIS Desktop that encapsulates MGI's cartographic rules within object-oriented classes and leverages ArcPy to automatically process and refine sheet-boundary geometries according to user-defined tolerances. Three specialized tools support unidirectional, bidirectional, and preparatory alignment scenarios, and automatically flag any residual misalignments for operator review. Deployment as a native ArcGIS toolbox lowers the barrier for non-programmers, while the modular architecture permits future enhancements such as automated quality control or adaptation to other map scales. Experimental application across multiple DTM25 sheet pairs demonstrated substantial reductions in manual effort, improved topological consistency, and enhanced geometric precision. This approach offers a replicable model for GIS organizations seeking to modernize and scale their cartographic workflows.*

Keywords: *GIS, Python, Computer science, Geoprocessing Automation, Digital Topographic Maps.*

DRAFT VERSION

1. INTRODUCTION

The field of data science and spatial analysis has witnessed an rapid growth in the availability of geospatial information. Geographic Information Systems (GIS) have thus emerged as indispensable platforms across disciplines, facilitating the visualization, interpretation, and analysis of complex spatial phenomena. As both the volume and complexity of geospatial data continue to rise, traditional manual cartographic workflows struggle to keep pace. Python has distinguished itself as a versatile, open-source programming language well suited to GIS applications, thanks to its extensive ecosystem of libraries and tight integration with leading GIS software such as QGIS and ArcGIS. These libraries not only provide powerful geoprocessing and data-manipulation capabilities but also benefit from active communities that accelerate development and troubleshooting.

This paper explores a Python-based approach to automating significant cartographic processes in the production of digital topographic maps at 1:25,000 scale (DTM25). In particular, it focuses on the challenge of seamless content alignment across adjacent map sheets, a task traditionally performed manually and thus prone to geometric inconsistencies and topological errors. By leveraging object-oriented Python scripts and ArcPy geoprocessing tools, we demonstrate how partial automation can substantially reduce processing times, enhance geometric precision, and minimize operator fatigue and error.

2. APPLICATION OF PYTHON IN GEOGRAPHIC INFORMATION SYSTEMS

Python has emerged as one of the most widely adopted programming languages for geospatial analysis, owing to its versatility, clear syntax and a robust ecosystem of libraries. Its open-source philosophy fosters collaboration, enabling developers and researchers to continuously enhance tools and techniques for spatial data processing (Gowrishankaru Veena, 2018). The readability and ease of use that Python offers lower the barrier to entry, making it accessible to users with varying levels of programming expertise. Within GIS, Python is employed across a diverse array of domains—environmental monitoring, urban planning, disaster management, precision agriculture and more—wherever spatial data must be ingested, manipulated or

visualized. Core packages such as ArcPy (for ArcGIS), GDAL/OGR, Shapely, GeoPandas and PyProj provide high-level interfaces to geoprocessing algorithms, coordinate transformations and vector/raster operations, while active user communities contribute plugins and extend functionality (Diener, 2015). Moreover, Python's seamless integration with leading GIS platforms—ArcGIS through the ArcPy site-package and QGIS via the PyQGIS API—enables the automation of repetitive tasks and the construction of custom workflows tailored to project requirements. By scripting complex chains of geoprocessing tools, users can dramatically reduce manual effort, ensure consistency across large datasets and facilitate reproducible analyses (Garrard, 2016).

3. APPLICATION OF PYTHON WITHIN ARCGIS

ArcGIS provides three primary mechanisms for executing Python code (Zandbergen, 2020):

- Stand-alone Python scripts: Developers author .py files in any IDE (Integrated Development Environment) or directly within the Conda-managed Python console bundled with ArcGIS. This approach offers full access to the Python ecosystem (third-party packages, custom modules, and version control) but requires manual handling of parameters and execution.
- Python Script Tools: By registering a stand-alone script as a geoprocessing tool, users can expose it alongside native ArcGIS tools in a toolbox. A tool dialog box allows authors to define script metadata (name, description), input/output parameters (with drop-down lists, checkboxes, and validation), and the code itself. This method simplifies deployment for non-programmers and reduces input errors, as parameter constraints are enforced by the dialog interface, yet still requires separate dialog configuration for each script.
- Python Toolboxes (.pyt): Python Toolboxes group one or more tools into a single .pyt file, entirely defined in Python without a separate dialog designer. Each tool is implemented as a class with methods such as getParameterInfo() (for defining and validating parameters) and execute() (for running geoprocessing logic). From the end-user perspective, Python toolbox tools appear and behave identically to native ArcGIS tools, enabling seamless integration into models and workflows.

Each method leverages the ArcPy library to automate geoprocessing tasks, spatial analysis, data conversion, and dataset management. Regardless of the chosen method, users benefit from ArcPy's comprehensive documentation and tightly controlled licensing model, which ensures that scripts run under the same interpreter and license checks as the core ArcGIS application (Pimpler, 2021). Together, these options empower GIS professionals to embed Python-driven workflows directly in ArcGIS apllications, from simple batch processing scripts to fully featured custom toolboxes, streamlining repetitive tasks and enforcing consistency across projects.

4. AUTOMATIC CONTENT ALIGNMENT BETWEEN MAP SHEET BORDERS

The production of the 1 : 25,000 scale digital topographic map (DTM25) at the Military Geographic Institute "General Stevan Bošković" (MGI) follows a sequence of six major phases: editorial preparation; aerial photogrammetric survey; photogrammetric mapping; field verification and completion; cartographic processing; and reproduction. During the cartographic processing phase, after all content has passed editorial review, the focus shifts to aligning each sheet's features with those of its neighbors. Three alignment scenarios can occur:

- Neighbor already finalized: content is adjusted to match the final positions of features on the adjacent sheet (unidirectional alignment).
- Neighbor under concurrent processing: both sheets undergo mutual adjustments so that their features match in tandem (bidirectional alignment).
- Neighbor not yet started: preliminary content is prepared so that, once the adjacent sheet enters cartographic processing, its data can be seamlessly integrated (preparatory alignment).

For a flawless result, all vector features (points, lines, and polygons) along sheet edges must form a continuous, gap-free, non-overlapping, and offset-free topology. Even sub-meter mismatches (e.g. 0.5–2 m) can introduce serious errors: small offsets that are invisible at DTM25 scale become magnified in larger-scale maps or advanced GIS analyses. For instance, a tiny gap in a road network breaks network connectivity (shortest-path, flow routing), overlaps in land-use polygons distort area calculations, and misaligned contour lines generate "steps" or "holes" in digital terrain models, undermining hydrologic modeling and line-of-sight analyses.

Traditionally, alignment has been performed manually. Operators painstakingly inspect and correct every line, polygon, and point along the sheet boundary, selecting, nudging, and snapping each feature by eye. This

DRAFT VERSION

process is not only time-consuming, but also monotonous, increasing operator fatigue and the likelihood of overlooked errors. Detecting miniature offsets demands zooming to very large scales (1:500 or greater), further slowing the workflow and exacerbating human error (Figure 1).

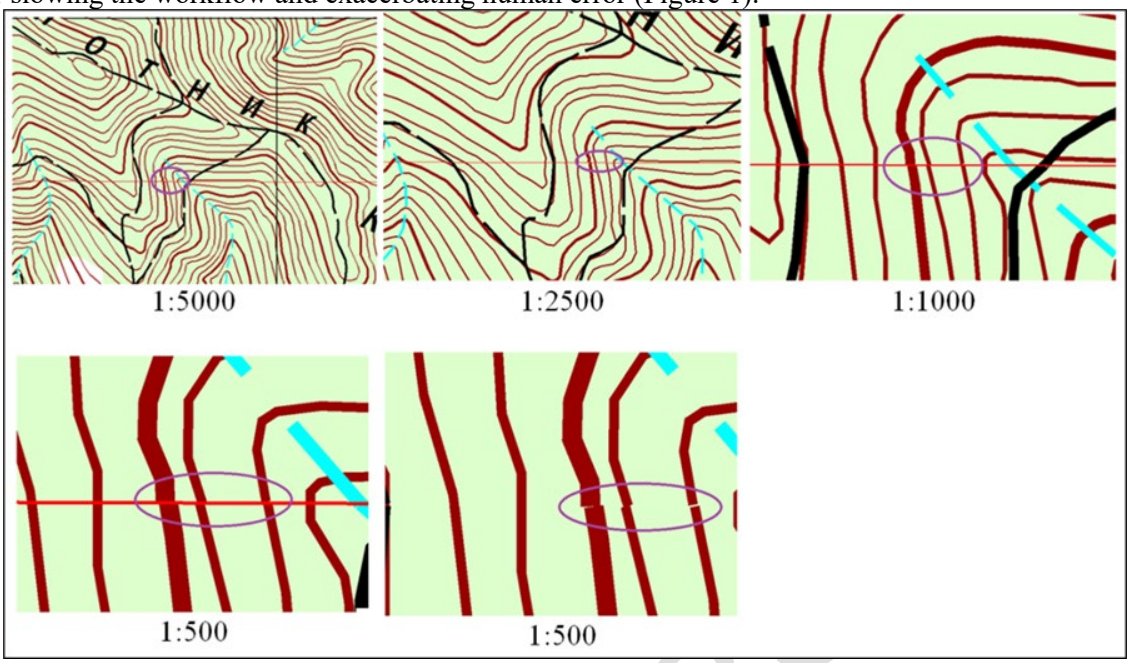


Figure 1: Subtle, hard-to-detect misalignments in line features along the boundary between two DTM25 sheets, illustrated at multiple scales

To address these challenges, we developed a semi-automated geoprocessing workflow implemented as a Python toolbox in ArcGIS Desktop. The system mimics the operator's actions: it identifies boundary features, checks their alignment, and based on user-defined tolerances automatically snaps or splits them as needed. Built on object-oriented Python scripts leveraging ArcPy tools and MGI cartographic rules, this solution reduces manual effort, accelerates processing, and ensures topological correctness and geometric consistency across all DTM25 sheets.

4.1. Python Toolbox for Semi-Automated Content Alignment

To accelerate and standardize the alignment of adjacent DTM25 map sheets, we developed a custom Python toolbox (.pyt) for ArcGIS Desktop 10.3.1. This single toolbox file bundles three distinct tools, each corresponding to one of the three alignment scenarios, into the familiar ArcGIS tool interface, so that operators with no prior programming experience can simply set parameters and run the process:

- Two Active Replicas: for bidirectional, concurrent alignment of two sheets in process.
- Active and Passive Replica: for unidirectional alignment against a finalized neighbor.
- Active Replica Only: for preparatory alignment when the neighbor sheet is not yet digitized.

Under the hood, each tool is implemented as an object-oriented Python class with getParameterInfo() defining the user's inputs (replica paths, sheet pairs, feature classes, tolerances, etc.), and an execute() method that sequences ArcPy geoprocessing operations (intersect, snap, merge, densify, etc.) according to MGI's cartographic rules.

The workflow remains semi-automated: before running the tool, the operator reviews and marks which feature classes need processing to optimize performance. After execution, any misalignments outside the chosen tolerances are flagged by automatically generated point features, guiding the operator to residual errors. In the "Active Replica Only" tool, a secondary point layer highlights line features within the user-defined snap distance that remain unconnected. This design delivers a balance of automation and human oversight for the few edge cases where manual intervention remains necessary. The dialog windows for launching both the "Active & Passive Replica" and "Active Replica Only" tools are shown in Figure 2.

DRAFT VERSION

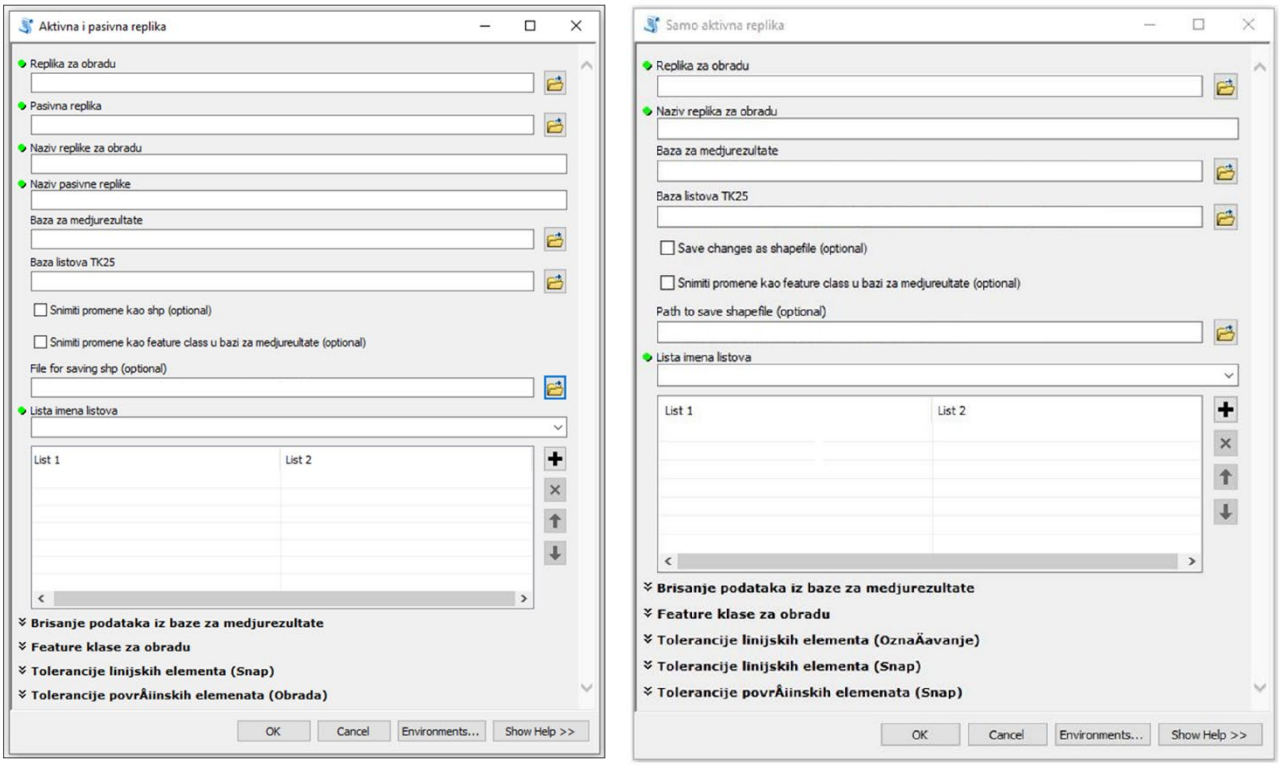


Figure 2: User interface for launching the "Active & Passive Replica" tool (left) and the "Active Replica Only" tool (right) in the ArcMap application

4.2. Script Architecture

The toolbox is built on the ArcPy library and follows an object-oriented programming (OOP) paradigm. In OOP, classes serve as "blueprints" for objects, encapsulating attributes (parameters and properties) and methods (functions that operate on those attributes). This organization groups related data and logic into cohesive units, facilitates inheritance of shared behavior, and promotes code reuse, resulting in shorter, more flexible scripts (Kindler & Krivy, 2011).

Each tool in the Python toolbox is implemented as a distinct class, recognized by ArcGIS via its .pyt structure, with two core methods:

- getParameterInfo(): defines and configures all user-input parameters in the dialog (e.g., replica paths, sheet pairs, feature classes, tolerance distances).
- execute(): contains the geoprocessing workflow, invoking ArcPy operations in sequence.

At the beginning of the script, two processing classes are defined: Linijski_Objekti and Poligonski_Objekti. Each class encapsulates a suite of ArcPy-based methods for handling its respective feature type. To ensure robustness, every method wraps its core logic in a try–except block, catching and logging errors without halting the entire process.

Within the execute() method, the workflow proceeds through the following stages:

- Parameter initialization: user inputs are assigned to internal variables, and paths to all relevant geodatabases (active/passive replicas, intermediate results, sheet index) are configured.
- Generating sheet-pair polygons: based on the DTM25 map sheet database, frame polygons representing sheet boundaries are created. A predefined list of sheet name pairs (each stored as a tuple) defines which adjacent sheets require alignment. A for loop iterates through these tuples, selects the corresponding boundary polygons for each sheet name, and appends each polygon pair to a new processing list. This structure enables simultaneous batch processing of multiple sheet pairs.
- Feature-object instantiation: Based on the defined classes, the script creates instances for each thematic layer. Line objects (contours, flowing water, roads, railways, public-utility features, vegetation, etc.) and polygon objects (vegetation patches, water bodies, marshlands, hydrologic features) are appended to two separate lists for line- and polygon-level processing.
- Two-stage hierarchical processing: A first loop iterates over each sheet pair, executing the full processing sequence independently for each pair. Within that loop, nested iterations traverse the list of line-feature classes and then the list of polygon-feature classes, calling the corresponding class methods for each element (Figure 3).

DRAFT VERSION

```
1863        for feature_class in povrsinski_elementi_za_obradu_1:
1864            arcpy.AddMessage("Obradjuje se {}".format(feature_class.ime))
1865            fc_za_selektovanje = feature_class.fc_za_menjanje
1866            for attribute_value in feature_class.attribute_value:
1867                layer_1 = arcpy.MakeFeatureLayer_management(fc_za_selektovanje, "replika_1_layer")
1868                layer_2 = arcpy.MakeFeatureLayer_management(feature_class.fc_pasivna, "replika_2_layer")
1869                feature_class.select_features(attribute_value, selection_type="NEW_SELECTION",
1870                                              select_aktivna=layer_1, select_pasivna=layer_2)
1871                intersection_line_parallel = feature_class.create_parallel_line_to_frame(TK25_1, TK25_2,
1872                                                                                         feature_class.paralelna_linija_udaljenost)
1873                intersection_line_fc_pasivna = feature_class.create_intersection_fc_pasivna_tk25_aktivna(TK25_1,
1874                                                                                                         pasivna_layer=layer_2)
1875                feature_class.select_fc_on_frame_border(intersection_line_parallel, select_aktivna=layer_1)
1876                feature_class.snap_fc_on_parallel_line(intersection_line_parallel,
1877                                                       feature_class.snap_paralelna_distance, snap_to=layer_1)
1878                feature_class.densify_fc(feature_class.densified_distance, densify_this=layer_1)
1879                feature_class.snap_fc_on_intersection_line(intersection_line_fc_pasivna,
1880                                                           feature_class.snap_za_okvir_end,
1881                                                           feature_class.snap_za_okvir_edge, snap_this=layer_1)
1882                feature_class.delete_feature_layers(delete_aktivna=layer_1, delete_pasivna=layer_2)
1883                ##########################################################################################
1884                feature_class.delete_intermediate_data()
1885
1886        for feature_class in povrsinski_elementi_za_obradu_2:
1887            arcpy.AddMessage("Obradjuje se {}".format(feature_class.ime))
1888            fc_za_selektovanje = feature_class.fc_za_menjanje
1889            layer_1 = arcpy.MakeFeatureLayer_management(fc_za_selektovanje, "replika_1_layer")
1890            layer_2 = arcpy.MakeFeatureLayer_management(feature_class.fc_pasivna, "replika_2_layer")
1891
1892            intersection_line_parallel = feature_class.create_parallel_line_to_frame(TK25_1, TK25_2,
1893                                                                                     feature_class.paralelna_linija_udaljenost)
1894            intersection_line_fc_pasivna = feature_class.create_intersection_fc_pasivna_tk25_aktivna(TK25_1,
1895                                                                                                     pasivna_layer=layer_2)
1896            feature_class.select_fc_on_frame_border(intersection_line_parallel, select_aktivna=layer_1)
1897            feature_class.snap_fc_on_parallel_line(intersection_line_parallel, feature_class.snap_paralelna_distance,
1898                                                   snap_to=layer_1)
1899            feature_class.densify_fc(feature_class.densified_distance, densify_this=layer_1)
1900            feature_class.snap_fc_on_intersection_line(intersection_line_fc_pasivna, feature_class.snap_za_okvir_end,
1901                                                       feature_class.snap_za_okvir_edge, snap_this=layer_1)
```

Figure 3: Python code snippet for processing polygon features within the "Active & Passive Replica" tool

This layered, class-based architecture delivers a flexible, scalable framework for semi-automated alignment of vector features across an arbitrary number of sheet pairs. Detailed flow diagrams for each tool's logic are provided in Figures 4–6.

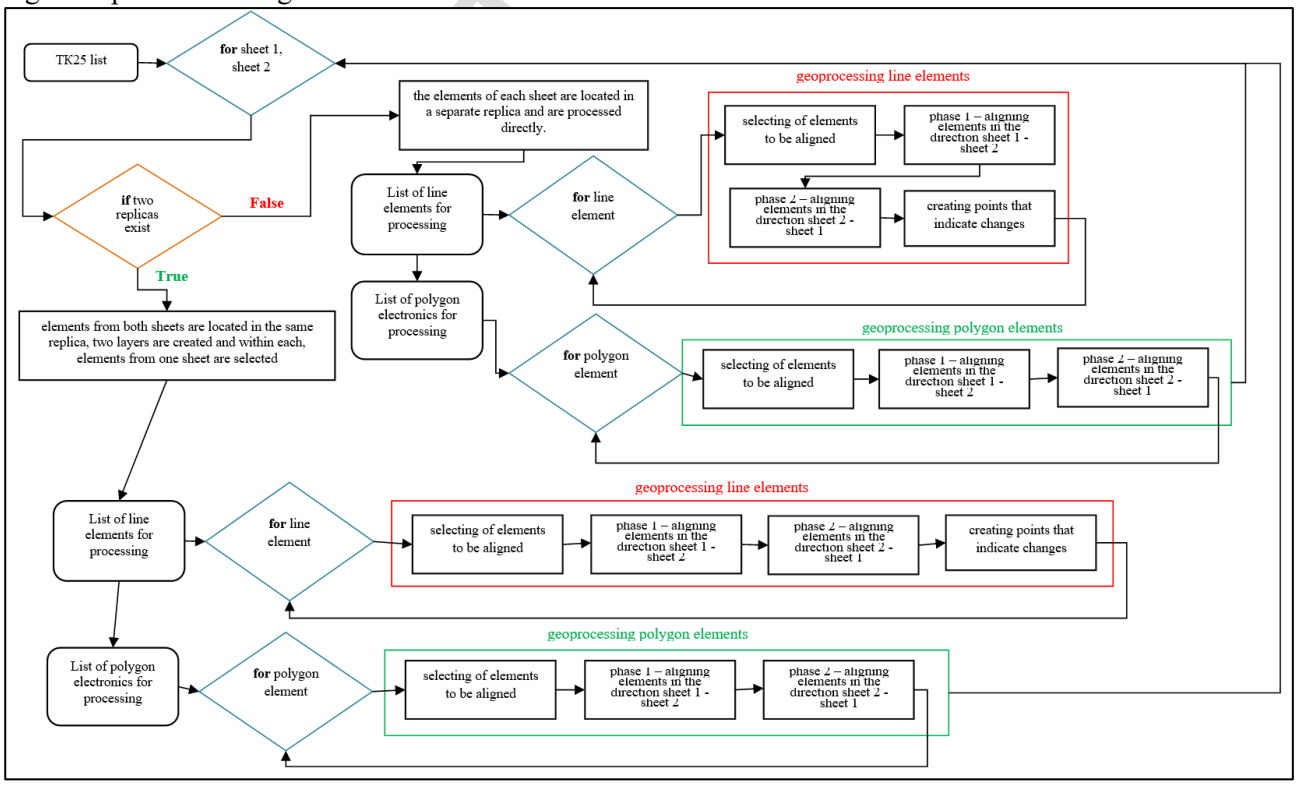


Figure 4: Diagram illustrating the content alignment workflow in the Two Active Replicas tool

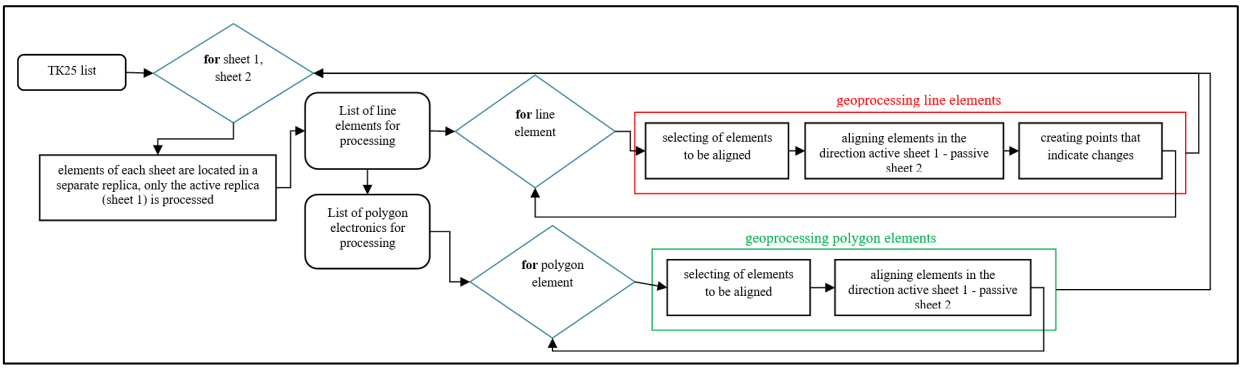


Figure 5: Diagram illustrating the content alignment workflow in the Active & Passive Replica tool

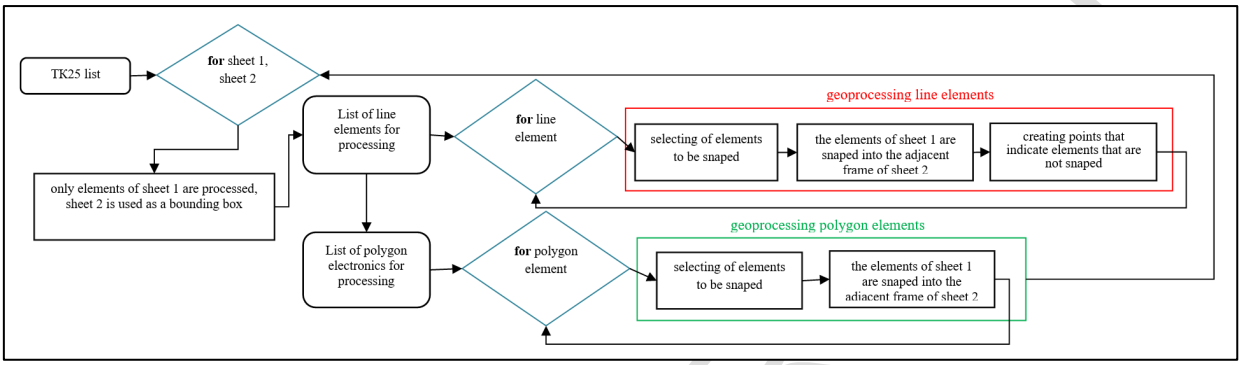


Figure 6: Diagram illustrating the content alignment workflow in the Active Replica Only tool

5. CONCLUSION

This work demonstrates that integrating Python-driven geoprocessing into the DTM25 cartographic workflow delivers substantial improvements in both efficiency and quality. By encapsulating MGI's cartographic rules within object-oriented Python classes and utilizing ArcPy geoprocessing tools, the semi-automated toolbox drastically reduces manual effort and processing times while enforcing topological integrity and geometric consistency. Automatic detection and correction of sub-meter offsets both speeds up the alignment task and minimizes operator fatigue and human errors.

Packaging the solution as a native ArcGIS toolbox ensures accessibility for GIS professionals without programming backgrounds, promoting rapid adoption and standardization within the institute's production chain. The modular, class-based design also paves the way for future enhancements, such as automated quality control, integration of additional data sources, or adaptation to other map scales, bringing us closer to fully automated topographic map generation.

Leveraging Python's versatile, open-source ecosystem and active community support, this approach offers a reproducible model for any GIS organization seeking to modernize manual cartographic tasks. As the volume and complexity of geospatial data continue to grow, Python-based automation will become increasingly vital for producing timely, accurate, and cost-effective cartographic products.

LITERATURE

Diener, M. (2015). Python Geospatial Analysis Cookbook. Birmingham: Packt Publishing Ltd. ISBN: 9781783555079

Garrard, C. (2016). *Geoprocessing with Python*. New York: Manning Publications Co. ISBN: 9781617292149

Gowrishankar, S., and Veena, A. (2018). *Introduction to Python Programming*. Miami: CRC Press. ISBN: 9780815394372

Kindler, E. and Krivy, I. (2011). *Object-Oriented Simulation of systems with sophisticated control*. International Journal of General Systems, 40/3, pp. 313–343.

Pimpler, E. (2021). *Programming ArcGIS Pro with Python - 2ndEdition*. Boerne: Geospatial Training Services. ISBN: 9798740349961

Zandbergen, A., P. (2020). *Advanced Python Scripting for ArcGIS Pro*. San Bernardino: Esri Press. ISBN: 9781589486188

DRAFT VER

SYM-OP-IS 2025
52. Simpozijum o operacionim istraživanjima
Palić, 7 - 10.9.2025.

UNIVERZITET U BEOGRADU
FAKULTET ORGANIZACIONIH NAUKA

Review article

REMOTE MONITORING OF WATER SURFACE POLLUTION

MARKO STOJANOVIĆ[1,2], MIRJANA KOKOTOVIĆ[3], SLAĐANA STANIŠIĆ[4], MIODRAG REGODIĆ[4]

[1] Military Geographical Institute, Belgrade, ; stojanovicm80@yahoo.com, 0000-0002-2193-1483
[2] University of Defence, Military Academy, Belgrade; stojanovicm80@yahoo.com, 0000-0002-2193-1483
[3] Georad d.o.o. Trebinje, mirjana.kokotovic@student.aggf.unibl.org, N/A
[4] University of Banja Luka, Faculty of Architecture, Civil Engineering and Geodesy, Banja Luka; sladjana.stanisic@aggf.unibl.org, miodrag.regodic@aggf.unibl.org N/A, 0000-0003-4675-4150

Abstract: *Remote sensing is a method that enables efficient and cost-effective monitoring of large and inaccessible areas. Remote sensing plays a significant role in detecting and analyzing pollutants present on the surface of water. Special emphasis in this paper is placed on the satellite missions SMOS, Aquarius/SAC-D, and GRACE, which were developed to monitor ocean salinity, soil moisture, changes in ocean circulation, and shifts in Earth's mass distribution. The paper also presents specific examples of water pollution detected through remote sensing. Examples include the environmental crisis in the Gulf of Mexico caused by the Deepwater Horizon oil rig explosion and an oil spill in a river near Norilsk, Russia. Satellite imagery allowed for precise tracking of the extent and impact of these events on the environment. The paper also demonstrates how digital images can be further enhanced using software tools, contributing to better interpretation and visualization of data.*

Keywords: *Remote sensing, SMOS, Aquarius/SAC-D, GRACE.*

DRAFT VERSION

1. INTRODUCTION

In the today's era of climate change and growing human influence on natural resources, monitoring the state of water bodies has become a key task in environmental protection (Rice, 2023). Remote sensing, as a method of collecting data without direct contact with the observed objects, enables continuous and comprehensive monitoring of large water surfaces, including seas, oceans, lakes, and rivers. Through the application of satellite missions including SMOS (Soil Moisture and Ocean Salinity), Aquarius/SAC-D, and GRACE (Gravity Recovery and Climate Experiment), data are collected on parameters like salinity, soil moisture, gravitational variability, and other indicators that enable the analysis of hydrological processes and potential sources of pollution (Dione & Faye, 2024). The purpose of this paper is to highlight the significance, capabilities, and applications of remote sensing in both ecological and scientific contexts, with a particular focus on specific examples and technologies used for monitoring the quality and changes of water resources around the world.

2. REMOTE SENSING OF THE STATE OF WATER BODIES

At the beginning of 1998, parallel efforts in Europe and the United States were directed toward exploring the concept of satellite sensors and mission design for measuring salinity as either a secondary or primary objective. Several campaigns were initiated to obtain new data and evaluate the models. Two research missions were approved: the SMOS and the Aquarius/SAC-D (Vernières et al., 2014).

2.1. SMOS mission

The research mission of the European Space Agency (ESA) is dedicated to enabling global observation of land surface soil moisture and ocean salinity (Weng, 2011). The SMOS satellite helps to clarify how the Earth's surface and atmosphere interact, supporting improvements in environmental prediction models (Bajiou Mroczkowska, 2018). The SMOS satellite was launched on November 2, 2009, and although it was

originally designed as a five-year mission, it has been extended, reflecting its versatility and new synergistic capabilities (European Space Agency, n.d.). The appearance of the SMOS satellite in orbit is shown in Figure 1.

Figure 1: Satellite in orbit (European Space Agency, n.d.)

The SMOS satellite is equipped innovative interferometric radiometer that operates within the L-band microwave spectrum, allowing for accurate temperature observations. This data is used to produce a global soil moisture map every three days. (United Nations Office for Outer Space Affairs, n.d.).

SMOS enables the monitoring of changes related to the amount of water present in the surface layers of soil and the concentration of salt in the upper layer of the sea. Figure 2 shows the average distribution of salinity in seas and oceans. Areas shown in red indicate high salinity levels, while green represents low salinity levels. The figure also displays a simplified global pattern of ocean circulation known as thermohaline circulation. Blue arrows are used to show cold deep currents, and red arrows depict warm surface currents (Sheldon et al., 2016).

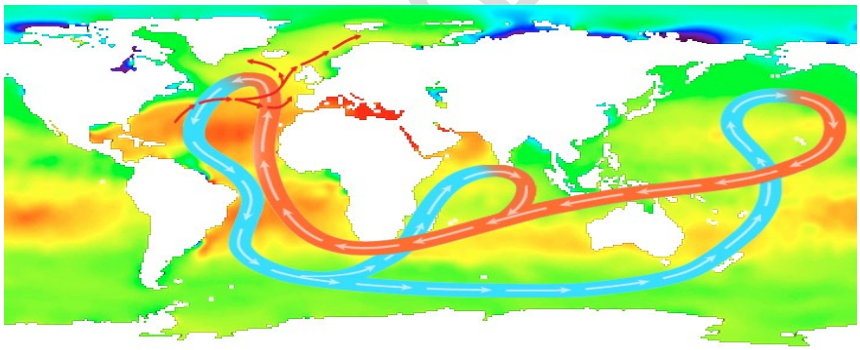


Figure 2: Sea and Ocean Salinity and Global Circulation (European Space Agency, n.d.)

2.2. AQUARIUS/SAC-D mission

NASA and CONAE partnered to develop the Aquarius/SAC-D mission. Figure 3 illustrates the design of the Aquarius/SAC-D spacecraft (Le Vine, Jacob, Dinnat, de Matthaeis, & Abraham, 2007).

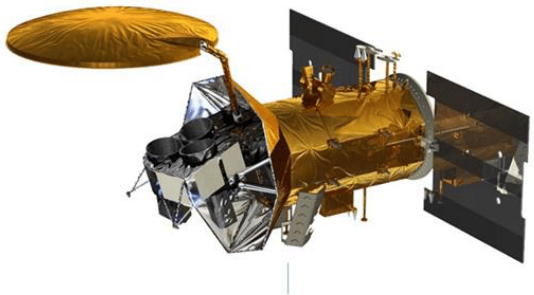


Figure 3: Aquarius/SAC-D spacecraft (Le Vine, Dinnat, Meissner, & Wentz, 2018)

DRAFT VERSION

The mission was launched on June 10, 2011, and concluded on June 8, 2015. SAC-D/Aquarius is a multi-sensor mission that covers the ocean, land, atmosphere, and space. Its main aim is to enhance to the understanding of the Earth's system (Sen, Caruso, & Lagerloef, 2008). The primary scientific objective is to examine the mechanisms that lead to changes in the water cycle and ocean circulation, as well as their impact on climate, by measuring global variations in sea surface salinity. Additional objectives include: observing environmental changes, natural hazards, and sea ice (Lagerloef & Torrusio, 2008).

Through measurements of ocean salinity variations, Aquarius/SAC-D contributed valuable data. Salinity, along with temperature, determines how dense and buoyant seawater is. This, in turn, provides information about the ocean water layers and how they mix.

2.3. GRACE mission

GRACE is a geodetic mission of the United States and Germany. Its overall goal is to obtain long-term, precise data on on both the average levels and the changing elements of Earth's gravitational field (European Space Agency, n.d.). The appearance of the GRACE satellite is shown in Figure 4.

Some of the scientific objectives are: to study ocean currents and heat flow, monitor seafloor pressure and track changes in ice sheets, land water storage and glaciers. (Tapley, Bettadpur, Watkins, & Reigber, 2004).

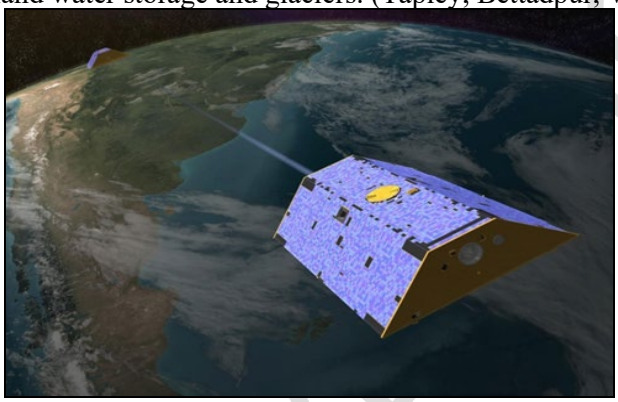


Figure 4: GRACE satellite (Zurich, 2025)

The mission concept is based on measuring variations in the distance among two coplanar satellites and their derivatives (Prost, 2025). The orbits of the two separate spacecraft vary as they pass through Earth's gravitational field, resulting in measurable variations between them. The GRACE satellites were launched into Earth orbit on March 17, 2002, and operated for three times longer than originally planned - until the launch of the new GRACE-2 satellites on October 27, 2017 (NASA Jet Propulsion Laboratory, n.d.).

3. EXAMPLES OF REMOTE SENSING OF WATER SURFACE POLLUTION

The use of satellite imagery is possible in all phases of environmental research, which requires the application of multiple methodological approaches, including both remote sensing and terrestrial investigations (Regodić, Sekulović, & Živković, 2010).

3.1. Disaster in the Gulf of Mexico

On April 20, 2010, a major environmental disaster occurred at the Deepwater Horizon oil rig in the Gulf of Mexico (Figure 5). The Deepwater Horizon oil spill, one of the largest marine oil disasters in history, polluted more than a thousand kilometers of Louisiana's marshy coastline, posing a severe threat to the local ecosystem (Bukharitsin, 2022).

After the explosion on the platform, crude oil began leaking into the sea from its well, located 1,500 meters below the ocean surface. The greatest victims of the disaster were the 11 workers who lost their lives, fisheries, tourism, marine flora and fauna, as well as various bird species that were severely affected. (Bukharitsin, 2022).

DRAFT VERSION

Figure 5: Oil spill caused by an offshore platform accident (NASA Earth Observatory, 2010)

Remote sensing provides key data for analyzing oil spill impacts on large, complex coastlines.

3.2. The Norilsk Disaster

On May 29, 2020, a major oil spill incident happened close to the city of Norilsk in Russia. (Sahu & Dutta, 2021). The spill happened because a fuel tank broke down, which caused a large area of water and soil to become polluted. The oil spill was a major industrial accident that realised 20,000 tons of fuel into the nearby Daldykan and Ambarnaya rivers. Figure 6 shows the area polluted by the oil spill from the fuel reservoir (Sahu & Dutta, 2021).

Figure 6: Satellite image of the area polluted by the oil spill from the fuel reservoir (NASA Earth Observatory, 2020)

4. IMPROVING THE QUALITY OF DIGITAL IMAGES

The significance of an image is considerably more complex and has a broader application than a photograph. Images provide useful information about changes in land and objects on it. In the satellite image of the area affected by the oil spill (Figure 6), spatial, radiometric, and spectral enhancement was performed using the MATLAB software package.

4.1. Spatial Improvement of Digital Image Quality

Spatial improvement of image quality refers to improving the quality of the visual presentation and emphasizing the desired information. As part of the spatial improvement of digital image quality in the MATLAB program, image filtering and edge enhancement of objects were performed.

Figures 7 and 8 show images corrupted by Gaussian, Poisson, salt-and-pepper, and multispectral noise, each affecting 10% of the image area.

DRAFT VERSION

Figure 7: Image affected by Gaussian noise (left) and Poisson noise (right). The images were processed in MATLAB.

Figure 8: Image affected by salt-and-pepper noise (left) and multispectral noise (right). The images were processed in MATLAB.

5. CONCLUSION

Remote sensing methods are used for the creation of orthophoto maps and charts derived from generated satellite images of a particular area, enabling efficient data collection along with a clear overview of the entire study area. Remote sensing can be viewed as a multidisciplinary system of "tools" that can be applied to define, analyze, and solve problems and challenges encountered in emergency situations. The ability to obtain accurate and timely information can assist in resolving field situations. By combining remote sensing methods, it is possible to create a pseudo-realistic representation of the entire situation and make strategic and tactical decisions quickly in accordance with the presented challenges.

In hazardous or hard-to-reach areas remote sensing enables data collection. The development of SMOS and other satellite Earth observation systems is of great importance for understanding the condition of seawater and oceans. By using images obtained through remote sensing, it is possible to successfully monitor and analyze water conditions, water movements, as well as track and assess water pollution from an ecological perspective. Further processing of satellite images, for example using software packages like MATLAB, allows for digital image processing. There are various digital image processing techniques, and images processed in this way are more suitable for use in specific fields. Although remote sensing is a effective tool for gathering data about the Earth's surface, it faces numerous limitations such as limited spatial and spectral resolution, atmospheric effects, complexity of data interpretation, and dependence on weather conditions and the availability of quality satellite images.

Recently, remote sensing has been rapidly advancing and developing, both in its methods and the technology used, so even greater applications can be expected in the future.

REFERENCES

Bajiou Mroczkowska, D. (2018). *Development of a system for detection, control and prevention of locust pests using UAV platforms* (Bachelor's thesis, Escola d'Enginyeria de Telecomunicació i Aeroespacial de Castelldefels, Universitat Politècnica de Catalunya). UPCommons. http://hdl.handle.net/2117/114519

DRAFT VERSION

Bukharitsin, P. (2022). *Studies of ice in the Caspian, Azov, and Aral Seas*. https://www.rgo.ru/sites/default/files/p._b_u_k_h_a_r_i_t_s_i_n_monografiya_studies_ice_caspian_azov_end_aral_sea.pdf

Dione, P. M., & Faye, C. (2024). GRACE characterization of spatiotemporal variation in terrestrial water storage in the AGA-FOUA-DJILAS watershed (Senegal). Environmental Geosciences and Sustainable Management, 2(1), 25–46. https://doi.org/10.34874/IMIST.PRSM/EGSM/53026

European Space Agency. (n.d.). *GRACE (Gravity Recovery and Climate Experiment)*. eoPortal. https://www.eoportal.org/satellite-missions/grace

European Space Agency, (2025). *Introducing SMOS*. Retrieved from https://www.esa.int/Applications/Observing_the_Earth/FutureEO/SMOS/Introducing_SMOS

Lagerloef, G. S., & Torrusio, S. E. (2008). The Aquarius/SAC-D mission: A new era in ocean salinity remote sensing. *Oceanography, 21*(1), 68-81. doi:https://doi.org/10.5670/oceanog.2008.68

Le Vine, D. M., Dinnat, E. P., Meissner, T., & Wentz, F. J. (2018). Status of Aquarius and Salinity Continuity. *Remote Sensing, 10*(10), 1585. doi:https://doi.org/10.3390/rs10101585

Le Vine, D. M., Jacob, S. D., Dinnat, E. P., de Matthaeis, P., & Abraham, S. (2007). The influence of antenna pattern on Faraday rotation in remote sensing at L-band. *IEEE Transactions on Geoscience and Remote Sensing*, 45(6), 1750–1759. https://doi.org/10.1109/TGRS.2007.895823

NASA Earth Observatory, (2010). *Oil slick in the Gulf of Mexico*. Retrieved from https://earthobservatory.nasa.gov/images/43877/oil-slick-in-the-gulf-of-mexico

NASA Earth Observatory, (2020). *Diesel spill in Ambarnaya River, Russia*. Retrieved from https://earthobservatory.nasa.gov/

NASA Jet Propulsion Laboratory, (2025). *GRACE-FO Mission Overview*. Retrieved from https://gracefo.jpl.nasa.gov/mission/overview/

Prost, G. L. (2025). *Remote sensing for geoscientists: Image analysis and integration* (4th ed.). CRC Press/Taylor & Francis.

https://www.taylorfrancis.com/books/mono/10.1201/9781003617310/remote-sensing-geoscientists-gary-prost

Regodić, M., Sekulović, D., & Živković, L. (2010). Application of Remote Sensing in the Study of Meteorological Parameters. *ENERGY, ECONOMY, ECOLOGY*, 323–328.

Rice, J. (2023). Monitoring Arctic pollution using satellite remote sensing. Uppsala University, Institute for Russian and Eurasian Studies. https://uu.diva-portal.org/smash/get/diva2:1795285/FULLTEXT01.pdf

Sahu, S. S., & Dutta, D. (2021). Monitoring oil spill in Norilsk, Russia using satellite data. *Scientific Reports, 11*, 1–14. doi:https://doi.org/10.1038/s41598-021-83260-7

Sen, A., Caruso, D., & Lagerloef, G. S. (2008). Aquarius/SAC-D mission. *Proceedings of SPIE*, 7106, 710605. https://doi.org/10.1117/12.804033

Sheldon, C. M., Seidenkrantz, M.-S., Pearce, C., Kuijpers, A., Hansen, M. J., & Christensen, E. Z. (2016). Holocene oceanographic changes in SW Labrador Sea, off Newfoundland. *The Holocene*, *26*(2), 274–289. https://doi.org/10.1177/0959683615608690

Tapley, B. D., Bettadpur, S., Watkins, M., & Reigber, C. (2004). The Gravity Recovery and Climate Experiment: Mission overview and early results. *Geophysical Research Letters, 31*(9), L09607. doi:https://doi.org/10.1029/2004GL019920

United Nations Office for Outer Space Affairs. (n.d.). *Hydrological*. In UN-SPIDER Knowledge Portal. Retrieved June 23, 2025, from https://un-spider.org/es/category/disaster-type/hidrol%C3%B3gicas

Vernières, G., Kovach, R., Keppenne, C., Akella, S., Brucker, L., & Dinnat, E. (2014). *The impact of the assimilation of Aquarius sea surface salinity data in the GEOS ocean data assimilation system. Journal of Geophysical Research: Oceans*, *119*, 6974–6987. https://doi.org/10.1002/2014JC010006

Weng, Q. (Ed.). (2011). *Advances in Environmental Remote Sensing: Sensors, Algorithms, and Applications* (Taylor & Francis Series in Remote Sensing Applications). CRC Press.

Zurich, E. (2025). *GRACE Level-1 ML*. Retrieved from https://space.igp.ethz.ch/research/grace-lv1-ml.html

DRAFT VERSION

GRAPHS AND NETWORKS

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

VORONOI DIAGRAM AS A TOOL FOR ANALYSING DEVELOPMENT IN RAILWAY AND AIR TRANSPORT INFRASTRUCTURE: A EUROPEAN PERSPECTIVE

MARTIN FALE[1], TEA VIZINGER[2], BOJAN RUPNIK[3], TOMAŽ KRAMBERGER[4]

[1] University of Maribor, Faculty of Logistics, Celje, Slovenia, martin.fale1@um.si 0009-0001-6014-1775
[2] University of Maribor, Faculty of Logistics, Celje, Slovenia, tea.vizinger@um.si 0000-0003-1539-9600
[3] University of Maribor, Faculty of Logistics, Celje, Slovenia, bojan.rupnik@um.si 0000-0001-9983-1178
[4] University of Maribor, Faculty of Logistics, Celje, Slovenia, tomaz.kramberger@um.si 0000-0001-7656-5444

Abstract: *This research examines the use of Voronoi diagram as an effective tool for analyzing the spatial dispersion and accessibility of mobility hubs in Europe, with a particular focus on railway and air infrastructure. Using ArcGIS Pro tool, we generated Voronoi sets representing selected railway stations and airports within the study area. Based on the size of the Voronoi sets, we assessed the development and coverage of each mode of transport and identified areas with less developed railway infrastructure. Such areas are in Northern, Eastern, and Southeastern Europe. Central Europe, Italy, and the Iberian Peninsula are areas where railway infrastructure is comparable to air infrastructure. The average Voronoi set for railway stations is larger due to their lower number, while airports have smaller average Voronoi set because they are more numerous and distributed more densely. The results show that Voronoi diagram provide clear visualization of accessibility to mobility hubs and are useful for spatial planning and as a decision support tool.*

Keywords: *Voronoi diagram, rail transportation, air transportation, Geographic Information System.*

1. INTRODUCTION

The tendency to use rail transportation as an alternative to air travel for distances up to 1000 km in the European Union (EU) is increasing. Currently, the environmental aspect is at the forefront, as (high-speed) rail is significantly less harmful to the environment than air travel (European Environmental Agency, 2021). Promoting the use of (high-speed) rail as a competitor to air travel faces the challenge of inadequate or non-existent infrastructure for high-speed trains. To fully benefit from rail transportation at the EU level, adequate rail infrastructure must be ensured. Everyone must have access to a railway station where a train stops that can take them to their destination in a travel time comparable to that of an airplane. If an individual does not have a railway station near their point of departure but does have access to an airport, the airport becomes the accessibility hub for continuing their journey toward their destination. Decision-makers must therefore identify areas where investments in specific types of transport infrastructure are necessary to provide an alternative to the current dominant mode of transport. This can be done using a Voronoi diagram.

Burkey et al. (2011) define Voronoi diagram as a simple geometric construct with a wide range of applications. A Voronoi diagram consists of Voronoi sets. Each set represents a region associated with a specific central point, and this region contains all points that are closer to the central point than to any other central point in the given space. Introducing weights to the central points leads to weighted central points, which in turn attract a larger set of points than they would otherwise. Dong (2008) notes that Voronoi sets in Geographic Information Systems (GIS) are typically based on unweighted central points. They divide the plane into distinct regions, a collection of planar sections. Marianov and Eiselt (2024), in their review of the history of location models, highlight the importance of Voronoi diagram in location theory. This is further supported by Boonprong et al. (2024), who describe the Voronoi diagram as a powerful tool for spatial analysis and decision support. In our research, we aim to explore whether Voronoi diagram can help identify areas where rail transport is at a disadvantage compared to air transport or vice versa.

2. RELATED WORK

Kloskowski et al. (2024) highlighted the relevance of using Voronoi diagrams as a tool for planning the spatial development of urban areas. Their study focused on determining optimal locations for shopping centers,

DRAFT VERSION

based on the Movement of People indicator. The Voronoi sets were generated for a selected area in Poland, covering the cities of Gdansk and Gdynia. Each Voronoi set represented the value of this indicator: smaller areas indicated higher levels of physical activity among the population, while larger areas signified lower levels of physical activity in a given region. Raghu Kisore and Koteswaraiah (2017) addressed the problem of ATM placement. They used Voronoi sets to determine the optimal number and distribution of ATMs. The sets represented potential locations, while waiting times at each ATM, affected by population density and the types of commercial activities in the area, were used as weights. Thus, the Voronoi set represented the influence zones of each ATM. The study did not specify the basis used for generating the Voronoi diagram.

Figurska et al. (2022) used Voronoi diagram to assess accessibility to selected facilities for the elderly in the city of Olsztyn, Poland. They divided the area into uniformly sized cells, Voronoi sets, and calculated the number of such facilities within each cell, focusing on those with high social significance for seniors. Jiao and Feng (2024) explored the problem of shelter location as a part of emergency planning. Their study included the use of weighted Voronoi diagram. Based on published figures, it appears they divided the selected area, Lindian County in Heilongjiang Province, into uniformly sized cells. They combined Voronoi sets with OD Cost Matrix analysis results to represent the coverage area of each shelter, considering the number of residences and a digital elevation model in the context of flood risk. Shahparvari et al. (2020) examined the distribution of fire stations in Melbourne. Based on their results, they proposed locations to establish new fire station or to relocate existing ones to achieve better coverage. Voronoi diagram was used as a tool to define the boundaries of areas that are closer to a particular fire station than to any other.

Several studies using Voronoi diagram tool have focused on the problem of locating electric vehicle (EV) charging stations. Prakobkaew and Sirisumrannukul (2022) divided selected provinces in Thailand into uniformly sized cells, to represent the number of electric vehicles in each area. They used Voronoi sets to illustrate the service areas of public EV charging stations. Jordan et al. (2022) investigated the problem of EV charging infrastructure in Valencia, Spain. They used Voronoi diagram to show how the coverage area of individual charging stations decreases as the number of stations increases. The coverage area of each charging station was defined as the intersection of a circle representing the station's influence zone and the Voronoi set. Boonprong et al. (2024) emphasized the advantage of using Voronoi diagram for spatial analysis. By analyzing coverage areas, they assessed the current distribution of EV charging stations and proposed locations for new ones. The Voronoi sets were generated using ArcGIS tool and were based on population density, proximity to roads of different categories, and land use patterns. The underlying data for generating the Voronoi sets were population data.

Mota et al. (2014) focused on determining the locations of suburban railway stations in the city of Brasília using weighted Voronoi sets, generated with ArcGIS tool. Each Voronoi set represented the relationship between a potential railway station site and the demand for rail transport. Based on their findings, the authors proposed a framework for establishing a railway corridor in the selected area.

3. METHODOLOGY

This study is part of a broader research project. The main idea is to create a model that determines, for any location within a given area, which mode of transport, airplane, train or another, is optimal within a predefined radius based on various factors such as time, cost and sustainability. Familiarity with the research context facilitated the development of the methodological framework of this research. First, we defined the concept of the Voronoi diagram and its unit, the Voronoi set. A literature review was conducted using Science Direct and Web of Science databases. Initially, we searched using the keywords Voronoi AND mobility hubs AND accessibility, but no results were found in either database. This indicate the novelty of research that uses Voronoi diagram to assess accessibility to mobility hubs. We then expanded our search with keywords Voronoi AND transportation, Voronoi AND hubs, and Voronoi AND GIS. Accordingly, we included the findings of nine studies in our review of the related work. The review of related work clearly indicates the importance of the Voronoi diagram in location theory.

The study area includes EU countries, the United Kingdom, Norway, and Switzerland. Within the study area, airports with more than 30 connections were considered, focusing on flight connections up to 1000 km in length. Railway stations were selected based on data from the Omio portal, considering additional information from the Eurail Association about railway stations currently designated as part of the existing high-speed rail network in Europe (Eurail, 2025; Flight Connections, 2025; Omio, 2025). The initial list of 299 airports and 171 railway stations was compiled in August 2023, and its accuracy was verified in May 2025.

For the spatial analysis, we used the software tool ArcGIS Pro, version 3.5.0. Locations of airports and railway stations were added using the Locate tool, creating point layers for airports and railway stations. Both

DRAFT VERSION

layers were created in the Geographic Coordinate System WGS 1984. The polygon layer of Europe from Geofabrik served as the base for selecting the study area (Geofabrik, 2025). To determine the influence zones of airports and railway stations, we created Voronoi sets using the Create Thiessen Polygons tool, one for airports and one for railway stations. To ensure that Voronoi sets were generated for the entire study area, we set the Processing Extent in Environments to Extent of a layer, selecting the polygon layer of the study area. For better readability of the Shape Area units of the Voronoi sets, we reprojected the generated Voronoi polygon layers into the projected coordinate system ETRS 1989 using the Project tool. This allowed us to obtain area data for each Voronoi set in metric units, meters. Several assumptions were also taken into consideration. The measure used to assess the development of a particular mode of transport is the area of the Voronoi sets, which represents the region served by an individual airport or railway station. The choice of the Omio platform is based on the following assumption: An individual's journeys between various locations gradually contribute to the recognition of major transport flows. This, in turn, leads to the establishment of regular connections using different modes of transport, which aim to meet the demand for travel along these routes. Omio, a platform designed to facilitate travel between frequently visited points, offers these locations for selection. Additionally, Omio can suggest alternative modes of transport, including air, rail, or road travel, which is useful for our research.

The presented example is based on the following algorithm: In Step 1, we define the analysis area as a bounded subset of the pane (by loading a polygon layer representing the selected area into ArcGIS Pro). In Step 2, we define the set of all generators or seeds (point layers representing airports and railway stations located within the analysis area). In Step 3, we generate Voronoi sets around these generators or seeds within the bounded subset of the pane (using the Create Thiessen Polygons tool in ArcGIS Pro).

The shape of the Voronoi sets is influenced by the bounded subset of the pane, the number of generators or seeds, airports and railway stations, and the software environment in which they are generated. Since this study was conducted using ArcGIS Pro, the final form of the Voronoi sets is determined by the specific method implemented in the software. In ArcGIS Pro, Voronoi sets are created by first connecting the input points into a triangulated irregular network that satisfies the Delaunay criterion. The perpendicular bisectors of the triangle edges are then calculated, defining the edges of the sets. Additionally, the final boundary of the output layer is slightly larger than the extent of the input points (ESRI, 2025).

4. FINDINGS

4.1 Results

In the study, we included 299 airports and 171 railway stations. On average, the Voronoi set area for an airport is 16,258.44 km², while for a railway station, it is 28,428.46 km². The smallest Voronoi set area for an airport is 1,267.27 km², covered by Ålesund Airport. The largest Voronoi set area for an airport is 68,285.55 km², covered by Tromsø Airport. The smallest Voronoi set area for a railway station is 1,052.51 km², covered by München railway station. The largest Voronoi set area for a railway station is 469,818.37 km², covered by Sundsvall railway station. The average area of a Voronoi set for railway stations is larger because there are fewer stations compared to airports. The average area of a Voronoi set for airports is smaller because there are more airports, resulting in a more uniformly distributed division of areas. The Voronoi diagram for air transportation is shown in Figure 1, while the Voronoi diagram for rail transportation is shown in Figure 2.

DRAFT VERSION

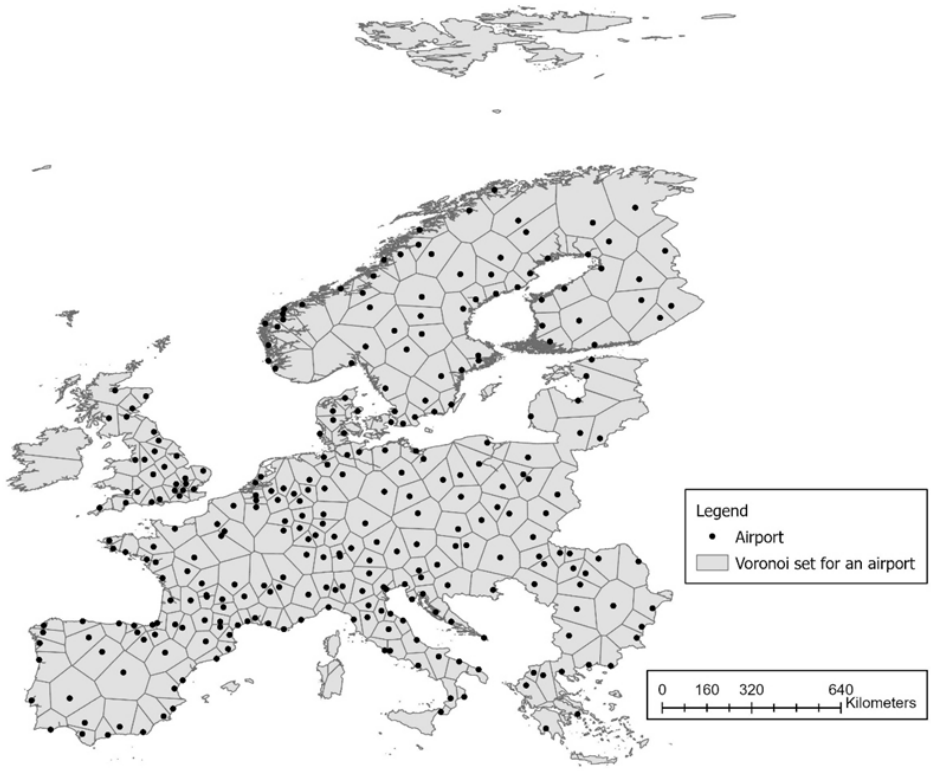


Figure 1: Voronoi diagram for air transportation

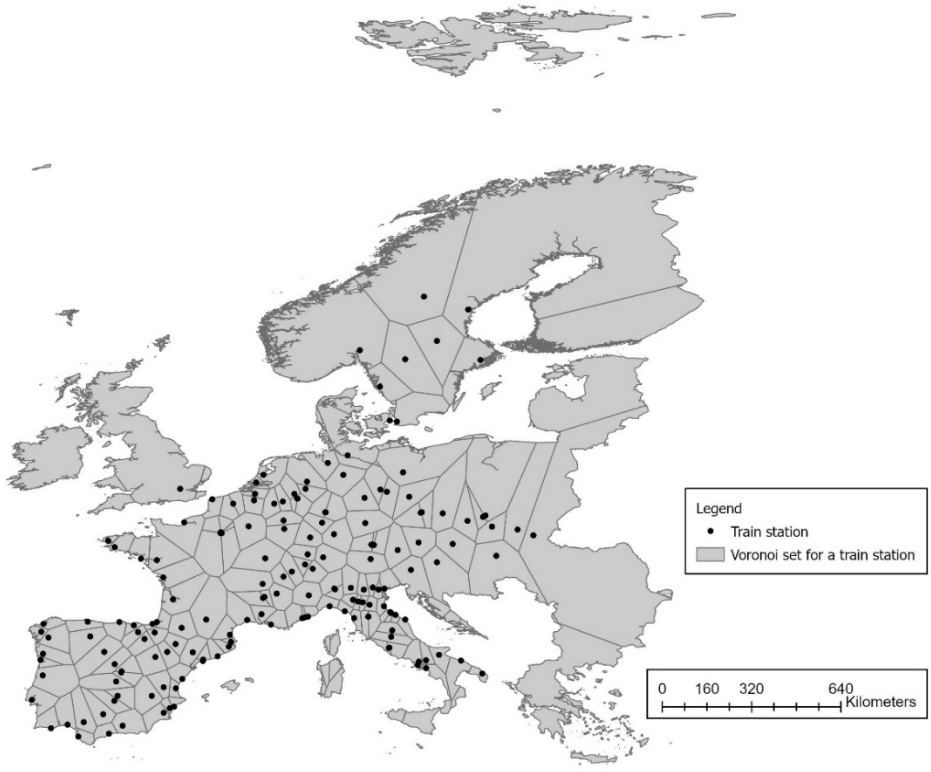


Figure 2: Voronoi diagram for rail transportation

4.2 Discussion

By comparing Figure 1 and Figure 2, we can identify areas where a lack of railway infrastructure is evident, as indicated by the larger area covered by individual Voronoi set. In Northern, Eastern, and Southeastern Europe, there is a noticeable underdevelopment of railway infrastructure. While the geographic and climatic conditions in Northern Europe act as limiting factors, Eastern and Southeastern Europe does not face such extreme conditions, suggesting that the underdevelopment of railway infrastructure in this region is primarily due to other factors. The dominant mode of transport in Northern, Eastern, and Southeastern Europe is therefore air travel, as indicated by the number of airports in these areas. In contrast, Central Europe, Italy, and the Iberian Peninsula have a much denser distribution of railway stations compared to other parts of the area covered in our research. It is relevant to highlight the case of the United Kingdom and its railway stations. The reason why only one railway station from the United Kingdom is included in the analysis is that this station serves as the starting or ending point for rail connections between the British Isles and mainland Europe. As a result, the Voronoi set representing the area around this station covers a significantly larger area than the average.

From the comparison between airport and railway station areas, Voronoi sets, we can observe that airport areas are much more uniform than those of railway stations. This is also reflected in the statistical data, with the standard deviation for airport areas being 11,566.92 km², while for railway stations it is 61,320.46 km². This can be also linked to the following fact: airport complexes occupy a much larger spatial area compared to railway stations, but they can more easily overcome geographical constraints such as bodies of water, hilly terrain, or highly urbanized areas, which on the other hand limit the construction of railway tracks. This enables a more uniform distribution of airports across the territory, as they can be more easily integrated into the spatial layout.

5. CONCLUSION

The primary value of this research lies in the visual insights provided by the graphical material, which can be useful for both decision-makers and the public. The size of Voronoi set area can be used to assess the state of infrastructure in a given region. Smaller polygon areas indicate a higher number of seeds, airports or railway stations, while larger polygon areas suggest a lower number of such facilities. This allows us to identify regions with well-developed or underdeveloped infrastructure. Furthermore, this approach can support modeling of transportation mode choice, particularly in the context of selecting between air travel and high-speed rail, a topic of increasing relevance within the EU. Based on the findings, we argue that the Voronoi diagram is an effective tool that can be applied within the framework of location theory.

One limitation of this study is that we focused exclusively on airports recognized as major mobility hubs for interregional and international travel, leaving out smaller airports that also serve passengers and play an important role in connecting certain areas. Similarly, for railway stations, we included only those recognized as mobility hubs for travel between frequently visited locations, omitting a significant number of other (smaller) stations. A potential extension of the research would involve incorporating an additional criterion, population distribution, which can reflect the demand for transport, into the analysis by using weighted Voronoi sets. We could exclude islands, such as Sardinia, from the covered area. This could be justified by the fact that, when comparing air and rail travel, we should analyze them based on comparable journeys. The ability of airplanes to cross bodies of water more easily gives them an advantage over trains. However, individuals living on islands can still use trains for their travels on the mainland of Europe once they have crossed the body of water separating the island from the mainland. Also, the results could be presented by individual countries/regions. This would allow for the development of measures at the national/regional level, as there are significant differences between countries/regions. Measures aimed at promoting the use of (high-speed) rail would certainly differ for Spain and Italy compared to Poland, Romania, or Bulgaria, as the current development of high-speed rail infrastructure gives the former a clear advantage.

We would also like to highlight a general methodological issue observed in other studies. Many studies do not specify the basis for generating the Voronoi diagram layers. These layers may be created from random points to represent the influence area of a certain feature, or as a uniform grid where each cell has the same area and serves as a baseline for calculating specific indicators. While experts in the field may understand the basis of the Voronoi diagram from context, this may not be clear to other readers, which can reduce the clarity and transparency of the research results. As an example, in the studies by Raghu Kisore and Koteswaraiah (2017) and Jordan et al. (2022), it was not specified on what basis the Voronoi sets were generated, and based on the research results, we were also unable to determine the procedure.

DRAFT VERSION

ACKNOWLEDGMENT

This research received no external funding.

LITERATURE

Boonprong, S., Punturasan, N., Varnakovida, P., & Prechathamwong, W. (2024). Towards sustainable urban mobility: Voronoi-based spatial analysis of EV charging stations in Bangkok. *Sustainability*, 16(11), 4729. https://doi.org/10.3390/su16114729

Burkey, M. L., Bhadury, J., & Eiselt, H. A. (2011). Voronoi diagrams and their uses. *Foundations of Location Analysis*, 445-470. https://doi.org/10.1007/978-1-4419-7572-0_19

Geofabrik. (2025). Europe. OpenStreetMap Data Extracts. Retrieved May 24, 2025, from https://download.geofabrik.de/

Dong, P. (2008). Generating and updating multiplicatively weighted Voronoi diagrams for point, line and polygon features in GIS. *Computers & Geosciences*, 34(4), 411-421. https://doi.org/10.1016/j.cageo.2007.04.005

ESRI. (2025). *Create Thiessen Polygons (Analysis)*. Retrieved May 24, 2025, from https://pro.arcgis.com/en/pro-app/latest/tool-reference/analysis/create-thiessen-polygons.htm

Eurail. (2025). *High-Speed Trains*. Retrieved July 25, 2025, from https://www.eurail.com/en/plan-your-trip/trip-ideas/trains-europe/high-speed-trains

European Environmental Agency (2021). *Transport and environment report 2020. Train or plane?* Retrieved May 24, 2025, from https://www.eea.europa.eu/publications/transport-and-environment-report-2020

Figurska, M., Dawidowicz, A., & Zysk, E. (2022). Voronoi Diagrams for Senior-Friendly Cities. International *Journal of Environmental Research and Public Health*, 19(12), 7447. https://doi.org/10.3390/ijerph19127447

Flight connections. (2025). FlightConnections: All flights worldwide on a map! https://www.flightconnections.com/

Jiao, H., & Feng, S. (2024). Towards resilient cities: optimizing shelter site selection and disaster prevention life circle construction using GIS and supply-demand considerations. *Sustainability*, 16(6), 2345. https://doi.org/10.3390/su16062345

Jordán, J., Palanca, J., Martí, P., & Julian, V. (2022). Electric vehicle charging stations emplacement using genetic algorithms and agent-based simulation. *Expert Systems with Applications*, 197, 116739. https://doi.org/10.1016/j.eswa.2022.116739

Kisore, N. R., & Koteswaraiah, C. B. (2017). Improving ATM coverage area using density based clustering algorithm and voronoi diagrams. *Information Sciences*, 376, 1-20. https://doi.org/10.1016/j.ins.2016.09.058

Kloskowski, D., Chamier-Gliszczynski, N., & Królikowski, T. (2024). Optimization of shopping centers location using GIS methods. *Procedia Computer Science*, 246, 3938-3946. https://doi.org/10.1016/j.procs.2024.09.168

Marianov, V., & Eiselt, H. A. (2024). Fifty years of location theory-A selective review. *European Journal of Operational Research*, 318(3), 701-718. https://doi.org/10.1016/j.ejor.2024.01.036

Mota, D. R., Takano, M., & Taco, P. W. G. (2014). A method using GIS integrated Voronoi diagrams for commuter rail station identification: a case study from Brasilia (Brazil). *Procedia-Social and Behavioral Sciences*, 162, 477-486. https://doi.org/10.1016/j.sbspro.2014.12.229

Omio. (2025). Book, search & compare trains, buses, flights & ferries – Omio. https://www.omio.com/

Prakobkaew, P., & Sirisumrannukul, S. (2022). Practical grid-based spatial estimation of number of electric vehicles and public chargers for country-level planning with utilization of GIS data. *Energies*, 15(11), 3859. https://doi.org/10.3390/en15113859

Shahparvari, S., Fadaki, M., & Chhetri, P. (2020). Spatial accessibility of fire stations for enhancing operational response in Melbourne. *Fire safety journal*, 117, 103149. https://doi.org/10.1016/j.firesaf.2020.103149

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

SIGNED TOTAL ROMAN DOMINATION OF COMPLETE SPLIT GRAPHS

ALEKSANDAR SAVIĆ[1], ZORAN MAKSIMOVIĆ[2]
[1] University of Belgrade, Faculty of Mathematics, Belgrade, aleks3rd@gmail.com, 0009-0003-8568-4260 [2]
University of defence, Military Academy, Belgrade, zoran.maksimovic@gmail.com, 0000-0001-9827-2490

Abstract: *In this paper, we investigate the concept of signed total Roman domination (STRD) in the context of complete split graphs. Domination in graph theory is a well-studied area with numerous variants, each imposing different constraints on how vertices dominate or are dominated by others. The signed total Roman domination is a recently introduced variant that combines elements of signed domination and Roman domination, incorporating both vertex labeling and edge-based conditions. Optimal value for complete spit graph is equal to 3, which is proved using integer linear programming.*

Keywords: *Graphs, Domination, Signed Total Roman Domination, Complete split graphs*

1. INTRODUCTION

Problem of Roman domination was introduced in context of military strategy concerning defence of Roman Empire in the time of famous emperor Constantine the Great. Falling fortunes of Roman empire during the 4th century AD needed accepting new defensive strategy in depth. To much legions in one province could lead to rebelion and rise of pretenders for the imperial purple and to small military detachments meant that province was undefended. Solution strategy, considered in papers (ReVelle and Rosing, 2000) and (Stewart, 1999), required that in any province there are 0, 1 or 2 legions. Province was considered defended if it has inside its border at least 1 legion or that in some neighboring province were stationed 2 legions. Naturally, goal was to defend the whole Empire with as small number of legions as possible.

Mathematically, in graph interpretation, Roman Empire could be represented as a graph, where provinces are vertices and edges exist if the provinces have common border or well traveled sea rout between them. Now, problem of Roman domination (RDP) could be interpreted as follows:

For a given simple graph $G(V,E)$, find $\min \sum_{v\in V} f(v)$, where Roman dominating function $f : V \to \{0,1,2\}$ subject to that for every v such that $f(v) = 0$, there exists at least one of its adjacent vertex w such that $f(w) = 2$.

There are many modifications of the problem of Roman domination in literature. In this paper we will consider problem of signed total Roman domination (STRDP) introduced in (Volkmann, 2016b). In graph terms STRDP can be formulated as follows:

For a given simple graph $G(V,E)$, find $\min \sum_{v\in V} f(v)$, where signed total Roman dominating function $f : V \to \{-1,1,2\}$ subject to that for every vertex v its adjacent vertices called open neighborhood and denoted as $N(v)$, must satisfy $\sum_{w\in N(v)} f(w) \geq 1$, and every vertex u for which $f(u) = -1$ is adjacent to at least one vertex v for which $f(v) = 2$. This minimmum is denoted $\gamma_{stR}(G)$.

Instead of Roman domination, signed Roman domination problems emphasize the risk of attack on the province were is no legion but some small detachments commonly called auxiliary troops. This signed modification is more realistic than just indicating that in a province there are no troops and putting the value of that vertex by 0. Indicating that in the province is small garrison by putting -1 make it defense more demanding.

In this paper we will consider solving STRDP on special class of graphs - complete split graphs.

Complete split graphs $K^*_{k,n-k}$ are defined as simple graphs $G(V,E)$ where n is the order of graph i.e. $|V| = n$. Formally, $V_1 = \{v_i | 1 \leq i \leq k\}, V_2 = \{v_i | k+1 \leq i \leq n\}, V = V_1 \bigcup V_2, F_1 = \{v_i v_j | 1 \leq i < j \leq k\}$, $F_2 = \{v_i v_j | 1 \leq i \leq k, k+1 \leq j \leq n\}$ and $E = F_1 \bigcup F_2$. In other words set V is partitioned in 2 subsets V_1 and V_2 where $|V_1| = k$, $|V_2| = n-k$, $V_1 \cup V_2 = V$ and $V_1 \cap V_2 = \emptyset$. Graph induced by subset V_1 is a clique, while vertices from V_2 are adjacent only to vertices from V_1, and $deg(v) = \begin{cases} n-1, & v \in V_1 \\ k, & v \in V_2 \end{cases}$.

Example. $K^*_{3,3}$ is given on the following figure with optimal solution of STRDP given by numbers attached to all vertices.

DRAFT VERSION

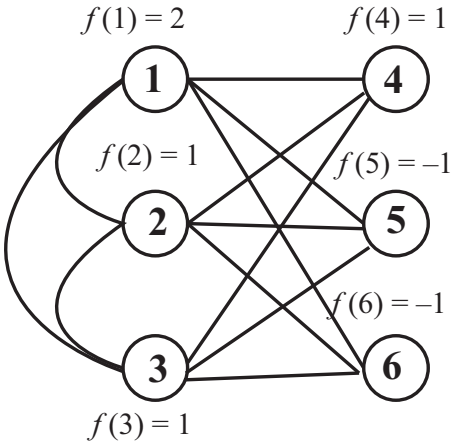


Figure 1: Signed total Roman domination for $K^*_{3,3}$

2. Previous works

Literature on Roman domination is huge and is growing over years so it is out of scope of this paper. Similarly, there is more than 15 modifications of Roman domination problem so, only introduction papers are numbered over 20 and this also is out of scope of this paper. We will limit previous work only on papers directly linked to STRDP.

In paper (Volkmann, 2016b) STRDP is introduced and is discussed. As modification of RDP it is also NP hard. Additionally, the author presented optimal solution for some classes of graphs, notably: stars, complete graphs, complete bipartite graphs $K_{p,p}$, cycles and paths.

The same author is broadening research of STRDP problem in (Volkmann, 2016a). Author introduces some new concepts. First, a set $\{f_1, f_2, \ldots, f_d\}$ of distinct signed total Roman dominating functions on G with the property that $\sum_{i=1}^{d} f_i(v) \leq 1$ for each $v \in V(G)$, is called a *signed total Roman dominating family (of functions) on G*. The maximum number of functions in a signed total Roman dominating family on G is the *signed total Roman domatic number of G*, denoted by $d_{stR}(G)$. Author presented sharp bounds for $\gamma_{stR}(G)$ and $d_{stR}(G)$. In addition, the total signed Roman domatic number of some graphs is determined.

The following paper (Volkmann, 2017) presents research of STRDP in another direction, that is on the digraphs $D(V,E)$. In the paper are given bounds on $\gamma_{stR}(D)$, and also signed total Roman domination on some classes of digraphs.

Another approach for solving STRDP is given in (Filipović et al., 2022). For each problem authors proposed two integer linear programming (ILP) formulations, the constraint programming (CP) formulation and variable neighborhood search (VNS) method. Additionally, proofs for the correctness of the ILP formulations and a polyhedral study in which was shown that the polyhedrons of the two model relaxations are equivalent.

Also, for solving STRDP, VNS uses specifically designed penalty function that allows the appearance of slightly infeasible solutions. All proposed approaches were tested on the large number of instances. Experimental results indicate that all of them reach optimal solutions for the most of small and middle scale instances. Both ILP models have proven to be more successful than the other two methods.

3. New results

Let us now consider complete split graphs $K^*_{k,n-k}$ where $k \geq 2$, $n-k \geq 2$. We define the following integer variables:

- $x_1 = |\{u \in V_1 | f(u) = -1\}|$;
- $x_2 = |\{u \in V_2 | f(u) = -1\}|$;
- $y_1 = |\{u \in V_1 | f(u) = 1\}|$;
- $y_2 = |\{u \in V_2 | f(u) = 1\}|$;
- $z_1 = |\{u \in V_1 | f(u) = 2\}|$;
- $z_2 = |\{u \in V_2 | f(u) = 2\}|$.

Let us consider following integer linear program (ILP):

$$\min -x_1 - x_2 + y_1 + y_2 + 2z_1 + 2z_2 \tag{1}$$

DRAFT VERSION

subject to:

$$-x_1 - x_2 + y_1 + y_2 + 2(z_1 - 1) + 2z_2 \geq 1 \quad (2)$$
$$-x_1 + y_1 + 2z_1 \geq 1 \quad (3)$$
$$-x_1 - x_2 + (y_1 - 1) + y_2 + 2z_1 + 2z_2 \geq 1 \quad (4)$$
$$-x_1 + y_1 + 2z_1 \geq 1 \quad (5)$$
$$-(x_1 - 1) - x_2 + y_1 + y_2 + 2z_1 + 2z_2 \geq 1 \quad (6)$$
$$-x_1 + y_1 + 2z_1 \geq 1 \quad (7)$$
$$x_1 + x_2 \leq nz_1 \quad (8)$$
$$x_2 \leq kz_2 \quad (9)$$
$$x_1, x_2, y_1, y_2, z_1, z_2, \in Z^+ \cup \{0\} \quad (10)$$

Term (1) is equal to $\min \sum_{v \in V} f(v)$. Inequalities (2-7) correspond to fact that in every open neighborhood $\sum_{w \in N(v)} f(w)$ must be greater or equal to 1. Inequalities (2) and (3) correspond to vertices with $f(v) = 2$, inequalities (4) and (5) correspond to vertices with $f(v) = 1$ and inequalities (6) and (7) correspond to vertices with $f(v) = -1$.

Let us show that if there is $v \in V_1$ such that $f(v) = -1$ than in optimal solution there must be at least one vertex $w \in V_1$ such that $f(w) = 2$. This situation corresponds to $x_1 \geq 0$, $x_2 = z_1 = 0$ and $z_2 > 0$.

Since $x_1 > 0$ and $z_1 = 0$ there must be at least one vertex $v \in v_2$ such that $f(v) = 2$. Without loss of generality we can assign $f(v_{k+1}) = 2$.

Since $x_2 = 0$ $\min \sum_{v \in V_2} f(v) = n - k + 1$ which implies $f(v_{k+2}) = ... = f(v_n) = 1$. We will consider two cases: when k is odd and when k is even.

$k = 2m + 1$:

Then $\min \sum_{v \in V_1} f(v) = 1$ (it follows from condition $\min \sum_{w \in N(v), v \in V_2} f(w) \geq 1$) which implies, without loss of generality, $f(v_1) = \ldots f(v_{m+1}) = 1$ and $f(v_{m+2}) = \ldots f(v_k) = -1$. It is easy to see that if we assign $f(v_1) = 2$ and $f(v_{k+2}) = -1$ then all conditions for STRDP will be satisfied and objective function will be lower. Which proves our statement.

$k = 2m$:

Similarly as in case when k is odd, $\min \sum_{v \in V_1} f(v) = 2$ which implies $f(v_1) = \ldots f(v_{m+1}) = 1$ and $f(v_{m+2}) = \ldots f(v_k) = -1$. There is $m + 1$ vertices with $f(v) = 1$ and $m - 1$ vertices with $f(v) = -1$. Again, if we assign $f(v_1) = 2$ and $f(v_{k+2}) = -1$ then all conditions for STRDP will be satisfied and objective function will be lower. Which proves our statement.

Inequality (8) implies that in the open neighborhood of the vertex $v \in V_1$ $f(v) = -1$ there must be at least one vertex w with $f(w) = 2$.

Let us show that if there is $v \in V_2$ such that $f(v) = -1$ than in optimal solution there must be at least one vertex $w \in V_2$ such that $f(w) = 2$. This situation corresponds to $x_2 \geq 0$, $x_1 = z_2 = 0$ and $z_1 > 0$.

Since $x_2 > 0$ and $z_1 = 0$ there must be at least one vertex $v \in V_1$ such that $f(v) = 2$. Without loss of generality we can assign $f(v_1) = 2$.

Since $x_1 = 0$ $\min \sum_{v \in V_1} f(v) = k + 1$ which implies $f(v_2) = ... = f(v_k) = 1$.

Let us consider three cases: $n > 2k - 2$, $n = 2k - 2$ and $n < 2k - 2$.

$n > 2k - 2$:

Since in V_1 there is $k - 1$ vertices with $f(v) = 1$ in V_2 maximum namber of vertices with $f(v) = -1$ is $k - 2$. This is required so that $\sum_{w \in N(v), v \in V_1} f(w) \geq 1$. Since $n > 2k - 2$, it follows that $n - k > k - 2$ which implies there is at least one vertex $v \in V_2$ such that $f(v) = 1$. Without loss of generality let it be v_{k+1}.

Now, if we assign $f(v_2) = -1$ and $f(v_{k+1}) = 2$ then all conditions for STRDP will be satisfied and objective function will be lower. Which proves our statement.

$n = 2k - 2$: Since in V_1 there is $k - 1$ vertices with $f(v) = 1$ in V_2 maximum namber of vertices with $f(v) = -1$ is $k - 2$. This is required so that $\sum_{w \in N(v), v \in V_1} f(w) \geq 1$. Since $n = 2k - 2$, it is not necessary that in V_2 there is v such that $f(v) = 1$. This implies that all vertices v in V_2 can have $f(v) = -1$. Now we can assign without loss of generality for $k \geq 5$, $f(v_2) = f(v_3) = f(v_4) = -1$, and $f(v_{k+1}) = f(v_{k+2}) = 2$ (in V_1 there is $k - 4$ vertices with $f(v) = 1$ and in V_2 there is $k - 4$ vertices with $f(v) = -1$). All conditions for STRDP will be satisfied and objective function will be equal, and though objective functions are equal our statement holds.

$n < 2k - 2$: Since in V_1 there is $k - 1$ vertices with $f(v) = 1$ in V_2 maximum namber of vertices with $f(v) = -1$ is $k - 3$. This is required so that $\sum_{w \in N(v), v \in V_1} f(w) \geq 1$. Since $n < 2k - 2$, it is not necessary that in V_2 there is v such that $f(v) = 1$. This implies that all vertices v in V_2 can have $f(v) = -1$. Now we can assign without loss of generality, $f(v_2) = f(v_3) = -1$, and $f(v_{k+1}) = 2$ (in V_1 there is $k - 3$ vertices with $f(v) = 1$ and

DRAFT VERSION

in V_2 there is $k-4$ vertices with $f(v)=-1$). All conditions for STRDP will be satisfied and objective function is lower and our statement holds.

Similarly, inequality (9) implies that in the open neighborhood of the vertex $v \in V_2$ with $f(v)=-1$ there must be at least one vertex w with $f(w)=2$.

On close inspection conditions (3), (5) and (7) are identical so two of these are redundant. Also, conditions (2), (4) and (6) can be transformed into:

$$-x_1-x_2+y_1+y_2+2z_1+2z_2 \geq 3 \tag{11}$$
$$-x_1-x_2+y_1+y_2+2z_1+2z_2 \geq 2 \tag{12}$$
$$-x_1-x_2+y_1+y_2+2z_1+2z_2 \geq 0 \tag{13}$$

From this it is easy to see that if condition (11) is satisfied then both conditions (12) and (13) are satisfied so they are also redundant. Since we are considering complete split graph it holds $x_1+y_1+z_1=k$ and $x_2+y_2+z_2=n-k$, i.e. $x_1=k-y_1-z_1$ and $x_2=n-k-y_2-z_2$. Now, if we include them in conditions (1), (2), (3) (8), (9) and (10), we obtain following new ILP model:

$$\min -n+2y_1+2y_2+3z_1+3z_2 \tag{14}$$

subject to:

$$2y_1+3z_1 \geq 1+k \tag{15}$$
$$2y_1+2y_2+3z_1+3z_2 \geq 3+n \tag{16}$$
$$y_1+y_2+(n+1)z_1+z_2 \geq n \tag{17}$$
$$y_2+(k+1)z_2 \geq n-k \tag{18}$$
$$y_1,y_2,z_1,z_2, \in Z^+ \cup \{0\} \tag{19}$$

With this new ILP model we can now prove following theorem for $K^*_{k,n-k}$.

Theorem. For $k \geq 3$ and $n \geq 4$, it holds $\gamma_{stR}(K^*_{k,n-k})=3$.

Proof. Let us consider dual problem of the program (14)-(19). Denoting the dual variables u_1,u_2,u_3,u_4 we obtain

$$\max -n+(1+k)u_1+(3+n)u_2+nu_3+(n-k)u_4 \tag{20}$$

subject to:

$$2u_1+2u_2+u_3 \leq 2 \tag{21}$$
$$2u_2+u_3+u_4 \leq 2 \tag{22}$$
$$3u_1+3u_2+(n+1)u_3 \leq 3 \tag{23}$$
$$3u_2+u_3+(k+1)u_4 \leq 3 \tag{24}$$
$$u_1,u_2,u_3,u_4 \geq 0 \tag{25}$$

In the primal problem x_1+x_2 must be greater than 1, since if $x_1=x_2=0$ means that $y_1+y_2+z_1+z_2=n$ implying that objective function in (14) has value at least n, which is obviously not minimum equal to 3.

So, we conclude that $x_1+x_2 \geq 1$ and this implies from constraint (8) that $z_1 \geq 1$. Now, since $z_1 \in Z$ we can obtain conclusion that constraint (17) can be interpreted as $y_1+y_2+(n+1)z_1+z_2 \geq n+1 > n$. It means that in complementary slackness condition $u_3(y_1+y_2+(n+1)z_1+z_2-n)=0$ dual variable u_3 must be equal to 0.

From dual constraint (21) we obtain that $u_1+u_2 \leq 1$. Since coefficients by u_1 and u_2 in the dual objective function are $1+k$ and $3+n$ respectively, and $3+n > 1+k$ for any n and k we can conclued that in optimal solution for the dual problem must be $u_1=0$ and $u_2=1$.

From complementary slackness condition $u_2(2y_1+2y_2+3z_1+3z_2-3-n)=0$ follows that $2y_1+2y_2+3z_1+3z_2=3+n$ which gives us primal and dual objective function optimal value equal to 3.

Finally, since $\gamma_{stR}(K^*_{k,n-k})$ is equal to minimum of (1) we conclude that $\gamma_{stR}(K^*_{k,n-k})=3$. ◄

It is interesting to see what happens in degenerated cases when $k=1$ or $n-k=1$. If $n-k=1$ graph is clique K_n, with $\gamma_{stR}(K_n)=3$ (Volkmann, 2016b). When $k=1$, graph is star S_{n-1} with $\gamma_{stR}(S_{n-1})=3$ (Volkmann, 2016b).

DRAFT VERSION

4. CONCLUSIONS

In this paper signed total Roman domination of complete split graphs are considered. Exact value is equal to 3, for $k \geq 2$ and $n - k \geq 2$. The result is obtained by solving integer linear programming formulation with 4 variables and 4 constraints.

Future investigations could be directed towards studying signed total Roman domination across various graph classes, along with examining analogous graph invariants in the context of complete split graphs.

References

Filipović, V., Matić, D., & Kartelj, A. (2022). Solving the signed roman domination and signed total roman domination problems with exact and heuristic methods [To appear].

ReVelle, C. S., & Rosing, K. E. (2000). Defendens imperium romanum: A classical problem in military strategy. *American Mathematical Monthly*, *107*(7), 585–594.

Stewart, I. (1999). Defend the roman empire! *Scientific American*, *281*, 136–138.

Volkmann, L. (2016a). On the signed total roman domination and domatic numbers of graphs. *Discrete Applied Mathematics*, *214*, 179–186.

Volkmann, L. (2016b). Signed total roman domination in graphs. *Journal of Combinatorial Optimization*, *32*(3), 855–871.

Volkmann, L. (2017). Signed total roman domination in digraphs. *Discussiones Mathematicae Graph Theory*, *37*, 261–272. https://doi.org/10.7151/dmgt.1929

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

DISTANCE-EDGE-MONITORING NUMBER OF COMPLETE SPLIT GRAPHS

MILICA MILIVOJEVIĆ DANAS[1], JOZEF KRATICA[2]

[1] University of Kragujevac, Faculty of Engineering, Kragujevac, milica.milivojevic@kg.ac.rs, 0000-0002-5182-1197
[2] Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade, jkratica@mi.sanu.ac.rs, 0000-0002-9752-0971

Abstract: *In this paper is studied the distance-edge-monitoring problem which is recently introduced, for special class of graphs, named complete split graphs $K^*_{k,n-k}$. We are stated and proved formula for the distance-edge-monitoring number for this graphs and its exact value is equal to k.*

Keywords: *Distance-edge-monitoring number, complete split graphs, graph theory, discrete mathematics.*

DRAFT VERSION

1. INTRODUCTION

The distance-edge-monitoring problem was first introduced in (Foucaud et al., 2022). Let $G = (V,E)$ be a simple connected undirected graph with vertex set V and edge set E. Then, a set M of vertices of a graph G is called a distance-edge-monitoring set if, for every edge e in G, there exists a vertex x from M and a vertex y from G such that e is part of all shortest paths between x and y. The set M with minimal cardinality is called the distance-edge-monitoring set of the graph and the smallest size of such set in graph G is denoted by $dem(G)$. The vertices in M serve as monitoring points within a network represented by G. When an edge e fails, the shortest path distance between x and y increases, allowing us to detect the failure. Interestingly, not only can we identify that an edge has failed, but we can also pinpoint exactly which edge is affected.

In (Foucaud et al., 2022), it is showed that for a nontrivial connected graph G of order n holds $1 \leq dem(G) \leq n-1$ and it is equal to 1 if and only if G is a tree, while it is equal to $n-1$ if and only if it is a complete graph. The authors computed the exact value of distance-edge-monitoring number for grid, hypercubes and complete bipartite graphs. Moreover, it is characterized graphs G with $dem(G) = 2$. Also, the authors provides both lower and upper bounds for $dem(G)$. Finally, the authors proved that determining $dem(G)$ for graph G is an NP-complete problem in a general case.

In (Kartelj et al., 2024), it presents an exact procedure, based on an integer linear programming model, to determine the probe set with the smallest number of elements. They proposed the first-ever integer programming model for distance-edge-monitoring problem and they presented empirical results for five classes of graphs, where the maximum number of vertices is 4096 and the maximum number of edges is 1599797. The effectiveness of this approach is demonstrated using random graphs that resemble real-world networks. However, the authors highlight that their method is not tailored to a specific type of graph, making it applicable in a broad range of case.

There are several related paper for distance-edge-monitoring problem in the literature. In the article (Li et al., 2025), the authors presented upper and lower bounds for Cartesian graph products, and in the cases where upper and lower bounds could be obtained, they provided a characterization of graphs that reach these bounds. In article (C. Yang, Klasing, Mao, and Deng, 2024), it also provides the lower and upper bounds for $P(M,e),EM(x))$ and $dem(G)$, as well as the characterization of the graphs on which they are reach.

After initial works, we will mention some additional papers that study distance-edge-monitoring problem from various theoretical and practical perspectives, such as (C. Yang, Klasing, He, and Mao, 2023), (G. Yang and He, 2023) and (Lepič, 2022).

Next, some theoretical properties of distance-edge-monitoring problem, which documented in the literature, will be presented, which is used in the following section.

1.1. Definitions and basic properties

For graph $G = (V,E)$, distance between two vertices u and v, where $u,v \in V$, is the length of the shortest path between this vertices in G and denoted as $d_G(u,v)$.

► Definition 1. (Foucaud et al., 2022) For a set M of vertices and an edge e of a graph G, let $P(M,e)$ be the set of pairs (x,y) with x a vertex of M and y a vertex of $V(G)$ such that $d_G(x,y) \neq d_{G-e}(x,y)$. In other words, e belongs to all shortest paths between x and y in G.

► Definition 2. (Foucaud et al., 2022) For a vertex x, let $EM(x)$ be the set of edges e such that there exists a vertex $v \in V(G)$ with $(x,v) \in P(x,e)$.

If $e \in EM(x)$, we say that e is monitored by x.

► Definition 3. (Foucaud et al., 2022) A set M of vertices of a graph G is *distance-edge-monitoring* if every edge e of $E(G)$ is monitored by some vertex of M, i.e., the set $P(M,e)$ is nonempty. Equivalently, $\bigcup_{x\in M} EM(x) = E(G)$.

An integer linear programming (ILP) model of the distance-edge-monitoring problem is given in the paper (Kartelj et al., 2024). The proposed integer programming model is:

$$\min \sum_{v\in V} x_v \tag{1}$$

s.t.

$$\sum_{v\in V} C_{v,e} \cdot x_v \geq 1, \quad for\, e \in E \tag{2}$$

$$x_v \in \{0,1\}, \quad for\, v \in V \tag{3}$$

The variable x_v, for $v \in V$, equals 1 if the vertex v is included in the distance-edge-monitoring set, otherwise, it equals 0. Hence, if $x_v = 1$ for some $v \in V$, then $v \in M$. Evidently, the objective function (1) relates to the formulation of a distance-edge-monitoring set problem.

The constant $C_{v,e}$ for $(v,e) \in V \times E$ equals 1 if for a pair (v,e) there exists a vertex $u \in V$ so that edge e belongs to all the shortest paths between vertices v and u. Practically, if $e \in EM(v)$ then $C_{v,e} = 1$, i.e. $C_{v,e} = 0$ when $e \notin EM(v)$. Constraints (2) impose the distance-edge-monitoring condition and that meaning that for every $e \in E$, there must be at least one $v \in M$ that monitors it.

It should be noted that $V(G)$ is always a distance-edge-monitoring set of G, so $dem(G)$ is always well-defined.

► Definition 4. A Complete split graph $K^*_{k,n-k}$ has vertices $V(K^*_{k,n-k}) = V_1 \bigcup V_2$, where $V_1 = \{v_1, ..., v_k\}$ and $V_2 = \{v_{k+1}, ..., v_n\}$. Edges $E(K^*_{k,n-k}) = E_1 \bigcup E_2$, where $E_1 = \{v_i v_j | 1 \leq i < j \leq k\}$ and $E_2 = \{v_i v_j | 1 \leq i \leq k, k+1 \leq j \leq n\}$.

It is easy to calculate that these graphs have n vertices and $\binom{n}{2}$ - $\binom{n-k}{2}$ edges. It should be noted that the diameter of these graphs is two and only shortest paths between two vertices from V_2 has length 2. All other shortest paths are trivial, with length 1.

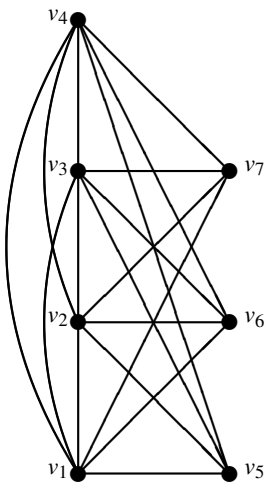


Figure 1: Graph $K^*_{4,3}$

The distance-edge-monitoring sets are relaxations of vertex covers. A vertex cover C of a graph G is defined as a set of vertices where each edge in G has at least one of its endpoints included in C. The minimum size of a

DRAFT VERSION

vertex cover for graph G is denotes as $vc(G)$. In some graphs, every distance-edge-monitoring set qualifies a vertex cover.

▶ Theorem 1. (Foucaud et al., 2022) In any graph G of order n, any vertex cover of G is a distance-edge-monitoring set, and thus $dem(G) \leq vc(G) \leq n-1$. Moreover, we have $dem(G) = n-1$ if and only if G is the complete graph of order n.

▶ Theorem 2. (Foucaud et al., 2022) Let G be a connected graph with at least one edge. We have $dem(G) = 1$ if and only if G is a tree.

▶ Corollary 1. (Foucaud et al., 2022) We have $dem(K_{a,b}) = vc(K_{a,b}) = min\{a,b\}$ and $dem(H_d) = vc(H_d) = 2^{d-1}$.

2. NEW THEORETICAL RESULTS

In the next theorem it will be obtained and proved distance-edge-monitoring number of any complete split graph $K^*_{k,n-k}$.

▶ Theorem 3. For $k \geq 2$ and $n \geq k+2$, it holds $dem(K^*_{k,n-k}) = k$.

Proof. *Step 1.* $dem(K^*_{k,n-k}) \leq k$
Let $M = V_1$. We will find $P(M,e)$ for all edges from $E(K^*_{k,n-k})$.

If $e \in E_1$ then it be noted as $e = uv$ with $u \in V_1$ and $v \in V_1$. Then, $P(M,e) = P(V_1,uv)$ is the set of pairs (x,y) with $x \in V_1$ and $y \in V_1 \bigcup V_2$ such that uv belongs to all shortest paths between x and y in G. Since $x \in V_1$ and $y \in V_1 \bigcup V_2$ so $d(x,y) = 1 \Rightarrow x = u, y = v$ implying $P(M,e) = P(V_1,uv) = \{(u,v)\}$.

If $e \in E_2$ then it be noted as $e = uw$ with $u \in V_1$ and $w \in V_2$. Then, $P(M,e) = P(V_1,uw)$ is the set of pairs (x,y) with $x \in V_1$ and $y \in V_1 \bigcup V_2$ such that uw belongs to all shortest paths between x and y in G. The fact $x \in V_1$ and $y \in V_1 \bigcup V_2$ so $d(x,y) = 1 \Rightarrow x = u, y = w$ implying $P(M,e) = P(V_1,uw) = \{(u,w)\}$.

Since, for each $e \in E(K^*_{k,n-k}) = E_1 \bigcup E_2$, it holds $P(M,e) = P(V_1,e) \neq \emptyset$, then by Definition 2 set $M = V_1$ is distance-edge-monitoring for graph $K^*_{k,n-k}$, implying $dem(K^*_{k,n-k}) \leq k$.

Step 2. $dem(K^*_{k,n-k}) \geq k$
Let we fix set M^* such that $|M^*| \leq k-1$. We will find $P(M^*,e)$ for each edge from $E(K^*_{k,n-k})$. Since $|M^*| \leq k-1$ then there exists vertex $u \in V_1$ such that $u \notin M^*$. Moreover, there exists another vertex $v \in V_1$, such that $v \notin M^*$ or vertex $w \in V_2$, such that $w \notin M^*$. In first case, for edge $e = uv$, it holds $P(M^*,e) = \emptyset$ since $u,v \notin M^*$ so edge $e = uv$ do not belongs to any shortest path of length l between $x \in M^*$ and $y \in V(K^*_{k,n-k})$ because:

- $l = 1 \Rightarrow x = u \wedge y = v$ but $x \in M^*$ while $u \notin M^*$;
- $l = 2 \Rightarrow$ edge uv with $u,v \in V_1$ could not includes in any shortest paths of length 2.

In second case, for edge $e = uw$, it holds $u,w \notin M^*$, so edge $e = uw$ could only belongs to shortest path $x-u-w$ with $y = w$ of length 2 between $x \in M^*$ and $y \in V(K^*_{k,n-k})$. On the other hand, for any $z \in V_1 \wedge z \neq u$ other shortest path $x-z-w$ between x and $y = w$ could not include edge uw, so $P(M^*,e) = \emptyset$.

In both cases is $P(M^*,e) = \emptyset$, so M^* is not a distance-edge-monitoring set for graph $K^*_{k,n-k}$. Since there no exist a distance-edge-monitoring set of cardinality less than k, then it holds $dem(K^*_{k,n-k}) \geq k$. ◀

Further, it should be discussed two special situations of degenerated complete split graphs:

- $k = 1$: It is easy to see that $K^*_{1,n-1} \cong S_n$, where S_n is a star with n vertices. Then, by Theorem 2, since S_n is a tree, we have the following result $dem(K^*_{1,n-1}) = dem(S_n) = 1$;
- $n-k = 1$: It is easy to see that $K^*_{n-1,1} \cong K_n$. Then, by Theorem 1, we have the following result $dem(K^*_{n-1,1}) = dem(K_n) = n-1$.

3. CONCLUSIONS

In this paper the exact value of distance-edge-monitoring number of the complete split graphs is stated and proved. For $k \geq 2$ and $n \geq k+2$ distance-edge-monitoring number of each complete split graph $K^*_{k,n-k}$ is equal k. Future work can be directed towards solving ongoing research on complete split graphs or determining the exact value of the distance-edge-monitoring number of some related graphs.

DRAFT VERSION

LITERATURE

Foucaud, F., Kao, S.-S., Klasing, R., Miller, M., & Ryan, J. (2022). Monitoring the edges of a graph using distances. *Discrete Applied Mathematics*, *319*, 424–438.

Kartelj, A., Filipović, V., & Kratica, J. (2024). Integer programming model for distance-edge-monitoring problem. *Yugoslav Journal of Operations Research*. https://doi.org/10.2298/YJOR230815016K

Lepič, V. (2022). Distance edge monitoring set problem with respect to structural parameters. *B.S. thesis, Faculty of Information Technology CTU in Prague, Department of Theoretical Computer Science*.

Li, W., Klasing, R., Mao, Y., & Ning, B. (2025). Monitoring the edges of product networks using distances. *Journal of Computer and System Sciences*, *148*, 103602.

Yang, C., Klasing, R., He, C., & Mao, Y. (2023). Perturbation results for distance-edge-monitoring numbers. *arXiv preprint arXiv:2301.02507*. https://doi.org/10.48550/arXiv.2301.02507.

Yang, C., Klasing, R., Mao, Y., & Deng, X. (2024). On the distance-edge-monitoring numbers of graphs. *Discrete Applied Mathematics*, *342*, 153–167.

Yang, G., & He, C. (2023). Distance-edge-monitoring sets in hierarchical and corona graphs. *Journal of Interconnection Networks*, *23*, 2250003.

DRAFT VERSION

HEURISTICS

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

MULTI-START LOCAL SEARCH FOR SUGAR BEET TRANSPORTATION PROBLEM WITH EQUITY REGULATIONS FOR GROWERS

ANA ANOKIĆ[1], ĐORĐE STAKIĆ[2], DRAGANA DRENOVAC[3], TATJANA DAVIDOVIĆ[4]

[1] Department School of Information and Communication Technologies, Academy of Technical and Art Applied Studies Belgrade, ana.anokic@ict.edu.rs, 0000- 0002-9159-3491
[2] Faculty of Economics and Business, University of Belgrade, djordje.stakic@ekof.bg.ac.rs, 0000- 0002-3241-4289
[3] University of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, drenovac@sf.bg.ac.rs, 0000-0002-6817-0538
[4] Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade, tanjad@mi.sanu.ac.rs, 0000-0001-9561-5339

Abstract: *The sugar beet transportation problem under growers' equity regulations is considered. The solution of the problem represents the optimal plan for the delivery of sugar beet collected by several growers at different storage piles to a single sugar mill. The vehicle fleet is homogeneous and vehicles are used several times during the day. Requirements in amount of transported beet must be satisfied for each day of the planning period. Equity regulations are related to fairness among growers, providing the equal opportunities in delivering sugar beet when it is fresh, contains large amount of sucrose, and therefore, is more profitable. Constraints related to equity regulations are formulated as the minimum percentage of beet that must be transported on the harvesting day for each grower. The goal is to minimize the number of used vehicles while maximizing the total sucrose content of the collected sugar beet. The Multi-start Local Search (MLS) metaheuristic is proposed for the considered problem, evaluated on the set of generated instances against CPLEX exact solver and an existing metaheuristic approach based on the Greedy Randomized Adaptive Search Procedure (GRASP). Our experimental results show that MLS outperforms GRASP in average and CPLEX in the case of large-sized instances.*

Keywords: *Sugar Beet Transportation, Growers' Equity, Multi-start, Metaheuristics.*

1. INTRODUCTION

Sugar beet, usually cultivated by many individual growers, is an important raw material for sugar production. It contains sucrose that decreases over time from the moment of harvesting and storing at storage piles until the beet is processed in a sugar mill. From the economic point of view, it is profitable to deliver the collected beet and process it as soon as possible. However, the resources available for the transport organization have certain limits such as: number of vehicles, their capacity, working hours, etc. Using minimal homogeneous vehicle fleet, within the defined working hours, the daily requirements of the mill in the total amount of delivered beet must be satisfied. Although the phase of sugar beet delivering and processing, known as the campaign, lasts four to six months, the plan is usually made for several days. During this planning period growers prepare the sugar beet on storage piles and hundreds of tours should be performed each day in order to supply mill with the required amounts.

The transport is organized by sugar mill company that tends to achieve the largest possible profit, but also to keep growers, as their major cooperators, satisfied. For both participant sides, growers and the company, the large amount of sucrose collected from recently harvested sugar beet is valuable. Therefore, it is necessary to create a transportation plan that provides equal opportunities for growers to transport their sugar beet shortly after it is harvested. Besides maximizing the quality of sugar beet, the other important goal refers to the transportation costs, as a direct consequence of the number of used vehicles, that should be minimized.

The concept of fairness in terms of providing equal opportunities for growers in achieving profit is included in several studies related to sugar cane production. Growers in sugar cane production in Australia (Higgins, 1999) are grouped for joint harvesting to guarantee equity between them. A large-scale Integer Programming model is proposed for the considered problem on planning harvesting terms, determining groups, meet the transportation conditions and production requirements with the goal to maximize the total net revenue of growers and miller for the period of several years. The author proposed a Local Search method in combination with the Tabu Search technique as a solution approach for the considered problem. An application software

DRAFT VERSION

tool named SugarMax was developed (Jiao et al., 2005) to improve the profitability in sugar cane production with equity among growers provided by a constant amount of cane harvested by each harvesting group of growers throughout the harvest season. Their aim was to maximize the gain in sucrose content in three Australian sugar mill regions. The equity constrains were ensured in the problem of Sugar Cane Harvest Scheduling (Sethanan et al., 2014) by the condition that all groups of growers should have the same average sucrose content of cane per ton. A heuristic algorithm improved by Tabu Search technique was proposed for this problem. In the study (Thuankaewsing et al., 2015) in Thailand, equity is guaranteed by a harvest plan in which the gap between the productivity of each grower from the maximum one is lower by an equal percentage. The authors developed an optimization model and a Tabu Search for the considered problem. In the research (Jarumaneeroj et al., 2021) the authors included the three objectives to be maximized: sugar production, growers' equity in terms of a fair harvesting time-slot distribution, and efficiency in terms of balancing resource usage during the season for the supply chain of sugar cane. A solution approach based on a genetic algorithm is proposed and tested on real-life examples in Thailand.

In literature, several papers related to different sugar beet optimization problems (Anokić et al., 2021; Brcanov et al., 2025; Drenovac et al., 2020) can be found. To the best of our knowledge, the authors of the research (Drenovac et al., 2025) are the only authors that introduced equity constraints in the sugar beet transportation problem. They considered the problem denoted by the Sugar Beet Transportation Problem under Growers' Equity Regulations (SBT-GER). The problem consists of determining the optimal transportation plan for sugar beet collected by several growers at different storage piles. The planning period includes three days, while the equity regulations are provided by the request that a minimum percentage of the quantity of sugar beet must be collected from each grower on the day of harvest. The remaining constraints refer to the daily requirements in amount of sugar beet delivered to the mill, vehicle fleet and working hours. The objective function was composed of the two opposite goals, to maximize the total sucrose content while minimizing the number of used vehicles. For this purpose, the weighted sum method is used to combine the two objective functions and transform the bi-objective problem into a single-objective one. The authors proposed an Integer Linear Programming (ILP) model and developed two metaheuristic algorithms, based on Variable Neighborhood Search (VNS) and Greedy Randomized Adaptive Search Procedure (GRASP) for the considered problem. The experimental results of metaheuristics are mutually compared, as well as against CPLEX exact solver using a set of 80 generated instances of different sizes. For the majority of examples, VNS was superior over GRASP and the exact solver.

Multi-start Local Search (MLS) is an iterative metaheuristic algorithm consisting of two steps. The first one represents a stochastic procedure for generating an initial solution, which can be improved by the second step, Local Search. The best found solution is preserved and one iteration of the method is completed. Iterations are performed until the termination criterion is satisfied. The concept of multi-start search was proposed in (Boender et al., 1982) as a stochastic method for global optimization. A review of this simple metaheuristics is given in (Martí, 2003). Different Multi-start (MS) algorithms have been successfully applied to a variety of optimization problems. The Response Time Variability Problem that refers to arranging a list of clients in order to minimize the variability in the waiting time for resource access, was considered in (García et al., 2006). The authors developed MS, GRASP, and Particle Swarm Optimization (PSO) algorithms, where MS appears to be the best solution approach to small-sized instances, while PSO outperformed the remaining two algorithms for large-sized instances. MS and GRASP were later improved by the same authors in (Corominas et al., 2008). The MS Evolutionary Local Search, developed in (Villegas et al., 2010) for The Single Truck and Trailer Routing Problem with Satellite Depots proved to be more accurate than the hybrid GRASP and Variable Neighborhood Descent (VND), proposed as an alternative in the same study. Numerous of successful applications of MS method to different Vehicle Routing Problems can be found in the literature (López-Sánchez et al., 2014; Molina et al., 2018; Matijević et al., 2024).

This study proposes an efficient Multi-start Local Search heuristic approach developed for SBT-GER that uses the same Local Search procedure and test examples as in the GRASP approach, proposed in Drenovac et al., 2025. The reminder of the paper is organized as follows. Section 2 contains the problem definition. The description of the proposed MS metaheuristic approach is given in Section 3, while the computational results are presented and discussed in Section 4. Finally, the main conclusions are highlighted in Section 5.

2. PROBLEM DEFINITION

In the considered SBT-GER problem, the sugar mill organizes the collection and transportation of the sugar beet, prepared by several growers at different locations. It is coordinated with respect to the available vehicle fleet and the processing capacity of the mill. As the sugar beet can stay at the defined locations, i.e., storage piles for several days or even a week, depending on the weather conditions, these periods are considered as

DRAFT VERSION

planning periods. The goal in SBT-GER is to determine the optimal vehicle assignment to a set of available storage piles for each day of the planning period to ensure the continuity of sugar mill processes and to achieve a minimum possible loss of sucrose content while using the smallest number of vehicles.

We summarize the characteristics of the SBT-GER problem as follows:

- plan refers to the period of several days known as planning period
- the homogeneous vehicles perform multiple trips (tours) during the day
- each vehicle serves only one location in each tour
- each tour starts and ends it at the single depot located at the mill
- daily requirements of the mill must be satisfied for each day during the planning period
- the level of sucrose in the delivered sugar beet is monitored at the end of the working day
- a predefined minimum percentage of beet should be transported from each grower on the day of storage
- the objective is to maximize the total amount of collected sucrose from the collected sugar beet, while simultaneously minimize the number of used vehicles.

The ILP formulation of the problem with detailed notations and descriptions is presented in (Drenovac et al., 2025). However, we explain here some details required for the interpretation and understanding of the obtained computational results. The objective function represents the weighted sum of two components. The first one is calculated based on the total level of sucrose over the sugar beet amount collected during the planning period. This value is obtained according to the formula proposed in (Asadi, 2007), that represents the decay function of the sucrose content depending on its initial value, the daily loss rate coefficient, and the number of days the sugar beet is staying at a storage pile. The total sucrose content is divided by the maximum level of sucrose (initial, without any loss) that could be obtained from the collected amount. The second component of the objective function is the total number of vehicles used, also divided by the number of available vehicles and then subtracted from 1. The division by the maximum value for both components is used for normalization of their values into the range $[0,1]$, so they can be combined, while the purpose of the subtraction is to direct the two opposite goals toward maximization. In this way, it is possible to provide the objective function as the weighted sum of these two components. In addition, the equity constraints are ensured as the minimum percentage, denoted by θ, of the quantity of sugar beet that must be collected from each grower on the day of harvest.

3. THE PROPOSED MULTI-START LOCAL SEARCH APPROACH

We develop an MLS algorithm to approach the considered SBT-GER problem. This method depends on the appropriate balance between diversification provided by Random Construction step and intensification of the search represented by Local Search (LS) step. Each iteration starts with a feasible solution constructed using certain randomness. Usually, the elements of the solution are chosen at random or they are composed by random moves to avoid performing the same iterations during the search. Then, the second step (LS) intensifies the search starting from the constructed solution, to find the local optimum that is evaluated and compared with the current best. If an improvement of the current best solution is achieved, an update is made and the next iteration starts. These steps are performed until the termination criterion is satisfied. The structure of the proposed method is presented in Algorithm1.

Algorithm 1 MLS for SBT-GER
procedure MLS ($Problem\ Data$)
$L \leftarrow List\ of\ storage\ piles$;
$first \leftarrow$ true;
repeat
  $S \leftarrow RandomConstruction(L)$; //Random construction
  if (*S is not feasible***) then**
    $S \leftarrow Repair(S)$;
  end if
  if $(first)$ **then**
    $S^* \leftarrow S$;
    $first \leftarrow$ false;
  end if
  $S \leftarrow LocalSearch(S, niter)$; //Local Search
  if $(f(S) > f(S^*))$ **then**
    $S^* \leftarrow S$;
  end if
until $(The\ termination\ criterion\ is\ satisfied)$
return S^*;

DRAFT VERSION

The solution S of the SBT-GER problem is represented as a three-dimensional matrix. It is indexed by the ordinal numbers of the day d, the vehicle k used on day d, and the tour i assigned to vehicle k. The elements of the matrix S are defined as follows: $S(d,k,i)$ equals to the index of a storage pile to which a tour i of vehicle k on a day d is assigned. Empty tours of a vehicle are denoted by zeros. The dimension of the matrix S is equal to $|D| \times |K| \times G$, where $|D|$ and $|K|$ represent the number of days in the planning period and the maximum number of vehicles, respectively, while G denotes the maximum daily number of tours estimated as the floor of the quotient of the vehicle's working hours and the duration of the shortest tour.

The Algorithm 1 uses the problem data to create the list of available storage piles. A boolean variable denoted by *first* indicates the first pass through the main loop when the current best solution is initialized. In each iteration, the matrix S is constructed by the procedure *RandomConstruction***.** If it appears to be infeasible with respect to the equity constraints or the working hours, S is repaired to feasibility by the procedure *Repair*. The procedure *LocalSearch* is then applied to the solution S to improve it through the fixed number of iterations, *niter*, which represents a parameter of the procedure. If the solution obtained after LS is better than the current best, i.e., its objective function value, denoted by $f(S)$ is greater than the corresponding value of the current best $f(S^*)$, an update is completed. When the termination criterion is satisfied, the MLS procedure returns the best found solution S^*. The Random Construction procedure is described in Algorithm 2.

Algorithm 2 RANDOM CONSTRUCTION of the proposed MLS for SBT-GER
procedure $RandomConstruction$ (L)
Initialize S;
$i \leftarrow 0;\ k \leftarrow 1;\ d \leftarrow 1;$
Generate candidate list cl *for day d;*
repeat
  Randomly select j from cl;
  repeat
    if (*vehicle* k *can be assigned to* j *at day* d) **then**
      $i \leftarrow i+1;\ S(d,k,i) \leftarrow j;$
    else
      $k \leftarrow k+1; i \leftarrow 1; S(d,k,i) \leftarrow j;$
    end if
    Decrease number of tours for storage pile j*;*
    if (*no tours to* j *left*) **then**
      $cl \leftarrow cl \backslash \{j\};$
    end if
  until ((*no tours to* j *left*) or (*j is not urgent*) or (*daily requirements are satisfied*))
  if (*daily requirements are satisfied*) **then**
    $i \leftarrow 0;\ k \leftarrow 1;\ d \leftarrow d+1;$
    Generate candidate list cl *for day* d**;**
  end if
until $(d > |D|)$
return S;

At the beginning, the *Initialize* procedure sets all entries of the matrix S to zero. Then, the index variables d, k, and i, which correspond to the dimensions of the matrix, are initialized, and the candidate list cl of storage piles available on the first day is generated. Within the main loop, a storage pile is selected at random from cl and the tours to that storage pile are assigned to the current vehicle if possible, otherwise the next vehicle and its first tour are considered. The selected storage pile is considered until it is emptied, or no longer urgent (equity constraints for its grower are satisfied) or the daily requirements are fulfilled. If the corresponding storage pile is emptied, it is removed from the candidate list and if requirements of the mill are satisfied, the next day is considered. When the plan for the last day of the planning period is completed, the procedure for solution construction ends returning the matrix S.

The LS step of the proposed MLS is the same as LS used and described in details in (Drenovac et al., 2025) for VNS and GRASP implementations. Therefore, its description is omitted in this paper.

4. EXPERIMENTAL EVALUATION

Experimental tests are performed on Intel Xeon CPU E5-2620 v3, 2.40 GHz with 32GB RAM memory, under Linux operating system. MLS is implemented in C programming language. The set of instances generated in (Drenovac et al., 2025) is used for the evaluation. Test instances differ in two characteristics: the mill's daily requirements and the value of θ (the minimum percentage of sugar beet that must be collected from each grower on the day of harvest). The first characteristic can take 4 values: 1000, 3000, 5000, and 7000

DRAFT VERSION

tons, while θ could be either 20% or 30%, giving 80 instances in total for our experiments. The parameter tuning tests are conducted on the subset of instances, yielding best performing value for LS parameter $niter = 1$, the same as in (Drenovac et al., 2025). CPLEX solver execution time is limited to 1h, while MLS and GRASP are tested 30 times on each instance with the time limit of 120s in each execution.

For simplicity, the computational results presented in Table 1, contain average values per group of instances. The name of each group of 10 instances is given in the first column. The CPLEX solver results, consisting of the average objective function value over all 10 instances (best) and the average execution time (t) in seconds, are given in the next two columns. The remaining columns refer to the results of the proposed MLS algorithm and GRASP from (Drenovac et al., 2025). As they are both stochastic, each instance is tested 30 times and the best obtained objective function value is recorded along with the average gap, agap (in percent), of the obtained solution in each run from the best obtained and the average execution time, atbest (in seconds), required for reaching the best solutions. These values are additionally averaged per group of instances, leading to the values in the columns denoted by: best, agap and atbest, for both metaheuristics. The best average objective function values are bolded in each group.

Table 1: Computational results and comparisons

Group of instances	CPLEX		MLS			GRASP		
	best	t(s)	best	agap(%)	atbest(s)	best	agap(%)	atbest(s)
$G_{1000,20\%}$	**0.830649**	1705.65	0.830623	0.002	51.81	0.823115	0.417	59.74
$G_{1000,30\%}$	**0.830638**	1308.59	0.830612	0.001	50.88	0.825609	0.413	61.82
$G_{3000,20\%}$	0.836094	2712.42	**0.838869**	0.105	51.86	0.833047	0.174	55.37
$G_{3000,30\%}$	0.836912	2653.47	**0.838461**	0.060	47.36	0.833873	0.176	55.70
$G_{5000,20\%}$	0.838291	3600.30	**0.843518**	0.162	57.85	0.839015	0.124	61.04
$G_{5000,30\%}$	0.840276	3600.30	**0.843506**	0.150	60.11	0.839010	0.117	56.77
$G_{7000,20\%}$	0.825986	3600.78	**0.835032**	0.062	57.84	0.830181	0.276	62.79
$G_{7000,30\%}$	0.826705	3600.41	**0.835208**	0.083	60.01	0.830364	0.272	60.57

As presented in Table 1, with the exception of the first two groups where exact solver is expected to dominate, MLS outperforms both CPLEX and GRASP with respect to average objective function value. In addition, its average time needed to provide the best solution and the average gap from the best are less than the corresponding GRASP values for almost all groups of instances, confirming its better stability and efficiency. The number of instances per group of 10 instances, on which MLS provides a better quality solution than GRASP are: 9, 9, 6, 6, 7, 7, 10, 10. On the remaining 16 instances, the two metaheuristics coincide in the case of two instances belonging to the group $G_{3000,20\%}$, and in the case of 14 out of 80 instances, GRASP obtains solution of slightly better quality. CPLEX provided 7 optimal solutions out of 10 instances on each of the two G_{1000} groups and 3 optimal solutions out of 10 instances on each of the two G_{3000} groups. The detailed CPLEX results can be found in Drenovac et al., 2025, while the detailed MLS results on all instances are available on request.

5. CONCLUSION

An efficient Multi-start Local Search (MSL) metaheuristic is proposed for the Sugar Beet Transportation Problem under Growers' Equity Regulations (SBT-GER). It uses the same Local Search procedure as the GRASP based algorithm, previously developed for the same problem. Due to the completely random construction of the initial solution in each MSL iteration, compared to the solution construction in GRASP, guided by a restricted candidate list, GRASP is expected to be more efficient than MLS. However, computational experiments on the same set of instances and the same computer platform, have shown superiority of MLS over GRASP, with respect to average solution quality, stability and time needed for achieving the best obtained solutions for majority of test examples, as well as in comparison with CPLEX exact solver, with exception of the small-sized instances. We believe it is a consequence of the problem characteristics that require higher level of randomized diversification of the search. The future work may include adaption of the proposed MLS for similar optimization problems in different supply chains or hybridization of the method with exact solver or other metaheuristics.

DRAFT VERSION

ACKNOWLEDGMENT

This work has been supported by the Serbian Ministry of Science, Technological Development and Innovation, Agreement Nos. 451-03-137/2025-03/200128, 451-03-137/2025-03/200097, and 451-03-136/2025-03/ 200029.

LITERATURE

Anokić, A., Stanimirović, Z., Stakić, Đ., & Davidović, T. (2021). Metaheuristic approaches to a vehicle scheduling problem in sugar beet transportation. *Operational Research, 21*(3), 2021–2053.

Asadi, M. (2007). *Beet-sugar handbook.* Hoboken, New Jersey: John Wiley and Sons, Inc.

Boender, C. G., Rinnooy Kan, A. H., Stougie, L., & Timmer, G. T. (1982). A Stochastic Method for Global Optimization. *Mathematical Programming, 22*, 125-140.

Brcanov, D., Dakić, S., Đokić, D., Gvozdenović, N., & Zekić, S. (2025). Optimization of transport activities in the sugar beet harvesting campaign. *Ekonomika Poljoprivrede, 72*(1), 155-169.

Corominas, A., García-Villoria, A., & Pastor, R. (2008). Solving the Response Time Variability Problem by means of Multi-start and GRASP metaheuristics. In *Artificial Intelligence Research and Development* (pp. 128-137). IOS Press.

Drenovac, D., Stakić, Đ., Anokić, A., Davidović, T., & Vidović, M. (2025). Sugar beet transportation problem under growers' equity regulations: Metaheuristic approach. *International Journal of Industrial Engineeering Computations*, 16, 1123-1142 doi: 10.5267/j.ijiec.2025.6.011

Drenovac, D., Vidović, M., & Bjelić, N. (2020). Optimization and simulation approach to optimal scheduling of deteriorating goods collection vehicles respectiong stochastic service and transport times. *Simulation Modelling Practice and Theory, 103*, 102097:1-23.

García, A., Pastor, R., & Corominas, A. (2006). Solving the Response Time Variability Problem by means of metaheuristics. In *Frontiers in Artificial Intelligence and Applications* (Vol. 146, pp. 187-194). Retrieved from https://ebooks.iospress.nl/volume/artificial-intelligence-research-and-development-2

Higgins, A. J. (1999). Optimizing cane supply decisions within a sugar mill region. *Journal of Scheduling, 2*(5), 229–244.

Jarumaneeroj, P., Laosareewatthanakul, N., & Akkerman, R. (2021). A multi-objective approach to sugarcane harvest planning in Thailand: Balancing output maximization, grower equity, and supply chain efficiency. *Computers & Industrial Engineering, 154*, 107129:1-13.

Jiao, Z., Higgins, A. J., & Prestwidge, D. B. (2005). An integrated statistical and optimisation approach to increasing sugar production within a mill region. *Electronics in Agriculture, 48*(2), 170–181.

López-Sánchez, A. D., Hernández-Díaz, A. G., Vigo, D., Caballero, R., & Molina, J. (2014). A multi-start algorithm for a balanced real-world Open Vehicle Routing Problem. *European Journal of Operational Research, 238*(1), 104-113.

Martí, R. (2003). Multi-start methods. In F. Glover, & G. A. Kochenberger, *Handbook of Metaheuristics* (pp. 355-368). Boston: Springer.

Matijević, L., Ilin, V., Davidović, T., Jakšić-Krüger, T., & Pardalos, P. M. (2024). General VNS for asymmetric vehicle routing problem with time and capacity constraints. *Computers & Operations Research, 167*, 106630.

Molina, J., López-Sánchez, A. D., Hernández-Díaz, A. G., & Martínez-Salazar, I. (2018). A Multi-start Algorithm with Intelligent Neighborhood Selection for solving multi-objective humanitarian vehicle routing problems. *Journal of Heuristics, 24*, 111-133.

Sethanan, K., Theerakulpisut, S., & Neungmatcha, W. (2014). Sugarcane Harvest Scheduling to Maximize Total Sugar Yield with Consideration of Equity in Quality Among the Growers. In P. Golinska, *Logistics Operations, Supply Chain Management and Sustainability* (pp. 341-352). Cham: Springer.

Thuankaewsing, S., Khamjan, S., Piewthongngam, K., & Pathumnakul, S. (2015). Harvest scheduling algorithm to equalize supplier benefits: A case study from the Thai sugar cane industry. *Computers and Electronics in Agriculture, 110*, 42–55.

Villegas, J. G., Prins, C., Prodhon, C., Medaglia, A. L., & Velasco, N. (2010). GRASP/VND and multi-start evolutionary local search for the single truck and trailer routing problem with satellite depots. *Engineering Applications of Artificial Intelligence, 23*(5), 780-794.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

VNS APPROACH FOR TOTAL TARDINESS MINIMIZATION IN A SINGLE MACHINE SCHEDULING PROBLEM WITH PERIODIC RESOURCE CONSTRAINTS

MAJA RISTIĆ[1], ZORICA DRAŽIĆ[2]

[1] University of Belgrade, Faculty of Mining and Geology, Belgrade, maja.ristic@rgf.bg.ac.rs, 0000-0002-3539-018X
[2] University of Belgrade, Faculty of Mathematics, Belgrade, zorica.drazci@matf.bg.ac.rs, 0000-0002-3434-6734

Abstract: *In this paper, we consider a recently introduced variant of a single machine scheduling problem with periodical resource constraints. The goal is to minimize the total tardiness of all jobs that need to be scheduled on a single machine, taking into account the time and resource consumption constraints per production period. Since the considered problem is NP-hard, we propose a metaheuristic approach using the Variable neighborhood search (VNS). The performance of the VNS method is evaluated on a set of test instances from the literature with up to 1000 jobs. The obtained results are compared with the results of other methods from the literature and show the efficiency of the proposed VNS approach over the other algorithms.*

Keywords: *Combinatorial optimization, Metaheuristics, Variable neighborhood search, Scheduling problems, Single machine, Resource consumption, Tardiness.*

1. INTRODUCTION

Scheduling problems have received considerable attention for many years. In the general case, the task is to schedule the execution of n jobs on m machines under certain conditions and constraints. These are real-life factors such as machine failure during operation, delay in starting machine operation, pre-emption, time constraints to be met, resource constraints (e.g. financial budget, material quantity), etc. In the case where $m = 1$, the task is reduced to scheduling jobs on a single machine, and several variants of this problem have been studied in the literature in recent decades (Bai et al., 2020; Cui and Lu, 2017; Karhi and Shabtay, 2018). Despite the simplicity of their formulation, these problems usually belong to the class of the most difficult (NP-hard) discrete optimization problems, even in the case of a single machine. The most common objectives of these problems are to minimize the total time for the execution of all jobs, to minimize the total time during which the machines are actively working, to minimize the total delay when the jobs are not executed on time, etc. Tardiness has a negative impact on profitability and therefore needs to be minimized. It leads to a loss of reputation with customers and reduces the possibility of attracting new customers. In addition, tardiness usually leads to late payments and penalties, resulting in additional operating costs. Some of the many scheduling problems that take tardiness into account can be found in (Chen, 2006; Gupta and Smith, 2006; Herr and Goel, 2016).

In this paper, we consider a variant of a single-machine scheduling problem with periodic resource constraints (SMPRC) with the objective of minimizing the total tardiness. In the classical SMPRC, there is a single machine environment with multiple production periods. Each period has a duration of fixed time units (e.g. working day or week) and in each period there is a fixed resource constraint that limits resource consumption. The goal is to minimize the makespan, i.e. the total time required to complete all jobs. The classical SMPRC was introduced in the paper (B. d. A. Prata et al., 2021), where the authors proved that the problem is NP-hard, described it as a mixed-integer linear programming problem (MILP) and presented the solutions of the problem obtained using 22 different methods. The Variable neighborhood search for solving the same problem was proposed in (Ristić and Dražić, 2021). In the paper (B. Prata et al., 2022), the authors presented the new variant of SMPRC with an objective function that minimizes the total tardiness. The authors developed a MILP model for the addressed variant of the problem and presented two matheuristics as solution methods: matheuristic Size Reduction with Simulated Annealing (SRSA) and matheuristic Relax-and-Fix with Variable Fixing Search (RFVFS).

Given the NP-hard nature and complexity of the problem, a metaheuristic approach is employed. In particular, we propose a Variable Neighborhood Search (VNS) method, selected for its simplicity and previously demonstrated success in solving related scheduling problems. The elements of the proposed VNS have been specifically adapted to address the structure and constraints of the considered problem variant, including

DRAFT VERSION

appropriate modifications of the neighborhood structures to ensure alignment with the problem's requirements. Furthermore, the computational results demonstrate that the proposed VNS outperforms existing methods on a variety of test instances from the literautre, confirming its effectiveness and practical applicability.

This paper is organized as follows. In the second chapter a description of the problem is given. The third chapter describes the Variable neighborhood search method proposed to solve this problem. The computational results obtained and their comparison with previous results from the literature are presented in Chapter 4. Chapter 5 contains conclusions and suggestions for further research.

2. PROBLEM DEFINITION

The SMPRC with total tardiness can be described as follows. At time zero, there are n jobs $J_1, \ldots, J_n$ that need to be processed on a single machine. Each job J_i, $i = 1, \ldots, n$ requires the processing time p_i, the amount of resources r_i and should ideally be completed before its due date d_i. The total available time is divided into limited time periods/blocks (e.g. weekdays) of length P. For each time block, there is a maximum amount of available resources R, and the resources that are not used in one block cannot be used in another. In general, P is not sufficient to process all jobs and it is assumed that $P \geq p_i$ and $R \geq r_i$ for all $i = 1, \ldots, n$. All jobs must be produced on a single machine, which cannot process more than one job at a time. Interruptions in the production process are not allowed, i.e. it is not allowed to start the production of a job in one time block and finish it in another. In this variant of a problem, the due dates d_i, $i = 1, ..., n$ are determined as the serial number (index) of a block period by which the job i should ideally be completed. If we denote with C_i the index of a time block in which a product i was produced, then its tardiness can be calculated as $T_i = \max\{0, C_i - d_i\}$. The goal is to schedule the production of all jobs to the time blocks in such a way that the total tardiness for all jobs $\sum_{i=1}^{n} T_i$ is minimized.

An illustrative example with 7 jobs, $p = \{4,4,3,3,2,2,1\}$, $r = \{4,3,2,3,3,1,1\}$, $d = \{3,3,1,1,2,2,1\}$, $P = 8$ and $R = 6$ is shown in Figure 1. For two different schedules $x_1 = \{1,3,2,4,5,6,7\}$ and $x_2 = \{3,4,7,2,5,6,1\}$ the Gantt chart diagrams for processing the jobs in these orders are shown. Figure 1a shows the optimal solution of SMPRC with the objective of minimizing the makespan. For a given job order, all jobs are scheduled in 3 blocks, and the time required to complete the entire process is 21 time units. Note that although J_7 requires only 1 time unit ($p_7 = 1$) and there is enough free time in the first two blocks, it cannot be scheduled there because both blocks have used up all the resources allocated to them. In this solution, the completion and tardiness vectors are given by $C = \{1,2,1,2,3,3,3\}$ and $T = \{0,0,0,1,1,1,2\}$, and thus the total tardiness for this schedule is 5. Figure 1b shows the optimal solution of SMPRC with the objective of minimizing the total tardiness. In this case, the vectors for completion time blocks and tardiness are $C = \{3,2,1,1,2,3,1\}$ and $T = \{0,0,0,0,0,1,0\}$. For this schedule, all jobs are also scheduled in 3 time blocks, the makespan is 22, but the objective of total tardiness is reduced to 1.

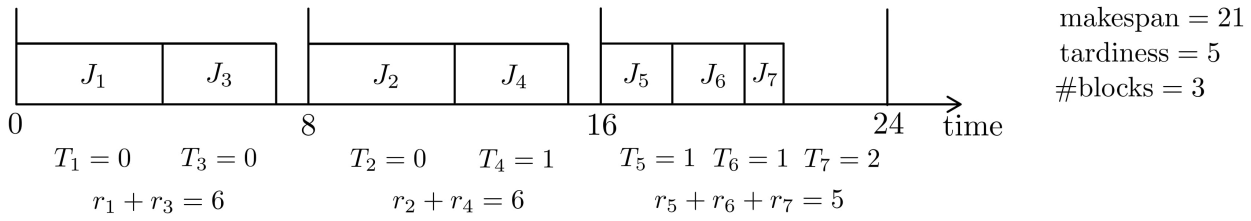
3. THE PROPOSED VARIABLE NEIGHBORHOOD SEARCH

The VNS metaheuristic originally proposed by Mladenovic and Hansen (Mladenović and Hansen, 1997) is based on the idea of systematically changing the neighborhood structures in order to escape from a possible local optimum. The basic variant of the VNS method consists of two phases that alternate together with the neighborhood change step until a stopping criterion is met. The first phase, called Shaking, is used to escape from local optimal solutions. The second phase is the Local search phase, in which some neighborhood of the solution obtained in the Shaking phase is explored in order to reach a local optimum. The VNS method has already been successfully used to solve many difficult optimization and scheduling problems. For example, VNS heuristics have been developed for scheduling problems with multiple machines (Adibi et al., 2010; Ahmadian et al., 2021; Wang and Li, 2014), as well as for single machine problems (Liao and Cheng, 2007; Pacheco et al., 2018).

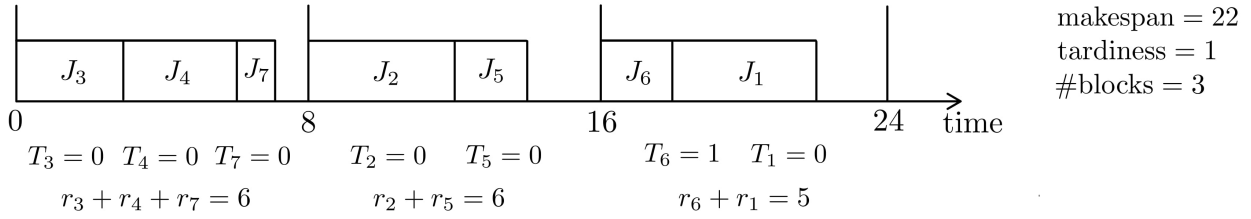
Algorithm 1 presents pseudocode of the proposed VNS algorithm for the SCPRC with minimizing tardiness. The solution of the problem is represented by an integer vector x of length n with unique integer values $1 \leq x_k \leq n$, $x_i \neq x_j$ which represent the order in which the jobs will be executed on the machine. The algorithm starts by generating an initial solution which is a random permutation of integers from 1 to n, representing the indices of the jobs.

Within a VNS main loop, the shaking procedure, the local search procedure and the neighborhood change step are applied alternately. For a given k, the neighborhood $N_k(x)$ contains all vectors with job indices that may differ from the current solution x in at most k index positions. The Shaking() function generates a new vector $x' \in N_k(x)$ following the next steps. First, a vector of k random integers from 1 to n is formed. This vector represents the indices of the solution vector x that are to be changed. The values at the selected positions are

DRAFT VERSION



(a) SMPRC with minimizing makespan



(b) SMPRC with minimizing total tardiness

Figure 1: Comparison of the optimal solutions for SMPRC with minimizing (a) makespan, (b) total tardiness

Algorithm 1 VNS Algorithm

```
 1: procedure VNS ALGORITHM(n, k_min, k_max, t_max, prob)
 2:     x ← RandomPermutation(1, n)
 3:     t ← CpuTime()
 4:     while t ≤ t_max do
 5:         k ← k_min
 6:         while k ≤ k_max do
 7:             x' ← Shaking(x, k)
 8:             x'' ← LocalSearch(x')
 9:             if ObjectiveFunction(x'') < ObjectiveFunction(x) then
10:                 x ← x''
11:             else
12:                 if ObjectiveFunction(x'') > ObjectiveFunction(x) then
13:                     k ← k + 1
14:                 else
15:                     if RandomProbability() < prob then
16:                         x ← x''
17:                     else
18:                         k ← k + 1
19:                     end if
20:                 end if
21:             end if
22:         end while
23:         t ← CpuTime()
24:     end while
25:     return x
26: end procedure
```

DRAFT VERSION

shifted cyclically by one position to the left, with the first value being shifted to the last position. This direction of rotation was chosen because it can be assumed that the tardiness of all but one of the selected jobs can be reduced when they are moved forwards. The values at the unselected positions remain unchanged. This is illustrated in Figure 2 using a small example.

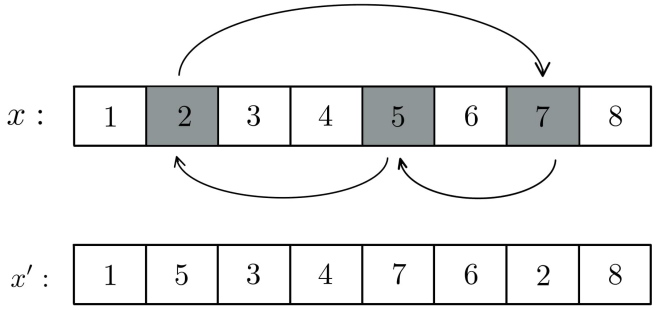


Figure 2: An example of shaking a solution vector of length 8, for $k = 3$

The LocalSearch() function explores the neighborhood of a solution x' in order to find a new solution with a smaller objective function value. The neighborhood to be explored consists of all solutions obtained from x' by swapping two of its elements. The first improvement strategy is used. The local search stops when the solution can no longer be improved and the best solution is denoted by x''.

Finally, it should be decided whether a new solution x'' should be accepted or not. If the local optimum x'' is better than the current solution x, the algorithm moves to this new solution and continues the search in the same neighborhood N_k. If the solution x'' is worse than x, the search is continued from the same x in the next neighborhood $N_{k+1}(x)$. Since the solution is represented as a permutation of jobs, there are various feasible sequences that lead to the same objective function value. For example, all jobs from the same block can change their position, leading to different solutions with the same objective function value. In such cases, where the problem has many local minima with the same objective function value, moving from one to another can expand the search and increase the chances of finding a better solution. On the other hand, there is a risk of being trapped in cycles if this step is performed every time. To avoid these problems, the solution x'' with the same objective function as x is accepted with a given probability *prob* and rejected with probability $1 - prob$. The VNS finishes its work once the maximum allowed CPU time, which is defined by the parameter $t_{\max}$, is reached.

4. EXPERIMENTAL STUDY

The VNS metaheuristic for solving the problem under consideration is implemented in the C programming language. The computational experiments were performed on a desktop computer with Intel Core i7-11700 3.6GHz processor and 16GB RAM in a 64bit Windows 10 environment. The experimental results were obtained on a set of 50 instances from the literature with up to 1000 jobs from (B. Prata et al., 2022). For each test instance, VNS was executed 10 times, each time starting from a different random initial solution. The VNS parameters were set to $k_{\min} = 2, k_{\max} = 10, t_{\max} = n$ and $prob = 0.4$.

In accordance with the experimental setup used in (B. Prata et al., 2022), the stopping criterion was defined as a fixed time limit of $t_{\max} = n$ seconds per run, where n is the number of jobs in the instance. This ensures that the algorithm runs proportionally to the problem size, enabling a meaningful comparison across different instance sizes. For example, an instance with 50 jobs was executed for 50 seconds, while an instance with 1000 jobs was allotted 1000 seconds.

The results of the proposed VNS were compared with the results of a MILP model and a hybrid matheuristic SRSA also presented in (B. Prata et al., 2022). Although the authors developed one more heuristic (RSVFS) in this paper, we did not have the solutions for RSVFS, so we only present the results of the comparison with the results of MILP and SRSA. The Relative Deviation Index (*RDI*) is used to compare the results obtained with these methods. Let H be the set of solution methods and T_{st} the best solution obtained with a method $s \in H$ for the instance t. RDI_{st} for the method $s \in H$ when solving the instance t is calculated as

$$RDI_{st} = \frac{T_{st} - \min\limits_{h \in H} T_{ht}}{\min\limits_{h \in H} T_{ht}}.$$

DRAFT VERSION

Table 1 shows for each of the compared methods $s \in \{MILP, SRSA, VNS\}$ the average values of *RDI* calculated over the set of instances with the same dimension $n \in \{50, 100, 250, 500, 1000\}$. Each instance group consists of 10 instances of the same dimension. The first column of the table contains the name of the method, while the following columns contain the average *RDI* values for all instances of the specified dimension in that set. The best values are in bold.

Table 1: Average values of RDI for MILP, SRSA and VNS

Method / Dimension	50	100	250	500	1000
MILP	0.65	3.24	3.94	9.97	12.74
SRSA	2.60	3.47	**0.12**	2.73	25.97
VNS	**0**	**0**	2.12	**0.54**	**0**

A comparison of the results presented in Table 1 shows that the VNS method achieves the best results according to this criterion in 4 out of 5 instance sets (test instances with 50, 100, 500 and 1000 jobs). In the case of the instance set with $n = 250$ VNS performs worse than SRSA, but better than MILP. Note that if the average RDI value in the table is 0, that means that the method under consideration has achieved the best known results for all instances from that group. This is the case for VNS and instance sets with 50, 100 and 1000 jobs.

Table 2 shows, for each set of 10 instances with the same dimensions, the total number of instances for which the MILP, SRSA and VNS methods achieved the best value $\min_{h \in H} T_{ht}$. The results are consistent with those of the *RDI* comparison. VNS achieves the highest number of best solutions for 50, 100, 500 and 1000 jobs, while in the case of $n = 250$ the SRSA method achieves the best solution for 9 out of 10 instances, VNS for only one instance and MILP for none.

Table 2: Total number of attained best known values for MILP, SRSA and VNS

Method / Dimension	50	100	250	500	1000
MILP	4	0	0	1	0
SRSA	1	0	**9**	4	0
VNS	**10**	**10**	1	**5**	**10**

Since the VNS method was executed 10 times for each instance, Table 3 shows the quality of the solutions obtained. For each set of 10 instances of the same dimension, the first row (labeled err_{avg}) shows the average value of the percentage relative error of the solutions obtained with the VNS method and the second row shows the average value of the standard deviation (σ_{avg}).

Table 3: Average values of computation time, percentual relative error and standard deviation for 10 executions of VNS

Parameter / Dimension	50	100	250	500	1000
err_{avg}	0.15	1.04	2.29	2.10	1.41
σ_{avg}	0.15	0.76	1.40	1.09	0.83

Only for the smallest instance set with 50 jobs VNS showed a constant performance and achieved the best solution for 9 out of 10 instances in all 10 runs. As the dimension of the instances increases, the number of solutions that are not the same (and best) in 10 runs increases for VNS. The worst performance is for instance set with 250 jobs, where err_{avg} and σ_{avg} have the highest values.

5. CONCLUSION

In this paper, we consider a variant of a single-machine scheduling problem with periodic resource constraints with the objective of minimizing the total tardiness. Since the considered problem is NP-hard, the Variable neighborhood search metaheuristic is proposed. The obtained VNS results were compared with the results of two methods (MILP, SRSA) from the literature. On average, the proposed VNS method achieves better results than other proposed methods from the literature in most cases. Future research is related to the introduction of new neighborhood, the acceleration of the local search procedure with additional adaptations of the components of the proposed method according to the characteristics of the problem under consideration, and the hybridization of the proposed method with other heuristics to solve this problem.

DRAFT VERSION

ACKNOWLEDGMENT

The research of the authors was partially funded by University of Belgrade, Faculty of Mathematics and Faculty of Mining and Geology, through the grant by the Ministry of Science, Technological Development, and Innovation of the Republic of Serbia (the contract no 451-03-136/2025-03/200104 and 451-03-136/2025-03/200126)

LITERATURE

Adibi, M. A., Zandieh, M., & Amiri, M. (2010). Multi-objective scheduling of dynamic job shop using variable neighborhood search. *Expert Systems with Applications*, *37*(1), 282–287.

Ahmadian, M. M., Salehipour, A., & Cheng, T. (2021). A meta-heuristic to solve the just-in-time job-shop scheduling problem. *European Journal of Operational Research*, *288*(1), 14–29.

Bai, D., Xue, H., Wang, L., Wu, C.-C., Lin, W.-C., & Abdulkadir, D. H. (2020). Effective algorithms for single-machine learning-effect scheduling to minimize completion-time-based criteria with release dates. *Expert Systems with Applications*, *156*, 113445.

Chen, J.-S. (2006). Optimization models for the machine scheduling problem with a single flexible maintenance activity. *Engineering Optimization*, *38*(1), 53–71.

Cui, W.-W., & Lu, Z. (2017). Minimizing the makespan on a single machine with flexible maintenances and jobs' release dates. *Computers & Operations Research*, *80*, 11–22.

Gupta, S. R., & Smith, J. S. (2006). Algorithms for single machine total tardiness scheduling with sequence dependent setups. *European Journal of Operational Research*, *175*(2), 722–739.

Herr, O., & Goel, A. (2016). Minimising total tardiness for a single machine scheduling problem with family setups and resource constraints. *European Journal of Operational Research*, *248*(1), 123–135.

Karhi, S., & Shabtay, D. (2018). Single machine scheduling to minimise resource consumption cost with a bound on scheduling plus due date assignment penalties. *International Journal of Production Research*, *56*(9), 3080–3096.

Liao, C.-J., & Cheng, C.-C. (2007). A variable neighborhood search for minimizing single machine weighted earliness and tardiness with common due date. *Computers & Industrial Engineering*, *52*(4), 404–413.

Mladenović, N., & Hansen, P. (1997). Variable neighborhood search. *Computers & Operations Research*, *24*(11), 1097–1100.

Pacheco, J., Porras, S., Casado, S., & Baruque, B. (2018). Variable neighborhood search with memory for a single-machine scheduling problem with periodic maintenance and sequence-dependent set-up times. *Knowledge-Based Systems*, *145*, 236–249.

Prata, B., de Abreu, L. R., Nagano, M. S., et al. (2022). Total tardiness minimization in a single-machine with periodical resource constraints. *Jurnal Optimasi Sistem Industri*, *21*(2), 97–105.

Prata, B. d. A., de Abreu, L. R., & Lima, J. Y. F. (2021). Heuristic methods for the single-machine scheduling problem with periodical resource constraints. *Top*, *29*(2), 524–546.

Ristić, M., & Dražić, Z. (2021). Solving the single-machine scheduling problem with periodical resource constraints. *Proceedings of the XLVIII International Symposium on Operations Research SYM-OP-IS 2021, Banja Koviljača, Serbia*.

Wang, S., & Li, Y. (2014). Variable neighbourhood search and mathematical programming for just-in-time job-shop scheduling problem. *Mathematical Problems in Engineering*, *2014*(1), 431325.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

VARIABLE NEIGHBORHOOD SEARCH FOR THE p-NEXT CENTER PROBLEM WITH DISCOUNT FACTOR

JELENA TASIĆ[1], ZORICA STANIMIROVIĆ[2], ZORICA DRAŽIĆ[3]

[1] University of Belgrade, Faculty of Mathematics, Belgrade, jelena.tasic@matf.bg.ac.rs, 0009-0002-0145-6985
[2] University of Belgrade, Faculty of Mathematics, Belgrade, zorica.stanimirovic@matf.bg.ac.rs, 0000-0001-5658-4111
[3] University of Belgrade, Faculty of Mathematics, Belgrade, zorica.drazci@matf.bg.ac.rs, 0000-0002-3434-6734

Abstract: *The p-next center problem (PNCP) is an extension of the well-known p-center problem, which captures the situation when one or more centers suddenly fail due to some technical problem. If a user is faced with the failure of its primary center, it is being redirected to its backup center - the center that is closest to the primary one. The goal of the PNCP is to minimize the maximum distance that a user must travel to its backup center via the primary center. In practice, the distance that the user travels is often expressed in travel cost or travel time, and the communication between the centers is cheaper or faster. In order to capture this real-world situation, we involve a discount factor for the travel cost or time between the centers in the objective function of the classical PNCP. A metaheuristic method based on variable neighborhood search is used as a solution method for the considered variant of the PNCP. The set of computational experiments on instances from the literature is performed to investigate the impact of the discount factor on the obtained solutions and the corresponding objective function values.*

Keywords: *Location analysis, p-next center problem, Discount factor, Technical failure, Emergency service network, Variable neighborhood search*

DRAFT VERSION

1. INTRODUCTION

The *p*-center problem (PCP) introduced in (Hakimi, 1965) is one of the most studied NP-hard combinatorial optimization problems in the literature, see (Biazaran and SeyediNezhad, 2009; Celik Turkoglu and Erol Genevois, 2020). Starting from the set of locations of users and potential centers, the goal of the PCP is to open exactly p centers in order to minimize the maximum distance a user travels to its nearest center. An extension of the PCP that covers situations where some of the centers suddenly fail is introduced in (Albareda-Sambola et al., 2015) and named the *p*-next center problem (PNCP). In the PNCP, the number of centers to be opened is set to p and each user is assigned to its closest opened center (primary center). In the case of a failure of its primary center, a user proceeds to an opened center that is closest to the primary one, denoted as backup center. It is assumed that a customer will pass through the primary center to reach its backup center. The goal of the PNCP is to find locations for opening exactly p centers such that the maximum distance from a user to its backup center via the primary center is minimized. The PNCP finds its application in the planning of emergency service networks (finding optimal locations of ambulances, hospitals, fire stations, police stations), but also in the optimizing customer service networks (determining optimal locations for ATMs in urban areas, opening gas stations of a gas distribution company, planning the network of supermarkets or drugstores of the same company, etc.).

There are several studies in the literature that deal with the PNCP. In (López-Sánchez et al., 2019), a Greedy Randomized Adaptive Procedure (GRASP), a Variable Neighborhood Search (VNS) and their hybridization (GRASP-VNS) were proposed to solve the PNCP. Computational experiments on instances with up to 200 nodes showed that the hybrid GRASP-VNS approach had the best performance. In the paper (Ristić et al., 2021), the authors presented the Filtered VNS method (FVNS), which was shown to be efficient in solving pmed instances with up to 200 nodes, and also the Additional instances with up to 900 nodes, introduced in the same study. In (Tasić, 2024), a Skewed VNS method (SVNS) was proposed for the PNCP, and it was shown that the SVNS reaches or improves the best FVNS solutions on both pmed and Additional data sets. Three variants of a Multi-Parent Biased Random-Key Genetic Algorithm with Implicit Path-Relinking (BRKGA) were introduced in (Londe et al., 2021) to solve the PNCP. The computational results on the set of 416 instances with up to 500 nodes showed the ability of BRKGAs to find excellent quality solutions, but with a significant running time consumption. A variant of the local search (LS) method with two strategies used to exploit the flat subspaces

was proposed in (Mousavi, 2023). Computational experiments on pmed instances with up to 200 nodes showed the advantages of the LS-based heuristic over the approaches from (López-Sánchez et al., 2019) in solving the PNCP. The study (Zhang et al., 2022) presented a weighting-based tabu search algorithm (WTS), which was tested on pmed-based instances with up to 900 nodes. The results of the comparative study indicate that WTS outperforms GRASP-VNS and the three BRKGA methods on the set of PNCP instances used in all three studies. A modified Basic VNS method proposed in (Ristić et al., 2023) was shown to be superior over all previous state-of-the-art PNCP methods in terms of solution quality on instances with up to 900 nodes. In addition, the authors introduced two new data sets of PNCP instances with up to 1000 and 2500 nodes, respectively.

In this study, we consider a variant of the PNCP that better reflects the situation in practice. Once the emergency or service centers are opened, efficient and reliable communication channels between them are also provided. Therefore, it is natural to assume that communication between the centers is more efficient (regarding travel time or transportation costs) than communication between a user and a center. For this reason, we involve parameter $\gamma \in (0,1)$ in the PNCP model, which represents the discount factor for the travel cost (or time) that a user travels from its primary to backup center. This problem is named the PNCP with discount factor and it belongs to the class of NP-hard problems, as a variant of PNCP that has been shown to be NP-hard (Albareda-Sambola et al., 2015). Due to the complexity of the problem under consideration, we use a Basic variant of the Variable Neighborhood Search metaheuristic (BVNS) as a solution approach. The choice of the solution method was motivated by previous successful VNS implementations for the PNCP (Ristić et al., 2021; Ristić et al., 2023; Tasić, 2024) and the necessity of having an efficient algorithm that is capable to solve instances of large dimensions. The elements of the proposed BVNS have been modified according to the characteristics of the problem considered. The BVNS uses Move Evaluation and Update procedures instead of the classical swap and/or update from scratch, which increases the efficiency of the algorithm.

The rest of the paper is structured as follows. In Section 2, mathematical formulation of the PNCP with discount factor is given. Section 3 explains the implementation steps of the BVNS metaheuristic for the problem under consideration. In Section 4, we present computational results obtained on pmed and Additional instances from the literature. The summary of the contributions and some future work directions are given in Section 5.

2. MATHEMATICAL FORMULATION

The p-next center problem with discount factor is defined on an undirected weighted graph $G = (A, E)$, where $A = \{1, 2, \ldots, n\}$ denotes the set of n nodes in the network, which represent both the locations of users and potential centers, and $E = \{(i, j) : i, j \in A, i \neq j\}$ is the set of edges. Each edge $(i, j) \in E$ has a weight d_{ij} corresponding to the travel cost or time along that edge. For each pair of nodes $i, j \in A$, $i \neq j$, the distance $d(i, j)$ is calculated as the length of the shortest path (in terms of travel cost or time) from i to j, and thus the distances satisfy the triangle inequality. The number of centers to be opened is set to p. Each user is assigned to its primary center, which is the closest center (in terms of travel cost or time), while the backup center is the closest center to its primary one. The parameter $\gamma \in (0,1)$ denotes the discount factor that multiplies the travel cost or time from a primary center to a backup center.

The goal of the problem is to find a set of centers $P \subset A$ with cardinality p such that the maximum travel cost (or time) from a user to its assigned backup center is minimized. The travel cost from a user to its allocated backup center is calculated as the sum of the travel cost from the user to its primary center and the travel cost from the primary center to its backup center, multiplied by the discount factor γ. More precisely, the objective is

$$\min_{P \subset A, |P| = p} f(P), \tag{1}$$

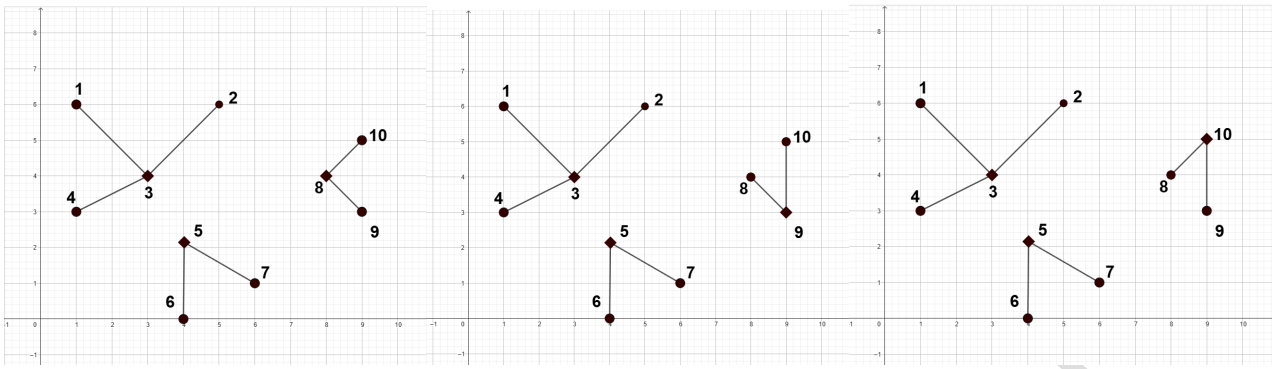
where

$$f(P) = \max_{i \in A} \{ \min_{j \in P} d(i, j) + \min_{j' \in argmin_{j \in P} d(i,j)} \min_{k \in P, k \neq j'} \gamma \cdot d(j', k) \}. \tag{2}$$

The formulation presented here is obtained from the PNCP formulation proposed in the paper (Albareda-Sambola et al., 2015) by including the discount factor γ in the second part of the function $f(P)$. In (Albareda-Sambola et al., 2015), the authors gave a formal proof that the PNCP is NP-hard, and therefore, the considered PNCP variant with discount factor is also NP-hard.

In order to graphically illustrate the effects of using the discount factor γ, let us consider a small example with $n = 10$ locations given by their coordinates in the plane and $p = 3$ centers to be opened. The opened centers are marked with ■, the users are marked with •, while the assignments of the customers to the centers are denoted with black lines (see Figure 1). In the optimal solution of the PNCP, placing centers at locations 3, 5, and 8 results in objective function value of 5.89, which differs from the solution obtained with a discount factor of $\gamma = 0.1$, where centers are placed at locations 3, 5, and 9, and the corresponding objective value is 3.05. When the discount factor is slightly increased to $\gamma = 0.2$, the resulting optimal solution again differs from both previous cases, with centers at locations 3, 5, and 10, and an objective function value of 3.28.

DRAFT VERSION



(a) PNCP **(b)** PNCP with $\gamma = 0.1$ **(c)** PNCP with $\gamma = 0.2$

Figure 1: Comparison of the optimal solutions for (a) PNCP, (b) PNCP with $\gamma = 0.1$ (c) PNCP with $\gamma = 0.2$ on a small-size network

3. THE PROPOSED BASIC VARIABLE NEIGHBORHOOD SEARCH

Variable Neighborhood Search is a metaheuristic introduced by Mladenović and Hansen (Mladenović and Hansen, 1997), based on the idea of systematic changes of neighborhood structures during the search. Up to now, many variants of VNS metaheuristics have been proposed in the literature and applied to a wide range of optimization problems. Examples of successful applications of VNS-based methods to the p-center problem and its variants can be found in the literature. For example, VNS heuristics have been developed for solving the p-center problem in (Mladenović et al., 2003), the probabilistic p-center problem in (Martínez-Merino et al., 2017), the unconditional and conditional vertex p-center problems in (Irawan et al., 2016), the capacitated vertex p-center problem in (Quevedo-Orozco and Ríos-Mercado, 2015), the p-next center problem in (López-Sánchez et al., 2019; Ristić et al., 2021; Ristić et al., 2023; Tasić, 2024), the conditional p-next center problem in (Tasić et al., 2025), etc.

To solve the PNCP with discount factor, we adapt the concept of the Basic Variable Neighborhood Search (BVNS) method described in Algorithm 1. The BVNS method starts from a randomly generated solution and the set of predefined neighborhood structures $N_k, k = 1, \ldots, k_{max}$. The solution of the problem is represented by an integer vector x of length p containing the indices of the locations with established centers. The solution is feasible if the corresponding p-dimensional vector contains indices that are mutually different. The indices of the locations with established centers are obtained from the integer vector x and each user is assigned to its nearest center (the primary one), while the user's backup center is determined as the one closest to its primary center. Once the primary and backup centers are known for each user, the objective function value $f(x)$ can be easily calculated. In the proposed BVNS, the neighborhood N_k, $k = 1, \ldots, k_{max}$ of a solution x consists of all feasible solutions x' satisfying $d(x,x') = k$, where $d(x,x')$ is defined as the total number of distinct locations chosen as centers for the solutions x and x'.

An iteration of BVNS consists of two main phases - Shaking and Local Search, as well as the decision about the neighborhood change (Move or Not step). In the Shaking phase, a feasible solution x' is randomly selected from the current neighborhood $N_k(x)$. More precisely, k locations that do not belong to the solution x are randomly selected to replace k randomly selected centers from x, and thus a neighbor $x' \in N_k(x)$ is obtained. The procedure Update is executed to find the primary and backup centers for each user in the new solution x' and to calculate its objective function value efficiently. During the Local Search phase, which is based on the Move evaluation and Update procedures, the neighborhood $N_1(x')$ is explored to reach a local optimum x''. For a given $i \notin x'$, the Move evaluation procedure finds a center j in the current solution x' to be exchanged with i such that the value of $f((x' \cup \{i\}) \setminus \{j\})$ is minimal. Swapping a specific pair of locations (i, j') can affect users in three different ways: (a) users may remain unaffected, (b) users may start using the new center i as their primary or backup center (if it is closer than their current primary or backup center), and (c) users may lose their existing primary or backup center and need to be reallocated. Taking all three scenarios into account, the potential swap of element $i \notin x$ with each element $j' \in x$ can be evaluated efficiently using specific auxiliary structures, allowing the best candidate $j \in x$ for elimination to be identified. This is achieved by considering each location only once and examining possible outcomes for that location: adopting i as the primary or backup center or changing its current primary and/or backup center. In this way, a complexity of $O(n)$ can be achieved for the Move Evaluation procedure.

The Local Search procedure examines each potential center $i \notin x'$ and applies the Move evaluation procedure to find the best candidate $j \in x'$ to be replaced by i. Once the best exchange pair is found, the procedure Update

DRAFT VERSION

performs the swap that leads to the improved solution, updates the set of primary and backup centers for each user and calculates the objective function value of the improved solution. The steps are repeated as long as an improvement can be found, and the output of the Local Search phase is the local optimum x''.

If the local optimum x'' obtained from the Local Search phase improves the objective function value of the current solution x, the algorithm moves on to this new solution and continues the search from there. Otherwise, the order of neighborhood to be explored is increased by 1 and the algorithm moves on to the Shaking phase. The algorithm is executed until the stopping criterion (the maximal number of iterations) is met.

Algorithm 1 BVNS algorithm

Initialization: select the set of neighborhoods N_k, $k = 1,2,...,k_{max}$
Randomly generate a feasible initial solution x and set $f \leftarrow f(x)$
while the stopping criterion is not met **do**
    $k \leftarrow 1$
    while $k \leq k_{max}$ and the stopping criterion is not met **do**
        $x' \leftarrow Shaking(x,k)$ ▷ Shaking phase
        $x'' \leftarrow Local_Search(x')$ ▷ Local Search phase
        if $f(x'') < f$ **then** ▷ Move or Not step
            $x \leftarrow x''$, $f \leftarrow f(x'')$, $k \leftarrow 1$
        else
            $k \leftarrow k+1$
        end if
    end while
end while

4. EXPERIMENTAL STUDY

The BVNS metaheuristic for solving the p-next center problem with discount factor is implemented in the C++ programming language. The computational experiments were performed on a desktop computer with Intel Core i7-11700 3.6GHz processor and 16GB RAM in a 64bit Windows 10 environment. Experimental results were obtained on a set of 50 small-size pmed instances from (Beasley, 1990) with up to 50 vertices and a set of Additional instances with up to 900 vertices from (Ristić et al., 2021). Values of 0.1, 0.4 and 0.6 were used for the parameter γ, representing different values of the discount factor. The proposed BVNS was executed 20 times for each test instance, each time starting from a different random initial solution. The maximum number of BVNS iterations was used as the stopping criterion: 500 for the instances with $n \leq 50$ and 5000 for instances with $100 \leq n \leq 900$. For all instances, the value of parameter $k_{\max}$ was set to 5. The quality of the results obtained with the proposed BVNS method was evaluated by comparing them with the optimal and suboptimal solutions obtained with the exact solver CPLEX 22.1.0, which uses the mathematical formulation from Section 2. The execution time of the CPLEX solver was limited to 2 hours for each tested instance.

Table 1: The summary of results obtained by CPLEX and BVNS on pmed ($n \leq 50$) and Additional ($100 \leq n \leq 900$) instances

Inst	***n***	**#Inst**	**γ**	**CPLEX**			**BVNS**	
				#Opt	**#Feas**	**avg.t(s)**	**%bk**	**avg.t(s)**
pmed	$n \leq 50$	50	0.1	50	-	56.35	100%	0.04
Additional	$100 \leq n \leq 900$	40	0.1	0	16	7200	100%	17.45
pmed	$n \leq 50$	50	0.4	50	-	18.90	100%	0.03
Additional	$100 \leq n \leq 900$	40	0.4	2	16	6602	100%	12.67
pmed	$n \leq 50$	50	0.6	50	-	11.73	98%	0.03
Additional	$100 \leq n \leq 900$	40	0.6	3	16	6170.63	100%	79.48

Table 1 contains an overview of the results of the CPLEX solver and the proposed BVNS. The columns Inst, n, #Inst, and γ contain the name of the data set, the dimension of instances belonging to this data set, the total number of tested instances, and the value of the discount factor. For each different γ, the number of optimal solutions (#Opt) and feasible solutions (#Feas) obtained by CPLEX are given. For the BVNS, we present the percentage of success in reaching the best known solution (%bk), i.e., the optimal solution for instances solved to optimality by CPLEX and the best solution obtained by either CPLEX or BVNS for the remaining instances. The columns labeled avg.t contain the average total execution time (in seconds) for CPLEX solver and BVNS

DRAFT VERSION

method, respectively. Note that the avg.t values are calculated only for instances where CPLEX reached either the optimal or the feasible solution. From the data presented in Table 1, it can bee seen that CPLEX successfully reaches all optimal solutions for small-size pmed instances. However, in the case of the Additional set with larger instance dimensions, CPLEX could not even reach a feasible solution for more than half of the tested instances due to time or memory limits (out of memory). On the other hand, BVNS reaches all optimal solutions obtained by CPLEX, except for one instance (*pmed*3_50_20 in case when $\gamma = 0.6$), and outperforms feasible CPLEX solutions in most cases in significantly less computing time.

Table 2: The detailed results obtained by BVNS on pmed ($n \leq 50$) and Additional ($100 \leq n \leq 900$) instances

n	**#Inst**	**UB**	**BVNS**								
			$\gamma = 0.1$			**$\gamma = 0.4$**			**$\gamma = 0.6$**		
			avg.f	**avg.t**	**sav(%)**	**avg.f**	**avg.t**	**sav(%)**	**avg.f**	**avg.t**	**sav(%)**
10	5	106.6	50.68	<0.01	53.99	72.88	<0.01	33.99	85.84	<0.01	21.87
20	10	117.7	68.46	0.01	42.73	89.04	0.01	25.10	99.36	0.01	16.27
30	10	134.9	85.96	0.03	36.20	102.5	0.02	23.99	115.06	0.02	14.77
40	15	121.27	79.13	0.04	36.22	95.39	0.04	22.62	105.2	0.04	14.16
50	10	107.7	70.35	0.07	34.91	85.02	0.06	21.49	92.26	0.06	14.60
100	5	131	94	2.02	29.61	107.12	1.78	19.70	114.60	1.73	14.26
200	5	82.60	55.38	8.10	35.35	63.64	7.10	25.33	69.36	6.09	17.66
300	5	58.60	39.70	19.10	33.96	44.52	16.29	25.44	47.20	15.46	20.68
400	5	44.40	31.08	34.69	31.64	35.28	29.09	21.42	37.28	29.87	16.74
500	5	41	27.62	56.53	33.39	31.32	48.35	24.13	34.24	46.83	16.93
600	5	43.60	24.40	84.31	43.17	27.68	72.31	35.29	32.06	67.26	26.13
700	4	42.50	23.28	66.30	38.74	25.95	59.69	31.11	31.05	55.38	20.18
800	3	37	25.47	47.35	31.54	28.13	43.15	24.19	29.60	46.35	20.10
900	3	45.67	23.40	67.36	40.79	26.73	56.02	33.04	32.53	53.45	23.23
Total					**37.32**			**25.20**			**17.23**

Table 2 shows the results obtained by the proposed BVNS metaheuristic for all considered instances. The first two columns of Table 2 contain the dimension of the instance group - *n* and the total number of instances in this group - #Inst. In the next column, denoted as UB, we present the average objective value of the best known PNCP solutions on the same group of instances. Note that the objective values of the PNCP problem are obtained for $\gamma = 1$ and represent the upper bound of the objective function values for the PNCP with discount factors $\gamma < 1$. For each different γ value, the data related to the BVNS solutions is presented in the following columns: avg.f - the average value of the objective function calculated over all instances in the considered group (for each instance in the group, the best solution value across 20 BVNS executions was used for calculating avg.f), avg.t - the average total execution time (given in seconds) and avg.sav - the average saving achieved with the given γ compared to UB (given in percentages). The results presented in Table 2 show that the proposed BVNS algorithm provides high quality solutions in short execution times considering the dimensions of the tested instances. From the last row of the Table 2, it can be concluded that the average total savings are 17.23% if the efficiency of communication between the centers was increased by 40% ($\gamma = 0.6$ instead of $\gamma = 1$). In the case of $\gamma = 0.4$ and $\gamma = 0.1$, the savings are 25.20% and 37.32%, respectively.

5. CONCLUSION

This paper proposes the PNCP with discount factor, which represents a variant of the well-known PNCP. The discount factor $\gamma \in (0,1)$ applies to the travel cost or time between the centers in the objective function. Due to the NP-hardness of the considered problem, the BVNS metaheuristic is implemented as solution method. The solution encoding, neighborhood structures, and other elements of the proposed BVNS are adapted to the problem's characteristics. Experimental study was conducted on pmed and Additional instances from the literature, and the obtained BVNS results were compared with the results of the exact CPLEX solver that uses the mathematical formulation of the problem. As it was expected, the CPLEX does not provide even feasible solutions for larger dimensions of the problem ($n > 400$). On the other hand, BVNS achieves all optimal solutions obtained by CPLEX (with exception of one case) in significantly less execution time than CPLEX, and the same or better solutions in the cases when CPLEX returns only feasible solutions. For large instances, which are out of reach of CPLEX due to limited time or memory resources, the implemented BVNS provides solutions of good quality in a relatively short execution time, considering the dimensions of the instances.

DRAFT VERSION

Further research may include the implementation of new metaheuristic algorithms, the extension of the model by adding new constraints, testing instances with larger dimensions, and additional adjustments of the components of the metaheuristic methods according to the characteristics of the considered problem.

ACKNOWLEDGMENT

The research of the authors was partially funded by University of Belgrade, Faculty of Mathematics through the grant by the Ministry of Science, Technological Development, and Innovation of the Republic of Serbia (the contract no 451-03-136/2025-03/ 200104)

LITERATURE

Albareda-Sambola, M., Hinojosa, Y., Marín, A., & Puerto, J. (2015). When centers can fail: A close second opportunity. *Computers & Operations Research*, *62*, 145–156. https://doi.org/10.1016/j.cor.2015.01.002

Beasley, J. E. (1990). OR-library: Distributing test problems by electronic mail. *Journal of the Operational Research Society*, *41*(11), 1069–1072. https://doi.org/10.1057/jors.1990.166

Biazaran, M., & SeyediNezhad, B. (2009). Facility location: Concepts, models, algorithms and case studies. In R. Z. Farahani & M. Hekmatfar (Eds.). Physica Heidelberg.

Celik Turkoglu, D., & Erol Genevois, M. (2020). A comparative survey of service facility location problems. *Annals of Operations Research*, *292*, 399–468. https://doi.org/0.1007/s10479-019-03385-x

Hakimi, S. L. (1965). Optimum distribution of switching centers in a communication network and some related graph theoretic problem. *Operations Research*, *13*(3), 462–475. https://doi.org/10.1287/opre.13.3.462

Irawan, C. A., Salhi, S., & Drezner, Z. (2016). Hybrid meta-heuristics with vns and exact methods: Application to large unconditional and conditional vertex p-centre problems. *Journal of Heuristics*, *22*(4), 507–537. https://doi.org/10.1007/s10732-014-9277-7

Londe, M. A., Andrade, C. E., & Pessoa, L. S. (2021). An evolutionary approach for the p-next center problem. *Expert Systems with Applications*, *175*, 114728. https://doi.org/10.1016/j.eswa.2021.114728

López-Sánchez, A. D., Sánchez-Oro, J., & Hernández-Díaz, A. G. (2019). GRASP and VNS for solving the p-next center problem. *Computers & Operations Research*, *104*, 295–303. https://doi.org/10.1016/j.cor.2018.12.017

Martínez-Merino, L. I., Albareda-Sambola, M., & Rodríguez-Chía, A. M. (2017). The probabilistic p-center problem: Planning service for potential customers. *European Journal of Operational Research*, *262*(2), 509–520. https://doi.org/10.1016/j.ejor.2017.03.043

Mladenović, N., & Hansen, P. (1997). Variable neighborhood search. *Computers & Operations Research*, *24*(11), 1097–1100. https://doi.org/10.1016/S0305-0548(97)00031-2

Mladenović, N., Labbé, M., & Hansen, P. (2003). Solving the p-center problem with tabu search and variable neighborhood search. *Networks: An International Journal*, *42*(1), 48–64. https://doi.org/10.1002/net.10081

Mousavi, S. R. (2023). Exploiting flat subspaces in local search for p-center problem and two fault-tolerant variants. *Computers & Operations Research*, *149*, 106023. https://doi.org/10.1016/j.cor.2022.106023

Quevedo-Orozco, D. R., & Ríos-Mercado, R. Z. (2015). Improving the quality of heuristic solutions for the capacitated vertex p-center problem through iterated greedy local search with variable neighborhood descent. *Computers & Operations Research*, *62*, 133–144. https://doi.org/10.1016/j.cor.2014.12.013

Ristić, D., Mladenovic, N., Todosijević, R., & Urošević, D. (2021). Filtered variable neighborhood search method for the p-next center problem. *International Journal for Traffic and Transport Engineering*, *11*(2), 294–309. https://doi.org/10.7708/ijtte.2021.11(2).09

Ristić, D., Urošević, D., Mladenović, N., & Todosijević, R. (2023). Solving the p-second center problem with variable neighborhood search. *Computer Science and Information Systems*, *20*(1), 95–115. https://doi.org/10.2298/CSIS210804049R

Tasić, J. (2024). An efficient solution approach to the p-next center problem. *Matematički Vesnik*, *76*(1), 66–83. https://doi.org/10.57016/MV-RBJS4362

Tasić, J., Dražić, Z., & Stanimirović, Z. (2025). A VNS method for the conditional p-next center problem. *Computers & Operations Research*, *175*, 106916. https://doi.org/10.1016/j.cor.2024.106916

Zhang, Q., Su, Z., Lü, Z., & Yang, L. (2022). A weighting-based tabu search algorithm for the p-next center problem. In L. De Raed (Ed.), *Proceedings of the thirty-first international joint conference on artificial intelligence (ijcai-22)* (pp. 4828–4834). International Joint Conferences on Artificial Intelligence.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

Comparative Study of a Basic and Modified Memetic Algorithm for the Electric Vehicle Routing Problem

LUKA MATIJEVIĆ[1]

[1] Mathematical Institute of the Serbian Academy of Sciences and Arts, luka@mi.sanu.ac.rs, 0000-0002-4575-6720

Abstract: *With increasing environmental concerns and the growing adoption of electric vehicles, green logistics has become a key area of research. This paper addresses the Electric Vehicle Routing Problem with time-dependent travel speeds, soft time windows, and partial recharging. The objective is to minimize total travel distance while penalizing deviations from customer time windows. We develop and evaluate two variants of a Memetic Algorithm, testing them on benchmark instances and presenting a comparative analysis of the results.*

Keywords: *GVRP, Green logistics, Metaheuristics*

1. INTRODUCTION

The *Vehicle Routing Problem* (**VRP**) is a well-studied combinatorial optimization problem, with numerous books and surveys dedicated to it (Golden et al., 2008; Toth and Vigo, 2014). It can be viewed as a generalization of the *Traveling Salesman Problem*, extended to multiple salesmen (vehicles). A notable variant of VRP, which focuses on routing electric vehicles, is known as the *Electric Vehicle Routing Problem* (**EVRP**). This variant was first introduced by (Erdoğan and Miller-Hooks, 2012) and has gained increasing attention in recent years due to growing environmental concerns and the push for sustainable transportation.

Unlike *internal combustion vehicles* (**ICV**s), *electric vehicles* (**EV**s) must visit *alternative fueling stations* (**AFS**s) to recharge their batteries. Moreover, while ICVs can refuel quickly, EVs require a longer time to recharge. This charging time must be explicitly considered when planning delivery routes, adding complexity to the problem.

This problem is computationally challenging to solve to optimality, as demonstrated in (Matijević, 2023a). For this reason, metaheuristic methods are commonly employed. A wide variety of metaheuristics have been proposed for solving similar problems—a comprehensive list can be found at https://www.mi.sanu.ac.rs/~luka/resources/phd/AppendixA.pdf.

In this paper, however, we focus primarily on the *Memetic Algorithm* (**MA**) and propose a slight modification to improve its performance on the EVRP. While the modification is tailored to this context, it is not entirely novel; we have previously applied it to enhance the training of neural networks (Matijević, 2023b).

The remainder of this paper is organized as follows. Section 2 defines the specific variant of the EVRP considered in this study. In Section 3, we describe the standard MA as well as our *Modified Memetic Algorithm* (**MMA**). Section 4 presents the computational results, and Section 5 concludes the paper.

2. PROBLEM DEFINITION

The problem addressed in this paper is an extension of the one studied in (Matijević, 2023a). It involves a delivery company operating a fleet of K identical EVs, all with the same speed, capacity, and energy consumption. Each day, the company receives a set of delivery requests from customers and must fulfill them as efficiently as possible.

Customers specify both the quantity of goods they require and a preferred time window for delivery. The company must then plan vehicle routes that respect these constraints, ensuring that the load on each vehicle does not exceed its capacity. Due to the limited range of EVs, vehicles may need to stop at AFSs during their routes. Unlike internal combustion vehicles, EVs take longer to recharge, so this factor must be carefully considered when scheduling arrival times at customer locations.

In the version of the problem considered here, partial recharging is allowed. This introduces additional complexity, as the algorithm must decide how much energy to recharge during each stop at an AFS.

Customer satisfaction is closely tied to adherence to the specified time windows. Delivering outside of this window, whether too early or too late, may negatively affect customer satisfaction and risk future business.

DRAFT VERSION

Arriving late may inconvenience the customer, while arriving too early could mean they are not prepared to receive the delivery. We consider *soft time windows*, meaning that vehicles are allowed to arrive outside the preferred delivery interval, but such violations incur a penalty in the objective function. To quantify these violations, we define a penalty function based on the arrival time τ_i at each customer $i \in V$. Early arrivals before the earliest allowable time e_i incur a penalty proportional to $max(0, e_i - \tau_i)$, while late arrivals after the latest allowable time l_i incur a penalty proportional to $max(0, \tau_i - l_i)$. The total penalty is calculated as a weighted sum over all customers, where α and β are user-defined parameters that control the relative importance of early versus late arrivals (in our case, both parameters were set to 1). This is expressed as:

$$\sum_{i \in V} p_i = \sum_{i \in V} [\alpha \cdot \max(0, e_i - \tau_i) + \beta \cdot \max(0, \tau_i - l_i)] \tag{1}$$

We also assume that all AFSs offer the same recharging rate and have unlimited capacity. Finally, this problem incorporates *time-dependent* vehicle routing, meaning that average vehicle speeds vary throughout the day, particularly relevant in urban environments where traffic conditions fluctuate significantly during rush hours.

3. MEMETIC ALGORITHM

The Memetic Algorithm is a variant of the Genetic Algorithm, originally proposed by (Moscato et al., 1989). Its distinguishing feature is the integration of a local search (or refinement) step within the evolutionary process. This local search can be interpreted as an individual learning mechanism or as the cultural transmission of knowledge. Each individual in the population undergoes a learning phase aimed at improving its quality.

In our implementation, solutions are encoded as chromosomes, as illustrated in Figure 1. For the crossover operator, we employed the *Order Crossover* (**OX**) (Eshelman and Schaffer, 1985), a method specifically designed for permutation-based problems in genetic algorithms. The core idea of OX is to preserve the relative ordering of elements from the parent chromosomes while ensuring that the resulting offspring represent valid permutations.

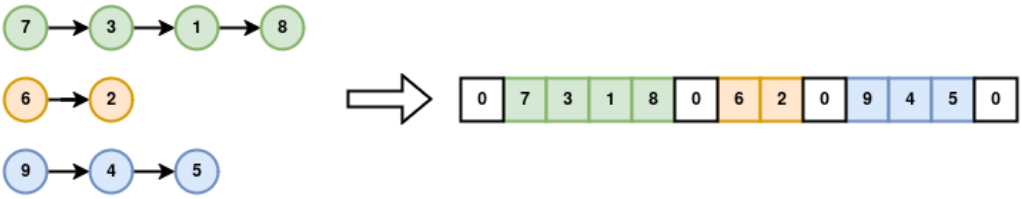


Figure 1: Chromosome representation of routes

For the mutation operator, we introduce controlled diversification by applying a random number of modification actions to the solution. Specifically, an integer is randomly selected from the range $[1, 3]$, indicating how many mutation actions will be performed. Each action is independently chosen at random from a predefined set of five mutation types designed to modify customer assignments and route structures. The available mutation operations are as follows: $M_1(S)$ — relocate a randomly selected customer to a different position within the same route; $M_2(S)$ — move a randomly selected customer to a different route and insert them at a random position; $M_3(S)$ — swap the positions of two randomly selected customers within the same route; $M_4(S)$ — swap two customers selected from different routes; and $M_5(S)$ — transfer a customer to an empty route, if one exists and the maximum number of routes has not been exceeded.

As the local search procedure, we employed *Variable Neighborhood Descent* (**VND**). The set of neighborhoods used within VND corresponds to those described in (Matijević, 2023a). Additionally, the methods for determining departure times from each node are also adopted from (Matijević, 2023a). We encourage the reader to consult that paper for further details on these methods.

The pseudocode for the our version of the MA is provided in Algorithm 1, where the population size (N), number of offspring (NO), mutation rate (MR), and time limit for local search (VND_{limit}) are given as input parameters. This approach slightly deviates from the standard formulation of the MA, which typically relies on a crossover rate as a hyperparameter rather than specifying the number of offspring directly. For more detailed information on MA, the reader is referred to (Neri et al., 2011).

DRAFT VERSION

Algorithm 1 Memetic Algorithm (MA)

```
1: procedure MA(N, NO, MR, VND_limit)
2:     Initialize population of N individuals
3:     Evaluate fitness of each individual
4:     while stopping criterion not met do
5:         offspring_set ← ∅
6:         for i = 1 to NO/2 do
7:             Select two parents using Roulette Wheel selection
8:             Generate two offspring via crossover
9:             for each offspring do
10:                Mutate with probability MR
11:                Create the departure schedule
12:                Apply local search with time limit VND_limit
13:                Add to offspring_set
14:            end for
15:        end for
16:        population ← population ∪ offspring_set
17:        Evaluate fitness of all individuals in the population
18:        Select the top N individuals to form the new population
19:    end while
20:    return the best individual from the population
21: end procedure
```

The algorithm begins by initializing a population of N individuals (line 2), followed by evaluating their fitness (line 3). The main loop (lines 4–19) is executed until a stopping criterion is met (time limit in our case).

In each generation, an empty set is created to store offspring (line 5). Then, $\frac{NO}{2}$ crossover operations are performed (lines 6-15). For each iteration, two parents are selected using *Roulette Wheel selection* (line 7), and two offspring are generated via crossover (line 8). Each offspring is then subjected to mutation with probability MR (line 10), after which the departure schedule is determined to minimize penalties from violating time windows (line 11). Finally, the solution is improved using VND with a time limit VND_{limit} (line 12). The refined offspring are added to the offspring set (line 13).

After generating all offspring, they are merged with the current population (line 16), and the fitness of all individuals is re-evaluated (line 17). The next generation is formed by selecting the top N individuals (line 18). Once the stopping criterion is satisfied, the algorithm returns the best individual found (line 20).

Our *Modified Memetic Algorithm* introduces a key difference in the way local search is applied, while retaining the overall structure of the original method presented in Algorithm 1. As in the original version, the algorithm begins by initializing a population of N individuals and evaluating their fitness. In each generation, new solutions are generated through selection, crossover, and mutation, and stored in a temporary offspring set. However, unlike the original approach, where local search is applied to every offspring individually, the modified version applies local search only to the best offspring in the current generation. By focusing local search on the most promising solution, the algorithm intensifies the search in high-quality regions of the solution space without sacrificing too much time.

4. RESULTS

The experiments were conducted on a personal laptop equipped with an *Intel i7-10750H* processor and 32 GB of RAM, running *Ubuntu 20.04*. All proposed algorithms were implemented in C++ and compiled using *GCC 9.4.0* with the O3 optimization flag to enhance performance. To ensure the reliability of our results, each test was repeated 30 times, demonstrating consistent outcomes.

We tested our methods on benchmark instances introduced in (Schneider et al., 2014), which are divided into two groups. The first group, referred to as I_1, consists of 35 instances with up to 15 customers each. The second group, I_2, contains 56 larger instances, each with 100 customers.

To model time-dependent travel speeds, we examined a scenario involving a goods transportation company based in Belgrade, which operates from 7 AM to 8 PM, totaling 13 hours of daily operation. We divided this delivery period into 13 equal subintervals, each assigned a distinct speed multiplier inversely related to traffic volume, as described earlier. Specifically, the first two intervals have a multiplier of 0.75, reflecting heavier traffic; the next eight intervals have a multiplier of 1, representing normal traffic conditions; the subsequent two

DRAFT VERSION

Table 1: The best-found values for each MA and MMA hyperparameter.

	Hyperparameters			
	N	NO	MR	VND_{LIMIT}
Possible values	[5, 40]	[2, 20]	{0.1, 0.2, 0.3}	[1, 9]
			MA	
I_1	25	17	0.3	4
I_2	6	3	0.1	7
			MMA	
I_1	18	17	0.3	3
I_2	5	13	0.1	8

intervals have a multiplier of 0.8, indicating another period of heavier traffic; and the final interval returns to a multiplier of 1, corresponding to regular traffic flow.

To determine the optimal hyperparameter values for both the MA and MMA, we used the iRace package[1] within the R programming language. Each algorithm was given a budget of 1000 experiments per dataset. Table 1: shows the best-found hyperparameter values for MA and MMA, identified by the iRace package. The row labeled *Possible values* specifies the ranges from which iRace could select during the tuning process. The remaining rows report the selected values that yielded the best performance for each algorithm on the I_1 and I_2 instance sets.

In the evaluation conducted on the I_1 dataset, our analysis showed that MMA outperformed MA in 23 instances when comparing the **average solution quality** across runs. For these cases, the average improvement in solution quality was 45.8%, with differences ranging from 0.15% to 520.35%[2]. Conversely, MMA produced inferior average results in 13 instances, with an average decrease in solution quality of 5.59%, and differences ranging from 0.56% to 27.75%.

A *Wilcoxon signed-rank test*, conducted at a 0.05 significance level and applied to the average solution values, confirmed that the performance difference between the two algorithms is statistically significant, with a p-value of $p < 0.01$.

When comparing the **best-found solutions** across all runs, a different pattern emerges: MA outperformed MMA in 27 instances, matched it in 4, and was outperformed in 5 instances. However, the magnitude of improvement reveals an important contrast: in cases where MA was superior, the average improvement in solution quality was 3.56%, while in the cases where MMA performed better, the average improvement reached a substantial 58.31%. This indicates that although MA more frequently found better best solutions, MMA achieved significantly larger gains in the instances where it excelled.

In terms of the **worst-found solutions**, MMA demonstrated superior performance, producing better results in 23 instances and worse results in 13. The average improvement in solution quality for cases where MMA outperformed MA was 109.36%, compared to just 7.05% when MA performed better.

A similar trend was observed in **average penalties**: MMA achieved lower-penalty solutions in 27 instances, matched MA in 5, and performed worse in 4. Interestingly, MA tended to find solutions with a lower number of routes, suggesting a different strategic preference in its search behavior.

When comparing the **average and maximum route durations**, no statistically significant differences were found between the two methods.

In the evaluation on the I_2 dataset, MMA demonstrated superior performance in 52 instances, while MA outperformed it in only 4. The average difference in solution quality also favored MMA. Specifically, in cases where MMA outperformed MA, the average improvement was 29.11%, with the smallest and largest gains being 3.12% and 55.41%, respectively. Conversely, in the few instances where MA performed better, the average improvement was 10.38%, with differences ranging from 1.25% to 18.62%.

[1] https://cran.r-project.org/web/packages/irace/index.html

[2] We compute the relative improvement using the following formula:

$$\text{Improvement } (\%) = \frac{Z_{\text{MA}} - Z_{\text{MMA}}}{Z_{\text{MMA}}} \cdot 100$$

where Z_{MA} is the average solution value of MA, and Z_{MMA} is the average solution value obtained by MMA. Since the objective is to minimize travel distance, this formulation expresses how many times the MMA result is smaller than the MA result. For instance, an improvement of 520.35% indicates that the MA result is more than five times worse than the one obtained by MMA.

DRAFT VERSION

These results were further supported by statistical analysis, which confirmed a statistically significant difference in average solution quality between the two algorithms, with a p-value of $p < 0.0001$.

MMA also outperformed MA in terms of best-found solutions, worst-found solutions, and the number of routes. However, MA achieved better results in average route duration, maximum route duration, and total penalties.

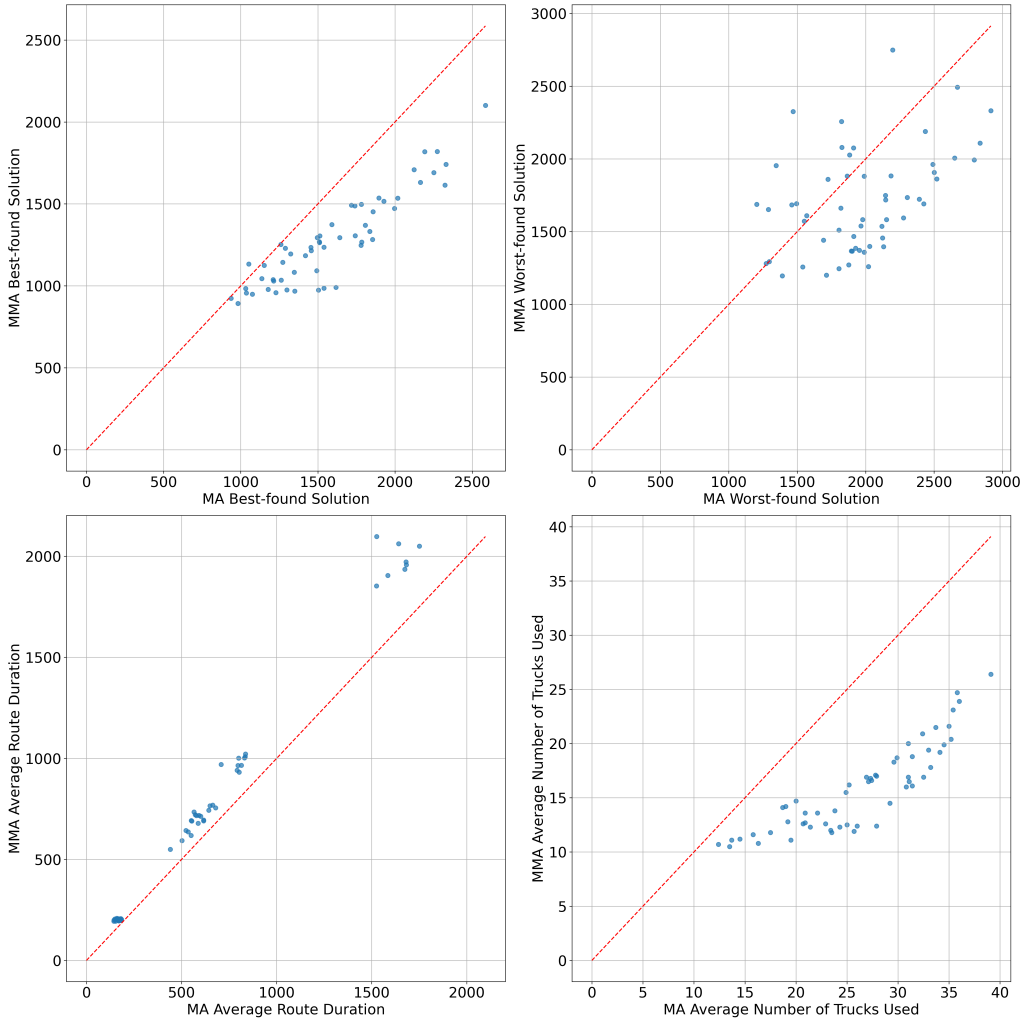


Figure 2: A comparison between MA and MMA across several metrics on I_2

Figure 2 presents scatter plots of the results obtained from testing the instance set I_2 across various metrics. Points above the red line indicate instances where MA outperformed MMA, while points below the line indicate the opposite.

5. CONCLUSION

In this paper, we addressed the Electric Vehicle Routing Problem with soft time windows, time-dependent travel speeds, and partial recharge. The objective was to minimize both the total distance traveled by all vehicles and the penalty costs associated with time window violations. To tackle this problem, we developed and evaluated two memetic algorithms, one standard and one modified, on two sets of benchmark instances. The proposed modified Memetic Algorithm outperformed the standard version in most cases, demonstrating its effectiveness and potential for solving complex routing problems.

However, this study would benefit from further research, particularly by exploring a variant of the Memetic Algorithm in which the top α solutions are refined using local search, a strategy that was not examined in this work.

DRAFT

ACKNOWLEDGMENT

This work was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, Agreement No. 451-03-136/2025-03/ 200029.

LITERATURE

Erdoğan, S., & Miller-Hooks, E. (2012). A green vehicle routing problem. *Transportation research part E: logistics and transportation review*, *48*(1), 100–114.

Eshelman, L. J., & Schaffer, J. D. (1985). Genetic algorithms for the traveling salesman problem. *Proceedings of the 1st International Conference on Genetic Algorithms*, 160–168.

Golden, B. L., Raghavan, S., & Wasil, E. A. (2008). *The vehicle routing problem: Latest advances and new challenges* (Vol. 43). Springer Science & Business Media.

Matijević, L. (2023a). General variable neighborhood search for electric vehicle routing problem with time-dependent speeds and soft time windows. *International Journal of Industrial Engineering Computations*, *14*(2), 275–292.

Matijević, L. (2023b). Utilizing metaheuristics to guide the training of neural networks [Held September 18–21, 2022]. *Proceedings of the L Symposium on Operational Research (SYMOPIS 2023)*, 1057–1062.

Moscato, P., et al. (1989). On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms. *Caltech concurrent computation program, C3P Report*, *826*(1989), 37.

Neri, F., Cotta, C., & Moscato, P. (2011). *Handbook of memetic algorithms* (Vol. 379). Springer.

Schneider, M., Stenger, A., & Goeke, D. (2014). The electric vehicle-routing problem with time windows and recharging stations. *Transportation science*, *48*(4), 500–520.

Toth, P., & Vigo, D. (2014). *Vehicle routing: Problems, methods, and applications*. SIAM.

DRAFT VERSION

INFORMATION SYSTEMS AND TECHNOLOGY

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

MODELING MONO AND MULTI-REPOSITORY DILEMMA FOR MICROSERVICE SOFTWARE ARHITECTURE DEVELOPMENT

RADOVAN ČIKIĆ[1], MILOŠ MILIĆ[2], SINIŠA VLAJIĆ[3]

[1] University of Belgrade, Faculty of Organizational Sciences, Belgrade, cikicsm@gmail.com, 0009-0005-2069-7022
[2] University of Belgrade, Faculty of Organizational Sciences, Belgrade, milos.milic@fon.bg.ac.rs, 0000-0002-2521-7607
[3] University of Belgrade, Faculty of Organizational Sciences, Belgrade, sinisa.vlajic@fon.bg.ac.rs, 0000-0003-3830-2170

Abstract: *Microservice software architecture represents one possible approach for designing complex software products. In industry, microservices have been identified as a common architecture solution, especially when all user requirements, or final product functionalities, are not known or not possible to be defined before the project implementation starts. In software development, source code organization can be structured using a monorepository model, a multirepository model, or a hybrid model, which integrates characteristics of both approaches. This paper presents the advantages and disadvantages of these approaches, provides a definition of microservices, and proposes a model for organizing source code repositories based on these approaches. Model in this research consists of variables that measure the characteristics of software quality, incorporated from ISO 25010 and ISO 25023. Through multiple criteria decision-making, it is possible to choose the alternative that best suits the software system whose development is being carried out. At the very end, final model will be presented with pros and cons of designed model and further research directions using it.*

Keywords: *Mono repository, Multi-repository, Microservices, Source code management, Multiple criteria decision-making.*

1. INTRODUCTION

In the development process of complex applications that solve real business problems, it is necessary to establish a correct software architecture model of such software. According to the Larman software development method (Larman, 2005), the first phase is to gather user requirements. Regardless of whether the software development life cycle (SDLC) model is waterfall, or iterative-incremental, the next phase – analysis – should occur if and only if the previous phase has been performed correctly (for the entire software or its current iteration, respectively). After the analysis is completed, and based on its results – including the system structure (conceptual and relational models) and the system behavior (sequence diagrams and defined system operations) – the subsequent development of the software product can proceed in a well-founded manner (Larman, 2005). From the system structure and behavior, software architect can design correct architecture for user interface, application logic, database broker, and other components. Software architect decides on the architecture of the complete software solution, so that it contributes to the simplicity of creating and maintaining the final product, but also makes it easier for the end user. This conceptual solution is the key for minimization, and ideally for prevention, the possibilities of errors on the client side. Due to the increasing complexity of end-user needs, but also the corporative desire to increase the portfolio of companies (both software and from other industries - retail, tourism, etc.), it is necessary to ensure the scalability and maintainability of the software product we are working on.

To ensure the adequate way for realization of the outputs from development phase, architect needs to choose the correct way for source code management. In the near past, adequate ways have been developed for the implementation of various conceptual solutions of software architecture, and especially VCS (Version Control Systems) - mono and multi-repository, which facilitate the manipulation of the entire source code, its versions, and changes by an unlimited number of development teams. During the early 2010s, the concept of microservices emerged, which further facilitates the programming of complex software solutions, where functionalities are implemented through separate units. This type of architecture enables development of each microservice completely or partially separate from another one, and has positive impact on development process of non-waterfall SDLC models.

DRAFT VERSION

This paper is organized in five sections. After this introduction, the second section will provide definitions, characteristics, and applications related to key concepts. Third section will provide the mathematical model, with adequate definitions of used characteristics and derived equations. Additionally, third section also provides a problem statement for microservices-based software. Fourth section is posed as a dilemma and discussion for one software architect about implementing one of the four source code management models (SCMs)(i.e. mono and multi repository, hybrid mono-as-multi and hybrid multi-as-mono). The dilemma will be examined in the context of the multi-criteria decision making model. Finally, the conclusions and further research directions are presented in Section 5.

2. BACKGROUND

This part of the paper will explain the concepts, characteristics and applications of mono and multi repositories, but also a hybrid approach that represents a combination of the two mentioned above. In addition, the microservice software architecture will be presented, which completes the theoretical foundations in this section.

Before we begin to analyze the concepts of mono and multi-repository, we must know that there are similarities between them: (1) both models ultimately track the same source code files, and do it by using SCM VCS platforms, such as Git or Mercurial; (2) both models are proven successful for projects of all sizes and (3) both models are straightforward to implement using any typical SCM VCS, up to a scaling limit (Henderson, 2019). However, when considering the dilemma at the level of complex projects, which have hundreds of commits and tens of thousands of code files in their codebase, we come to the point where we need to carefully choose one of them in order to make the solution more readable and functionable for the software architects and software developers involved in the project.

2.1. Mono repository

Mono repository (shortened as *monorepo*, or *single repository*) is a type of source control approach where all the components and units of source code are kept in one repository (Shakikhanli & Bilicki, 2022).

The main positive characteristics of monorepo projects are: (1) well-organized hierarchical structure, (2) ease of code reuse, (3) simplified dependency management, (4) atomic commits, (5) large-scale code refactoring and (6) collaboration across teams (Potvin & Levenberg, 2016). On the other side, problems that can be caused by complexity and size of monolithic code base are: (1) tooling investments for both development and execution, (2) codebase complexity – including unnecessary dependencies and difficulties with code discovery and (3) effort invested in code health (Potvin & Levenberg, 2016). Intuitively, those disadvantages are implications of mandatory trade-offs, and they need to be well-handled, or the whole product may not be properly driven. The most famous companies that use monorepo approach are: Google, Meta, Microsoft, Uber and Twitter (Brousse, 2019). In practical work, monorepos may seem inflexible to engineers accustomed to having more autonomy over the tools and dependencies they use in their projects (Jaspan et al., 2018).

2.2. Multi repository

Multi-repository, or *Poly-repository* (shortened as *multirepo* or *polyrepo*) approach manages all the packages and source code components in several repositories (Shakikhanli & Bilicki, 2022).

Multirepo approach offers benefits from using it, such as: (1) better security than monorepo, (2) more feasible and manageable scalability from the beginning, (3) almost instantaneous versioning, (4)easier ownership assigning, (5) better resource distribution, easier migrating than in monorepo, and (6) easier CI/CD, because of basic separation, which makes it easier to facilitate resources (Plaza, 2023). Worldwide companies that use multi-repo approach for their software are Netflix, Amazon, and Lyft, among the others. Engineers note at least two benefits of multi-repo codebases: ability to maintain stable, versioned dependencies. and the freedom and flexibility to select their own toolchain (Jaspan et al., 2018).

2.3. Hybrid approach

Hybrid approach tries to combine best characteristics of the both models from above, but it needs to stand on one of those mainly, and with another one as the “modification tool” for the first. Because of that, there are two hybrid models: hybrid Multi-as-Mono and hybrid Mono-as-Multi approach.

This kind of approach can be suitable in multiple situations. First of all, we can decide the best way to deal with codebase is to have more than one repository, and to use tool like Meta (Walters, 2021) to manage

DRAFT VERSION

them. There, we can execute commands against one, all or specific repositories, simply listing them or using some type of condition(e.g. name or logical). This approach is named as Multi-as-Mono. On the other side, Mono-as-Multi approach separates concerns: during development, code is managed like a monorepo, while for deployment, each library's code is copied into its own independent repository. This strategy is prevalent within the PHP ecosystem because Packagist (the main Composer repository) requires a public repository URL to publish a package, and it's not possible to indicate that the package is located within a subdirectory of the repository (Losoviz, 2021).

2.4. Microservices architecture

Microservices are an architectural and organizational approach to software development in which the software consists of several independent services that communicate with each other (AWS, 2024).Such services make applications easier to develop, accelerate innovation, and reduce the time required for deployment (AWS, 2024). Microservice is an independent unit of work, which is small enough to stand on its own in terms of functions.

That means there must be zero coordination for the deployment with other microservices. From this, we can see microservices work on componentized model – a component is a unit of software that is independently replaceable and upgradeable (Bakshi, 2017). Toleration of the each service failure must be integrated in microservice design, and this is one of the crucial differences from monolithic architecture. Services can fail at any time, therefore, it is important to detect failures quickly and automatically restore service. This is why microservice applications put a lot of emphasis on real-time monitoring of the application, checking both architectural elements (such as how many requests per second is the database receiving) and business relevant metrics (such as how many orders per minute are received) (Bakshi, 2017).

The most important principles that a structure should follow to be called a microservice are: separation of concerns, clear boundaries, portability, its own database, temporality, discoverability, and fault tolerance (Frye, 2023). Microservice architecture is based on: loose coupling, security, service registry, single sign-on, transactions, and communication (Redis, 2023). When it comes to implementation, this kind of architecture does not need the existence of full requirement specification at the beginning of project.

3. MODELING AND CRITERIA FOR DECISION-MAKING

This section will present the model and corresponding problem. They are presented through two subsections. First subsection is proposed for problem statement. The main model and variables within it are presented in second subsection. Additionally, in the second subsection we give the explanation of deriving mathematical model and criteria used.

3.1. Problem statement

A software company is developing a complex enterprise system intended for use in a large-scale organizational environment. The system is based on a microservice software architecture and is designed to support a wide range of business functionalities. These include standard data management operations (such as create, read, update, and delete) for various business entities, as well as support for managing their associated properties and relationships. All of this functionalities are arranged as microservices, to ensure autonomous developing and existence of one concrete functionality in case of deleting or changing the other ones. In decision-making time, the decision-maker (software architect) needs to find the best alternative for software product among the defined scope: mono-repository, multi-repository, hybrid mono-as-multi and hybrid multi-as-mono repository. The top management of software company is open for changing the SCM, but requests a brief explanation of the final solution, based on criteria officially used(i.e. standardized criteria).

3.2. Model and variables

Every multi-criteria decision model needs to have sets of alternatives *(A)* and criteria *(C)* (Čupić & Suknović, 2010). We will define them as vectors. First vector A is scope of available alternatives, where a_1 represents mono-repository, a_2 represents multi-repository, h_1 represents hybrid mono-as-multi, and h_2 represents hybrid multi-as-mono approach. Then, the alternatives vector A for our case looks like: $A = [a_1, a_2, h_1, h_2]$. Criteria vectors are scopes of software quality metrics. According to ISO 25010 (ISO/IEC, 2011), the nine quality characteristics of software product are: functional suitability, performance efficiency, compatibility, interaction capability, reliability, security, maintainability, flexibility and safety. Each of them

DRAFT VERSION

has their own sub-characteristics, and their utilization can be complete or partial, depending on goal which using stands for. Depending on use-case (currently, on purpose of the final software product), all or few of those characteristics can be used for quality evaluation. In the context of repository model selection, we will use sub-characteristics from the performance efficiency subsection of this standard. ISO 25010 describes Time behavior as a degree to which the response time and throughput rates of a product or system, when performing its functions, meet requirements (ISO/IEC, 2011). In same standard, Resource utilization is defined as a degree to which the amounts and types of resources used by a product or system, when performing its functions, meet requirements (ISO/IEC, 2011). Third sub-characteristic, Capacity, is degree to which the maximum limits of a product or system parameter meet requirements (ISO/IEC, 2011). The selected sub-characteristics and relevant software metrics are presented in Figure 1.

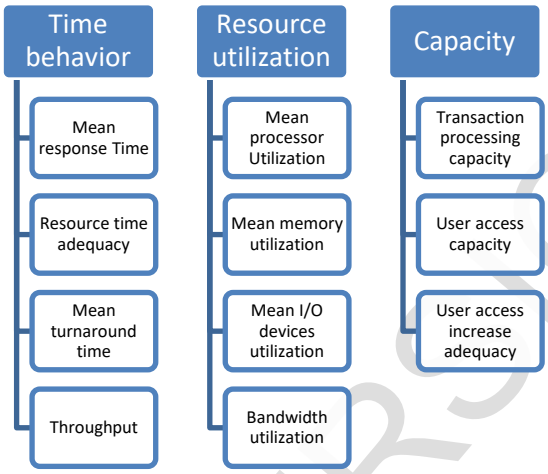


Figure 1. Sub-characteristics and software metrics related to Performance quality characteristic (adapted from ISO 25010 standard) (Karnouskos et al., 2018).

Criteria vectors for C_1, C_2 and C_3 , are the scopes of relevant metrics for evaluating overall validity for Time behavior, Resource utilization and Capacity (Figure 1), respectively. If we generalize these vectors, let $k \in \{1,2,3\}$ denote the index of each strategy vector, and let $m_1, m_2, \ldots m_n$ represent the set of metrics corresponding to each quality criterion, where n is the total number of metrics presented in Figure 1. Then, the general form of each vector can be defined as $C_k = [m_1, m_2, \ldots m_n]$.
Similarly, the general form of the matrix, where columns represent the repository strategy alternatives and rows correspond to the metrics $m_1, m_2, \ldots, m_n$, can be expressed as $O_k = [v_{ij}]$, where v_{ij} is value of metric i for alternative j. Then, there are three matrices in this case, for every sub-characteristic (see Figure 1). Values from matrices O_k follow the rule $v_{ij} >= 0$, because every metric is non-negative value.
After forming criteria vectors, we need to normalize values to adequate scope, preferable from 0 to 1. A common normalization method is the MAXMIN approach, as shown in equation (1).

$$v_{ij}^{norm} = \frac{v_{ij} - min\,(v_i)}{max(v_i) - min(v_i)} \tag{1}$$

However, when dealing with metrics where lower values are required (e.g., response time, turnaround time), the MAXMIN strategy must be adjusted. In such cases, we apply an inverted MAXMIN normalization, as defined in equation (2).

$$v_{ij}^{norm} = \frac{max(v_i) - v_{ij}}{max(v_i) - min(v_i)} \tag{2}$$

After this normalization, matrices O_k are transformed(normalized) into matrices O_k^{norm}. The next step is calculation of scores for every column(alternative), for each matrix(criteria scope) , which is presented in equation (3). Summary of every criteria scope is vector of same number of elements as the number of alternatives.

$$C_i^{norm} = \sum_{i=1}^{3} v_{ij}^{norm} \tag{3}$$

Following this rule, we will have three vectors with four elements each, C_1^{norm}, C_2^{norm} and C_3^{norm}. Those vectors are forming the final matrix F or F_w. In this step, we can introduce weights or not. If we prefer to introduce them, every value for each vector needs to be multiplied by corelated weight coefficient, following

DRAFT VERSION

with rule that summary of all weights needs to be equal to 1. Following all of those rules, the general form of the final weighted matrix F_w(or just final matrix F, if we decide for w_i = 1) can be expressed as $F_w = [w_i * v_{ij}^{norm}]$.

4. DILEMMA AND DISCUSSION

Evaluation of the mentioned metrics through the presented model is possible in two ways. The first is certainly more difficult - developing appropriate applications for each of the alternatives listed, and measuring and finally comparing metrics through the model. The first way is also the way for solving case study. In that sense, final solution needs to have the best scores of mentioned characteristics, and to manage source code to enable the fastest and safest way for its functionalities. Another way can be an isolated test, similar to beta versions, where a balance is achieved with sufficient abstraction - omitting complexity to the point where the abstracted model still realistically depicts the original product, but also including all the details related to the Performance efficiency components: Time behavior, Resource utilization and Capacity. Let's note that there are a large number of variables in the defined model, so the sensitivity of such a model is the primary challenge in this investigation. Examining a total of 12 listed metrics, there is a possibility of "making up" the results, in the sense that the superiority of one or more metrics in relation to the others "corrects" their simultaneous inferiority in terms of quantitative assessment. This situation is not impossible, especially if we consider the simultaneous presence of metrics related to capacity, which refer to the static capabilities of the system, and on the other hand, the presence of metrics related to Time behavior and Resource utilization, which relate to dynamic capabilities. Then too much positive deviation of some metrics can compensate for the lack of others. It should also be borne in mind that this case, although it has meaning and the possibility of realization in practice, still represents an extreme - borderline possibility. The model's response to this possibility is precisely the weighting coefficients that should be added to equation (3), in order not to give importance to that subgroup of metrics that are more relevant to the software product we are developing. In cases that are not borderline, the model provides clarification between four alternatives, ranking them in order of importance and combining the real performance of the system as recorded by tests and the assessment of the decision maker. The model incorporates expertise from the software architect. The decision maker can, using the explained model, determine the weights w_i based on his knowledge and the specific properties of the software system. The assignment of weights depends on: (1) the industry for which the software is developed; (2) the desire of the end user or the need emphasized by the end user during the definition of requirements; and (3) expertise-based evaluation of team members, mid-level management or other employees in the hierarchy that, in accordance with the specific corporate culture, have an impact on determining the greater or lesser participation of certain characteristics from subsection 3.2. in the final assessment of the mentioned software quality characteristics.

5. CONCLUSION

The software quality characteristics defined in the ISO 25010 standard and the metrics defined in ISO 25023 can be reformulated into a model that provides the ultimate software quality assessment. In this paper, such a model is simplified by using only metrics related to, a sub-characteristics of Performance efficiency - one of the characteristics defined by the standard. In addition, the model in this research is based on the fact that one person should make a final assessment (judgment) about which alternative (model) is the most acceptable for implementation. According to the existing model, that person must perform tests of the real performance of all alternatives (models) and assign adequate weights to the sub-characteristics based on the factors described at the end of the previous section, or on the basis of some others.

Practical application of this model is directly correlated with the problem statement from 3.1., especially if the model is more detailed. In the “limited” version from this paper, model can contribute to the utilization of software working-time, especially execution time, and hardware resources. The main model limitation is that it doesn’t incorporate all characteristics and metrics from the ISO standard itself. Additional metrics (e.g., from the ISO 25010 standard, other standards, or user-defined metrics) and their corresponding weights can be incorporated later as a custom model extension.

Further research can be carried out in three directions. (1) Expansion of the model to all nine attributes defined in ISO 25010. Then the model becomes more complex, which contributes to increasing its sensitivity through the increase of variables, but also comprehensiveness, since then its scope includes all characteristics of software quality that the international standard provides. (2) Inclusion of relevant decision makers, listed in Larman's or some other software development method. Then it becomes a group decision-making model. One must be careful here, because the model acquires the dimension of subjectivity of different decision-

DRAFT VERSION

makers, who may differ in expert knowledge, experience, level in the company's hierarchy, or in some other aspect of importance for authority in decision-making. (3) Rounding of the whole picture from the aspect of decision-making theory, where the payoff and/or regret tables are included in the model. They should take into account the quantified values of payoffs and losses for each alternative and, based on that, give the decision maker a complete insight into the financial background of the project. Also, an assessment of the time trade-off and other resources that the decision makers or their superiors consider to be important for the company developing the project is carried out.

ACKNOWLEDGMENT

The authors express their gratitude to the University of Belgrade – Faculty of Organizational Sciences for its support of this research.

LITERATURE

AWS, What Are Microservices (2024), https://goldenowl.asia/blog/what-are-microservices-a-complete-guide-in-2024 , access date: 16.06.2025.

Bakshi, K. (2017). Microservices-based software architecture and approaches. In 2017 IEEE Aerospace conference (pp. 1-8). IEEE. DOI: https://doi.org/10.1109/AERO.2017.7943959

Brousse, N. (2019), The Issue of Monorepo and Polyrepo In Large Enterprises, Programming '19: Companion Proceedings of the 3rd International Conference on the Art, Science, and Engineering of Programming, 1-4, https://doi.org/10.1145/3328433.3328435

Čupić, M., Suknović, M. (2010), Decision making: Formal approach, Faculty of Organizational sciences, Belgrade

Frye, M. (2023). 6 Fundamental principles of Microservice Design. Salesforce. https://www.salesforce.com/blog/microservice-design-principles/ , access date: 16.06.2025.

Henderson, J.P. (2019), Monorepo vs. polyrepo, Github, https://github.com/joelparkerhenderson/monorepo-vs-polyrepo , last access: 16.06.2025.

ISO/IEC 25010, https://iso25000.com/index.php/en/iso-25000-standards/iso-25010 , access date: 18.06.2025.

Jaspan, C., Jorde, M., Knight, A., Sadowski, C., Smith, E., Winter, C., Murphy-Hill, E. (2018), Advantages and Disadvantages of a Monolithic Repository, ACM/IEEE 40th International Conference on Software Engineering: Software Engineering in Practice (pp. 225-234), https://doi.org/10.1145/3183519.3183550

Karnouskos, S., Sinha, R., Leitão, P., Ribeiro, L., Strasser, T. (2018), The Applicability of ISO/IEC 25023 Measures to the Integration of Agents and Automation Systems , IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society, 2927-2934, https://doi.org/10.1109/IECON.2018.8592777

Larman, C. (2005), Applying UML and patterns: An introduction to object-oriented analysis and design and iterative development (3rd ed.) , Prentice Hall.

Losoviz, L. (2021), Monorepo vs Multi-Repo: Pros and Cons of Code Repository Strategies, https://kinsta.com/blog/monorepo-vs-multi-repo/#hybrid-monoaspoly-approach , access date: 18.06.2025.

Plaza, F. (2023), Monorepo vs. multi-repo: Different strategies for organizing repositories, https://www.thoughtworks.com/insights/blog/agile-engineering-practices/monorepo-vs-multirepo , access date: 18.06.2025.

Potvin, R., Levenberg, J. (2016), Why Google Stores Billions of Lines of Code in a Single Repository, Communications of the ACM, Volume 59, Issue 7, 78-87, https://doi.org/10.1145/2854146

Redis (2023). Microservice Architecture Key Concepts, access date: 16.06.2025. https://redis.io/blog/microservice-architecture-key-concepts/

Shakikhanli, U., Bilicki, V. (2022), Comparison between mono and multi repository structures, Pollack Periodica, Volume 17, Issue 3, 7-12, https://doi.org/10.1556/606.2022.00526

Walters, M. (2021), Meta, https://github.com/mateodelnorte/meta/blob/master/README.md , access date: 18.06.2025.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

DIRECTION OF ARRIVAL ESTIMATION IN FMCW RADAR USING THE MUSIC-2D ALGORITHM

BORKO ĐAKOVIĆ[1], SLOBODAN SIMIĆ[2], DANIJELA PROTIĆ[3], KRISTINA ŽIVANOVIĆ[4]

[1] Center for Applied Mathematics and Electronics, Belgrade, Serbia. borko.djakovic@vs.rs, 0009-0000-4662-3981
[2] University of Defense, Military Academy, Belgrade, slobodan.simic@va.mod.gov.rs, 0000-0002-6846-4886
[3] Center for Applied Mathematics and Electronics, Belgrade, Serbia, danijelaprotic318@gmail.com, 0000-0003-0827-2863
[4] Center for Applied Mathematics and Electronics, Belgrade, Serbia, kristina.zivanovic@vs.rs, 0009-0004-3468-3400

Abstract: *This paper presents the MUSIC-2D algorithm applied to determine the angle of targets of the FMCW radar. As a result, figures, two-dimensional heat maps, and Range-Angle (RA) maps are presented. The distance and angle at which the targets are located are observed. A target azimuth accuracy estimation and angular resolution using the Angle DFT and MUSIC-2D algorithms are compared. The system and the signal model are described. Simulations and data processing of several scenarios experimentally determined by the FMCW TDM MIMO commercial radar were performed in MATLAB.*

Keywords: *MUSIC-2D, Angle DFT, FMCW Radar, Range-Angle map.*

DRAFT VERSION

1. INTRODUCTION

The need for precise determination of key target parameters - such as range, velocity and angle at which the target is located has led to the widespread use of FMCW radars. The FMCW radars are usually used in medicine (Alizadeh et al., 2019; Bliss et al., 2006), in drone with low radar cross-section detection (de Quevedo et al., 2018), and in the automotive industry (Bilik et al., 2019). Due to its large Time-to-Bandwidth (tB), FMCW radar allows for accurate and precise measurement of target distance and velocity. Moreover, the angular estimation and resolution relate to the antenna aperture and the number of receive antennas (Richards, 2005). In real-world scenarios, due to limited space, cost reduction requirements and hardware resource optimization, the implementation of a large number of receive antennas is often impractical. When using conventional angle estimation algorithms such as Angle DFT, even radars which use multiple transmit antennas and form a virtual antenna array do not provide sufficiently accurate results. For this reason, it is necessary to apply high-resolution algorithms such as MUSIC.

This paper presents the MUSIC applied to determine the angle and distance to the target, resulting in a two-dimensional Range-Angle (RA) map. A system and signal model used for both simulation and experimental verification were presented. In addition, the AngleDFT and MUSIC-2D algorithms are described. Several scenarios are investigated to illustrate the performance of the algorithms, including the accuracy of angle estimation and angular resolution, along with a comparative analysis of the results obtained using the aforementioned algorithms.

2. SISTEM AND SIGNALS MODELS

The system and signals models are described in text that follows. First the time division multiplex (TDM), multiple input multiple output (MIMO), frequency modulated continuous wave FMCW radar system is presented. In addition, two-dimensional complex signals are described.

2.1. System model

A TDM MIMO FMCW radar system consists of two transmit antennas and four receive antennas (Figure 1). The FMCW signal is fed to the transmit antenna via a microwave switch, which controls the antennas, in terms of transmission time. The transmission occurs alternately between the two antennas. After the signal is reflected from the target, it is received by a linear antenna array. The target is assumed to be in the far field, so the reflected signal reaches the receive antennas as a plane wave. The received signal passes through a low noise amplifier (LNA) and is then mixed with the transmitted signal, as well as a version of the same signal

phase-shifted by 90°. An in-phase (I) and quadrature (Q) components are generated in this process. After passing through a low-pass filter, the intermediate frequency (IF) signals are generated and then sampled using an analog-to-digital converter (ADC) and processed using digital signal processing (DSP). The I and Q samples are combined to form complex signal samples. These samples are organized into a three-dimensional Radar Data Cube (RDC), where one axis contains samples originated from one chirp, the other contains multiple chirp samples, and the third axis contains samples of signals received from multiple antennas.

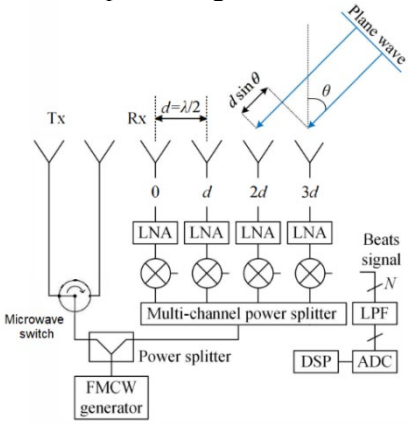


Figure 1: Radar system model

Two separate radar data cubes are formed, each containing samples of a signal from a single transmit antenna. Combination of samples creates a virtual array with one transmit antenna and eight virtual receive antennas; and one RDC is formed. Given the short duration of the chirp (a few milliseconds), it is assumed that the target remains within the same resolution cell, so that Doppler phase compensation is not required. By determining the beat frequency in the IF signal, which depends on the target position, information about the distance is extracted. The delay of the signal over the receive array is used to estimate the angle of arrival (Đaković et al., 2025).

2.2. Signal model

To generate a RA map at a given time, signal samples from a single chirp and multiple receive antennas are analyzed. These samples are analytically represented as a two-dimensional complex exponential signal (Manokhin et al., 2015):

$$Y(m,n) = \sum_{l=1}^{L} \alpha_l e^{j\varphi_l} e^{j2\pi \frac{2R_l B}{cT_c} mT_S} e^{j\frac{2\pi}{\lambda} dn \sin\theta_l} + n_n(m,n), \quad (1)$$

where $\alpha_l e^{j\varphi_l}$ denotes complex amplitude and phase, R_l is the distance, θ_l is the angle of l-th target out of the L total targets, B is the FMCW signal bandwith, c is the propagation speed of EM waves (considered equal the spped of light in the air), T_c is the duration of chirp, T_S denotes the sampling period, λ is the wavelength of transmission signal, the distance between recieving antennas is denoted as $d = \lambda/2$, $n_n(n,m)$ is the noise, $n=0...N_{R_x}$-1, where N_{R_x} is the number of recieving antennas and $m=0...N_o$-1, where N_o denotes the number of samples in the chirp ($N_o=T_c/T_S$). Matrix of samples is denoted as Y and is of dimension [N_o x N_{R_x}].

3. ANGLE DFT AND MUSIC-2D ALGORITHMS

3.1. Angle DFT

The Angle DFT algorithm is usually applied after the Range DFT algorithm, after the bit rate and range of the target have been determined using the Discrete Fourier Transform (DFT). The Angle DFT algorithm is based on the application of the DFT to estimate the phase shift between signals received by multiple receiving antennas. The phase shift is the result of the time delay introduced by the additional path the signal transmits when the target is at a certain angle to the radar (Đaković et al., 2025). The target angle is given as follows:

$$\theta_u = sin^{-1}(u\frac{\lambda}{Nd}), \quad (2)$$

where $-\frac{N}{2} \leq u < \frac{N}{2}$, and N represents the number of the DFTcoeficients. The measurement range of the angle and angular resolution can be determined as follows:

$$-sin^{-1}(\frac{\lambda}{2d}) \leq \theta < sin^{-1}(\frac{\lambda}{2d}), \quad (3)$$

DRAFT VERSION

$$\theta_{\text{res}} = 2 \cdot sin^{-1}(\frac{\lambda}{2N_{\text{R}_x} d\, cos(\tilde{\theta})}), \quad (4)$$

the central angle $\tilde{\theta}$, is the angle of the target arrival. The angular resolution does not depend on the number of DFT coefficients, but solely on the number of receiving antennas.

3.2. Music-2D

The matrix of the received signals can be determined as in Equation (5).

$$Y = AXS + N, \quad (5)$$

$$A = [\, a(\theta_1), a(\theta_2), ..., a(\theta_L)\,]_{N_{\text{R}_x} \text{ x } L}, \quad (6)$$

$$a(\theta_l) = [\, 1, e^{-j\frac{2\pi}{\lambda} d sin(\theta_l)}, ..., e^{-j\frac{2\pi}{\lambda} d(N-1) sin(\theta_l)}\,]^{\text{T}}_{1 \text{ x } N_{\text{R}_x}}, \quad (7)$$

$$X = diag\,[\, \alpha_1 e^{j\varphi_1}, \alpha_2 e^{j\varphi_2}, \dots, \alpha_L e^{j\varphi_L}]_{\text{L x L}}, \quad (8)$$

$$S = [\, s(R_1), s(R_2), ..., s(R_{\text{L}})\,]^{\text{T}}_{N_{\text{o}} \text{ x L}}, \quad (9)$$

$$s(R_l) = [\, 1, e^{-j2\pi\frac{2R_l B}{cT_{\text{c}}} T_{\text{s}}}, \dots, e^{-j2\pi\frac{2R_l B}{cT_{\text{c}}}(N_{\text{o}}-1)T_{\text{s}}}]^{\text{T}}_{1 \text{ x } N_{\text{o}}}, \quad (10)$$

where A is the matrix of the angular steering vector $a(\theta_l)$, X is the diagonal matrix of complex amplitudes and phases of the signal, S is the matrix of the time steering vector $s(R_l)$, and N is the additive noise matrix.

The MUSIC algorithm relies on the singular decomposition of the autocorrelation matrix of the signal C_Y, enabling the distinction between noise and signal subspaces. It is shown that the signal subspace depends on the number of targets. By comparing the steering vector to the noise subspace, the angle of arrival the signal can be estimated with high resolution. The MUSIC-2D algorithm follows further steps (Manokhin et al., 2015):

- Reorganize matrix Y into the column vector $\tilde{Y}$ by stacking all the rows sequentially, into a single column as follows:

$$\tilde{\text{Y}} = [\, \text{Y}(1 \dots N_{\text{R}_x}, 1), \text{Y}(1 \dots N_{\text{R}_x}, 2), \dots, \text{Y}(1 \dots N_{\text{R}_x}, N_{\text{o}})]^{\text{T}}_{1 \text{ x } N_{\text{R}_x} N_{\text{o}}}, \quad (11)$$

- Determine the correlation matrix $C_{\tilde{Y}}$ $[\, N_{R_x} N_o \text{ x } N_{R_x} N_o]$:

$$C_{\tilde{Y}} = \frac{1}{N_{\text{R}_x} N_{\text{o}}} \tilde{\text{Y}} \tilde{\text{Y}}^{\text{H}}, \quad (12)$$

- Perform singular value decomposition:

$$C_{\tilde{Y}} = \text{Q}\, \Lambda\, \text{Q}^{\text{H}}, \quad (13)$$

- Determine the number of targets L. It is common practice to compute Range-Doppler matrix firstly. Based on the structure and peak distribution within this matrix, the number of detected targets can be estimated.
- Construct the matrix Q_n, based on matrix Q, such that Q_n contains samples corresponding to the signal subspace.

$$\text{Q}_{\text{n}} = [\text{Q}\,(L+1), \dots, \text{Q}(N_{\text{R}_x} N_{\text{o}})]_{N_{\text{R}_x} N_{\text{o}} \text{ x } N_{\text{R}_x} N_{\text{o}} - L}, \quad (14)$$

where Q(n), is n^{th} column of matrix Q.

- Determine column vector V(R,Θ):

$$\text{V}(R, \Theta) = [\text{v}(1 \dots N_{\text{R}_x}, 1), \text{v}(1 \dots N_{\text{R}_x}, 2), \dots, \text{v}(1 \dots N_{\text{R}_x}, N_{\text{o}})]^{\text{T}}_{1 \text{ x } N_{\text{R}_x} N_{\text{o}}}, \quad (15)$$

$$\text{v} = \text{a}(\Theta)\text{s}(R)^{\text{T}}, \quad (16)$$

- Determine the value of function P(R,Θ):

$$\text{P}(R, \Theta) = \frac{1}{\text{V}(R,\Theta)^{\text{H}}\, \text{Q}_{\text{n}} \text{Q}_{\text{n}}^{\text{H}}\, \text{V}(R,\Theta)}, \quad (17)$$

- Compute the values across all angles and ranges within the search space. Determine the corresponding response matrix and determine the target's range and angle of arrival, based on its maximum value.

4. SIMULATION

For the simulation purposes, a given radar cube was generated using the MIMOFMCW.m function. The radar parameters used in both the simulation and experimental setup are: operating frequency f_c = 24 GHz, signal bandwidth B = 1 GHz, chirp duration T_c = 0,5 ms, sampling frequency f_s = 64 KHz, the number of samples in the chip N_o = 32, the number of antennas in the virtual antenna array N_{R_x} = 8. The parameters set in this way enable a unique radar range of up to 4.8 m and angle measurement in the range from -90º to 90º. Using the MUSIC method, a resolution step of 0.1 m per distance in 0.5º per angle was taken, which makes

DRAFT VERSION

the RA map of dimensions [97 x 361]. Therefore, for fair comparison, the DFT coefficients using the Angle DFT method were calculated for the same number of points.

CASE 1

The target parameters used in the simulation are presented in Table 1. The parameter A denotes the amplitude of the echo signal received from the target, which can be related to the radar reflective surface, r indicates the distance of the target, v is its speed and θ is the angle at which the target is located relative to the radar.

Table 1: Simulation parameters

Target No	A [V]	r [m]	v [m/s]	θ [º]
1	1	4.5	1	70
2	0.8	3	2	-70
3	0.8	3	3	-50
4	0.7	1.5	4	-10
5	0.7	1.5	5	10
6	0.6	0.5	6	0

In this case, the target speed is not considered critical, since the Range-Doppler map is not generated. The targets are generated so that target 1 is positioned at a large observing angle, while target 6 is at a observing angle of 0º. Targets 2 and 3, as in 4 and 5, are arranged in pairs, such that the angular difference between the pairs is equal, but the central angle of arrival for the first pair is equal to 60º, while that of the second pair is equal to 0º. It is considered to be ideal conditions and there is no influence of noise. Figure 2 shows the RA matrix obtained using the RangeDFT algorithm (left), and a heatmap image (right).

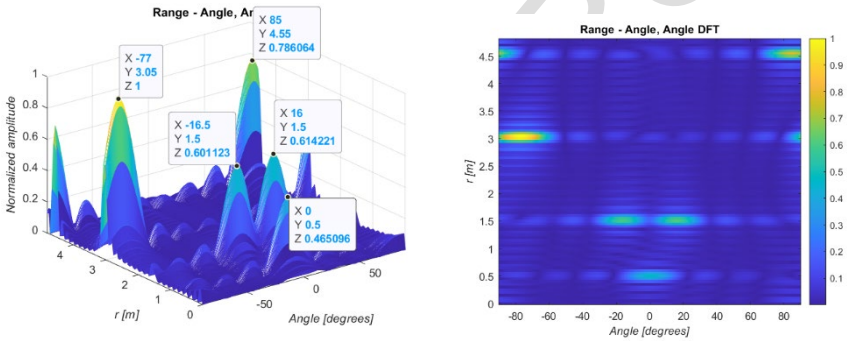


Figure 2: RA matrix (left), and RA map (right) (AngleDFT)

The matrix is normalized in accordance with its maximum value, which allows for consistent visual interpretation. By identifying the peak (maximum) values of the lobes, the angle and range of the targets can be determined. It can be concluded that the angles at which the targets are located are estimated with a certain error, because the algorithm does not offer high accuracy, due to the wide main lobes in the spectrum estimation. Due to the width of the lobes in both dimensions, a small error occurs in the range estimation for targets 1, 2 and 3. Targets 2 and 3 were not successfully separated and were located under one lobe, despite having the same angular distance as targets 4 and 5. The consequence was a decrease in the angular resolution with an increase in the viewing angle. Despite a number of targets in the space, the first singular value dominates, because of high correlation between the target signals due to the FMCW radar signal. Figure 3 shows the RA matrix obtained using the MUSIC-2D algorithm (left), and a heatmap image (right).

The maximum of the function very accurately determines the angle at which the target is located. The angle of target 1 is more difficult to determine, due to the large angle of arrival, the spectrum of the signal is mapped to the other end of the matrix. Targets 2 and 3 were not successfully separated because of reduced angular resolution cell. Unlike the previous method, the other parts of the spectrum had a high-level value; the maxima are distributed in the range of 0.6 to 1 of normalized value, so that targets with a low amplitude are more noticeable.

FMCW radars can detect signals even at low signal-to-noise ratios (SNR) (Simić, 2007). Figure 4 shows the RA maps corresponding to SNR levels of 5 dB, 0 dB, and -5 dB. It is shown that the MUSIC-2D algorithm is more sensitive to noise, compared to the AngleDFT algorithm, and already at 0 dB SNR ratios, lower amplitude targets cannot be detected.

DRAFT VERSION

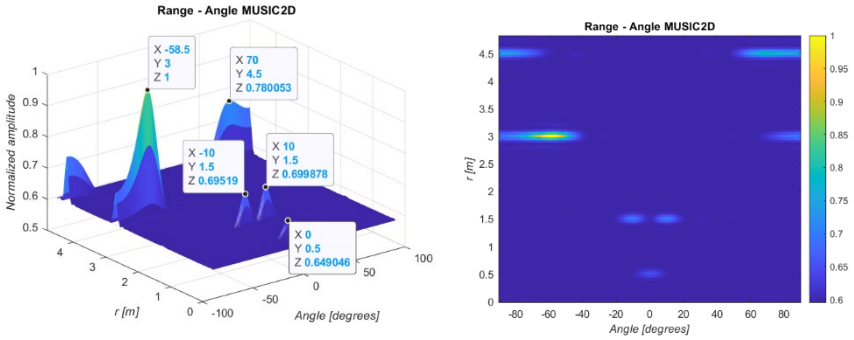


Figure 3: RA matrix (left) and RA map (right) (MUSIC-2D)

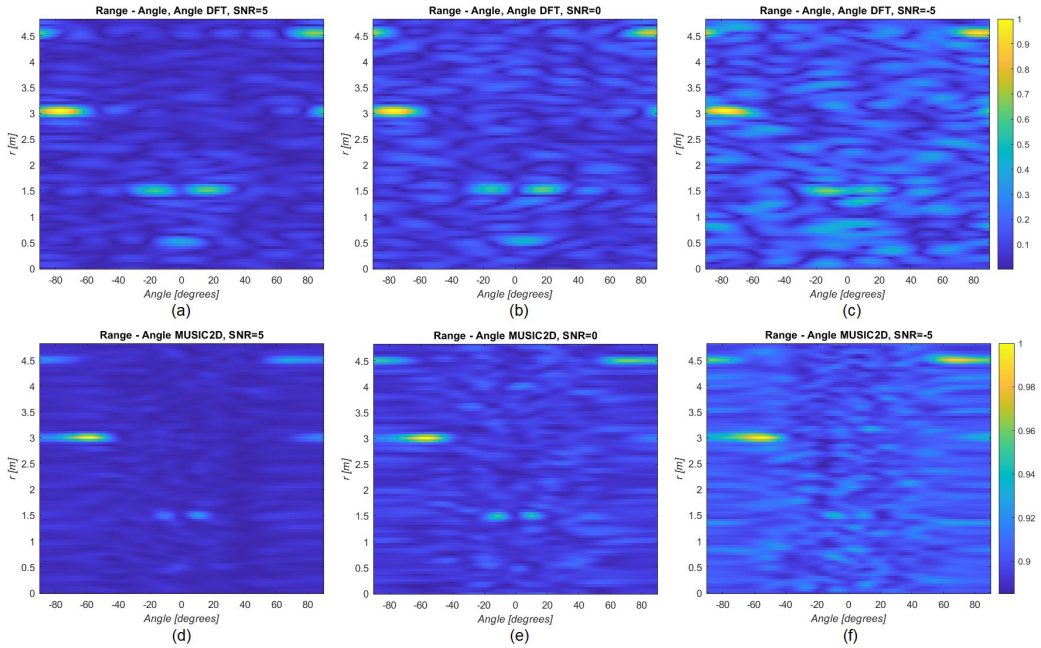


Figure 4: RA maps: a) AngleDFT, SNR = 5 dB, b) AngleDFT, SNR = 0 dB, c) AngleDFT, SNR = -5 dB, d) MUSIC-2D, SNR = 5 dB, e) MUSIC-2D, SNR = 0 dB, f) MUSIC-2D, SNR = -5 dB

At low SNR values, target visibility can be improved if they are highlighted. This procedure involves subtracting the mean value of the matrix from each element, which suppresses the dominant background values, while emphasizing the regions containing real targets. Afterwards, the rows corresponding to the target ranges have to be normalized with respect to the maximum value of each row. Data on the distance of the target can be determined from the Range-Doppler map (Manokhin et al., 2015).

CASE 2

In Case 2, only targets with low normalized amplitudes (0.6 and 0.5) are simulated, positioned at angles of −50° and 40°, with a SNR=−5 dB. As in the previous case, the MUSIC-2D algorithm estimates the angle more precisely, and targets remain detectable even at lower amplitude and SNR levels, which suggests that the efficiency of the MUSIC-2D algorithms is affected by the number of targets (Figure 5).

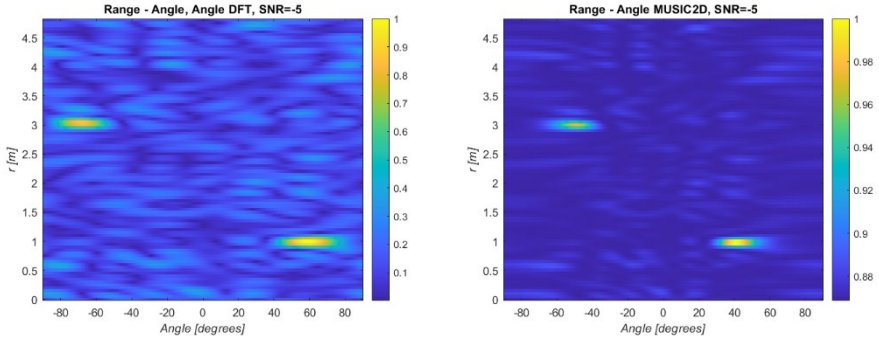


Figure 5: RA map AngleDFT (left) and RA map MUSIC-2D (right)

5. EXPERIMENTAL VERIFICATION

The Luswave MIMO radar platform PUP_DUAL24P_T2R4 was used to conduct the experiments. The platform generates the FMCW waveforms and allows changing parameters such as duration, number of samples in the chirp, operating frequency, frequency range of the signal, and similar (Luswave, 2020). The target was a man moving towards the radar at an angle of approximately -30º. In he carried a Lunenberg lens, a radar reflector, to amplify the reflection of the signal.

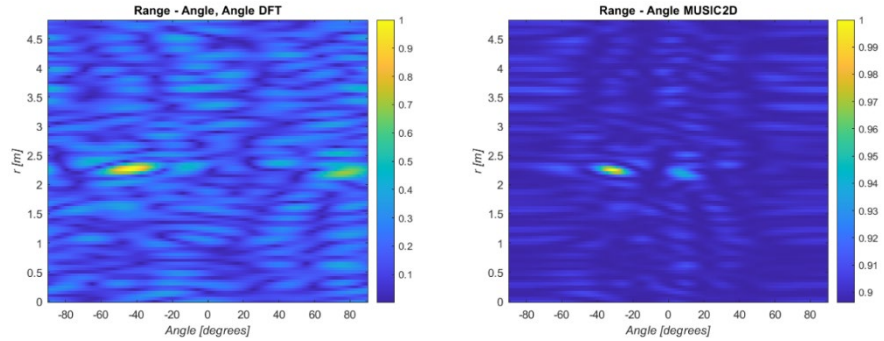


Figure 5: Experimental results: RA map AngleDFT (left) and RA map MUSIC-2D (right)

6. CONCLUSION

The MUSIC-2D algorithm was tested and compared to the AngleDFT algorithm in terms of determining the angle of target using FMCW radar. The MUSIC-2D algorithm is shown to be more accurate and computationally complex, and more sensitive to noise compared to the AngleDFT algorithm. By applying this algorithm, the angular resolution does not increase significantly, but still depends on the number of receiving antennas. Due to the high correlation and similarity between FMCW echo signals and with a larger number of targets, it is more difficult to detect targets of low amplitude. The AngleDFT proved to be a robust algorithm, which estimates the target angle with noticeable errors. However, its implementation is simpler and at low SNR ratios, and it detects low amplitude targets more successfully. Further research can be directed towards more advanced algorithms such as Decorrelation Algorithms, Spatial Smoothing, or Forward-Backward Averaging.

REFERENCES

Alizadeh, M., Shaker, G., Almeida, J., Morita, P., & Safavi-Naeini, S. (2019). Remote monitoring of human vital signs using mm-Wave FMCW radar. *IEEE Access*, *7*, 54958-54968. https://doi.org/10.1109/ACCESS.2019.2912956

Bliss, D., & Forsythe, K. (2006). MIMO Radar Medical Imaging: Self-Interference Mitigation for Breast Tumor Detection. *Proceedings of the Fortieth Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, 1558-1562. https://doi.org/10.1109/ACSSC.2006.355020

Bilik, I., Longman,O., Villeval, S., & Tabrikian, J. (2019). The Rise of Radar for Autonomous Vehicles: Signal Processing Solutions and Future Research Directions. *IEEE Signal Processing Magazine*, *36*(5), 20-31. https://doi.org/10.1109/MSP.2019.2926573

De Quevedo, A., Urzaiz, F., Menoyo, J., & López, A. (2018). Remotely Piloted Aircraft Detection with Persistent Radar. *Proceedings of the 15th European Radar Conference (EuRAD)*, Madrid, Spain, 317-320. http://dx.doi.org/10.23919/EuRAD.2018.8546546

Đaković, B. Simić, S., Trivundža, L., & Ristić, A. (2025). Digital Signal Processing in MIMO Radars with Time Multiplexed Transmit Signals. *Military Technical Courier, 73*(2), 519-552. https://doi.org/10.5937/vojtehg73-54972

Luswave Technology (2020). *PUP_DUAL24P_T2R4 Datasheet.* https://luswave.com/PUP_EN24P_T2R4

Manokhin, G., Erdyneev, Z., Geltser, A., & Monastyrev, E. (2015). MUSIC-based algorithm for range-azimuth FMCW radar data processing without estimating number of targets. *Proceedings of the 15th IEEE Mediterranean Microwave Symposium (MMS)*, Lecce, Italy, 1-4. https://doi.org/10.1109/MMS.2015.7375471

Richards, M. (2005). *Fundamentals of radar signal processing* (volume 1). New York: Mcgraw-hill, 2005.

Simić, S. (2007). FMCW radar signal analysis by time-frequency representations and correlation techniques. *Vojnotehnički glasnik*, *4*, 400-413. https://scispace.com/pdf/fmcw-radar-signal-analysis-by-time-frequency-representations-4ft0yli9f6.pdf

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

DETERMINING OPTIMAL REQUEST RATE TO MAINTAIN THERMAL EQUILIBRIUM IN SERVICE-ORIENTED SOFTWARE ARCHITECTURE

MILOŠ MILIĆ[1], NEBOJŠA NIKOLIĆ[2], DRAGANA MAKAJIĆ-NIKOLIĆ[3]

[1] University of Belgrade, Faculty of Organizational Sciences, Belgrade, milos.milic@fon.bg.ac.rs, 0000-0002-2521-7607
[2] University of Belgrade, Faculty of Organizational Sciences, Belgrade, nebojsa.nikolic@fon.bg.ac.rs, 0000-0003-3022-5617
[3] University of Belgrade, Faculty of Organizational Sciences, Belgrade, dragana.makajic-nikolic@fon.bg.ac.rs, 0000-0002-0790-6791

Abstract: *Service-oriented software architecture is a widely adopted architectural paradigm. This paper explores service-oriented architecture in the context of thermodynamic principles. Recognizing that services consume computational resources and interact via communication channels, we establish a conceptual mapping between key thermodynamic concepts and components of service-oriented systems. Within this framework, we introduce the notion of Service Temperature which is defined as a normalized abstraction of a service's runtime load or operational "stress", as well as Thermal Equilibrium defined as the state in which all interconnected services maintain an equal Service Temperature. Building on this foundation, we develop a mathematical model to determine the optimal request rate necessary to preserve thermal equilibrium. A set of experiments is conducted to validate the mathematical model. The results demonstrate that this approach enables software architects to design service-oriented systems which services are in runtime balance in terms of operational load (i.e., in thermal equilibrium).*

Keywords: *Service-Oriented Architecture, Thermodynamics, Software Metrics, Service Temperature, Service Optimization, Thermal Equilibrium.*

DRAFT VERSION

1. INTRODUCTION

Software systems play an important role in nearly every aspect of modern life and are utilized in various industries like communication, transportation, education, healthcare or finance. As these systems grow in scale and complexity, ensuring their non-functional quality requirements like reliability, performance, and adaptability become increasingly essential (Haigh, 2010). Understanding and maintaining an appropriate level of various non-functional requirements can be a complex and ongoing challenge. Achieving an appropriate balance among these requirements requires thoughtful design. In this context, software systems must be carefully architected to support and maintain these non-functional characteristics throughout their lifecycle.

One of the most critical artefacts produced during the software design phase is the software architecture, which serves as the foundational blueprint guiding all subsequent development and system evolution (Washizaki, 2024). Software architecture defines the fundamental concepts and properties of a software system within its environment, as reflected in its elements, their relationships, and the guiding principles of its design and evolution (ISO/IEC/IEEE 42010, 2022; Washizaki, 2024).

Service-oriented software architecture (SOA) represents one possible approach in software architecture design. This paper examines service-oriented software architecture in the context of thermodynamic principles. We have established a conceptual mapping between key thermodynamic concepts and components of service-oriented systems. Based on these mappings, a mathematical model to determine the optimal request rate to maintain thermal equilibrium is defined.

The rest of the paper is organized as follows. Section 2 introduces the key concepts of service-oriented architecture and establishes connections with principles from thermodynamics. Section 3 presents the formulation of the mathematical model. In Section 4, numerical results are discussed to illustrate the behaviour of the proposed model. Finally, Section 5 concludes the paper and outlines directions for future research.

2. SERVICE-ORIENTED SOFTWARE ARCHITECTURE AND THERMODYNAMICS

This section introduces the key concepts of service-oriented architecture and establishes their alignment with thermodynamic principles. These insights will form the basis for further modelling.

2.1. Service-Oriented Architecture

There is no single universal software architecture; instead, a wide range of architectural styles, patterns, and models exist to address different system goals and contexts (Hasselbring, 2018; Washizaki, 2024). One approach is service-oriented architecture (SOA), which is based on the principle of designing software as a collection of loosely coupled, reusable, and interoperable services that communicate over a network (Papazoglou & Van Den Heuvel, 2007). SOA promotes modularity and flexibility, making it well-suited for distributed and scalable systems. This approach can be realized in various concrete forms, including Microservice architecture (Alshuqayran et al., 2016), Cell-based architecture (Abeysinghe, 2025), or Function-as-a-Service (FaaS) serverless architecture (Shahrad et al., 2019) – each offering distinct advantages and trade-offs in terms of scalability, deployment, and granularity of services (Cordingly et al., 2020; de Oliveira Rosa et al., 2020; Ford et al., 2021; Hassan & Bahsoon, 2016).

A conceptual overview of service-oriented architecture is presented in Figure 1. Typically, this architectural approach includes components responsible for service management, such as service registration, discovery, orchestration, and choreography. In addition, a variety of services can be identified, each providing specific functionality. In addition, Figure 1 depicts that each service requires resources for operational work. Although these services are loosely coupled, they can communicate with each other (i.e., *Service A* to *Service C*, and *Service B* to *Service C* in the figure). Communication between services is realized through well-defined interfaces and standardized protocols.

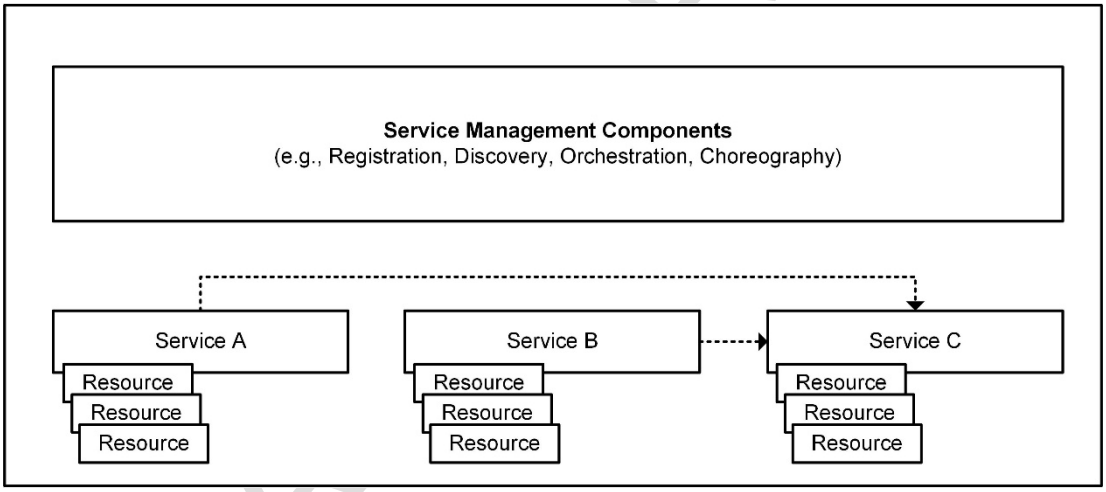


Figure 1: Conceptual overview of a service-oriented architecture

2.2. Thermal Equilibrium in the Context of Software Architecture

Thermodynamics, a fundamental branch of physics, studies the relationships between heat, temperature, work, and energy (Walker, 2017). One of its foundational principles, the Zeroth Law of Thermodynamics, establishes the basis for thermal equilibrium: if System A and System B are each in thermal equilibrium with the System C, they are also in equilibrium with each other (Walker, 2017). This implies that when systems are brought into thermal contact, they will eventually equalize in temperature if no external interference occurs (Walker, 2017).

This principle can be abstracted to the domain of service-oriented software architectures, where a system is composed of a number of interconnected software services (i.e., subsystems). In this context, each service can be seen as a thermodynamic object. Services consume computational resources to perform their operations and interact with one another through communication channels – which can be interpreted as the thermal contacts between subsystems. In this context, we introduce the concept of Service Temperature as a normalized abstraction representing the runtime "stress" or operational load of a service.

Various performance metrics can be used to characterize the operational load of a software service. Each service is typically deployed independently, often within a virtual machine (VM) or a containerized

DRAFT VERSION

environment, which provides isolation and allows fine-grained monitoring of resource consumption. In this direction, we will focus on four key software metrics: CPU Utilization, Memory Usage, Latency, and Request Rate:

- CPU Utilization reflects the consumption of processing capacity by a service over time (Pacifici et al., 2008),
- Memory Usage indicates the amount of physical or virtual memory allocated or actively used by the service to perform tasks and related procedures (ISO/IEC/IEEE 24765, 2017),
- Latency represents the time taken from the initiation of a service request and the start of actual processing (ISO/IEC/IEEE 24765, 2017),
- Request Rate corresponds to the number of requests received by the service per unit of time (Shen, 2010).

Given this model, Thermal Equilibrium in the context of service-oriented software architecture is defined as the state in which all interconnected software services maintain the same Service Temperature.

3. MATHEMATICAL MODEL FORMULATION

The main assumptions of the optimization problem under consideration in this paper are:

- The set of interconnected software services is given;
- Service Temperature is calculated as weighted sum of CPU Utilization, Memory Usage, Latency, and Request Rate;
- CPU Utilization, Memory Usage and Latency are known;
- For each service, the Request Rate needs to be determined such that the Service Temperature difference between each pair of services is less than a given small value;
- Temperature variations are allowed but must be minimized.

To formulate mathematical model of the observed optimization problem, the following notation is used:

- S – set of services

Parameters:

- cpu_j – j-th service's CPU utilization, $j \in S$
- mem_j – j-th service's memory usage, $j \in S$
- lat_j – j-th service's latency, $j \in S$
- wc_j – weight (importance) of the j-th service's CPU utilization, $j \in S$
- wm_j – weight of the j-th service's memory usage, $j \in S$
- wl_j – weight of the j-th service's latency, $j \in S$
- wr_j – weight of the j-th service's request rate, $j \in S$
- t_j – service temperature of the j-th service, $j \in S$
- lr_j, ur_j – lower and upper bound of the j-th service's request rate, $j \in S$
- e - required difference between service temperatures

Variables:

- x_j – j-th service's request rate, $j \in S$
- t_j – Service Temperature of the j-th service, $j \in S$
- d_{ij} – deviation variable that represents deviation of required difference between Services Temperature of i-th and j-th services, $i,j \in S$, $i \neq j$.

For each service $j \in S$, the temperature is calculated as:

$$t_j = wc_j cpu_j + wm_j mem_j + wl_j lat_j + wr_j x_j \quad (1)$$

Thermal Equilibrium between two services $i,j \in S$ is expressed by:

$$|t_i - t_j| \leq e \quad (2)$$

This nonlinear condition can be linearized using the following two conditions:

$$t_i - t_j \leq e \quad (3)$$

$$t_j - t_i \leq e \quad (4)$$

DRAFT VERSION

Mathematical model of the described optimization problem can now be formulated as follows:

$$\min f(x) = \sum_{i \in S} \sum_{j \in S, j \neq i} d_{ij} \tag{5}$$

s.t.

$$t_j = wc_j cpu_j + wm_j mem_j + wl_j lat_j + wr_j x_j,\ j \in S \tag{6}$$

$$t_i - t_j - d_{ij} = 0,\ \ i, j \in S, i \neq j \tag{7}$$

$$lr_j \leq x_j \leq ur_j,\ \ \ j \in S \tag{8}$$

$$x_j \in \mathrm{Z},\ \ \ j \in S \tag{9}$$

The objective (5) represents the sum of all deviations, that should be minimal. The constraint (6) refers to the connection between the variables related to the service temperature and the request rate. The constraint (7) is the linearisation of the thermal equilibrium condition (2), additionally extended with deviations variables. The last two constraints are related to the boundaries of the services' request rates and their integer nature.

4. NUMERICAL RESULTS

Six groups of experiments were conducted to examine the dimensions of the formulated problem that can be solved exactly. The number of services in these experiments were: 5, 10, 15, 20, 25, and 30. For each dimension, ten instances of data was generated. The summarized values of metrics are given in Table 1.

Table 1: Input data for the optimizations

Metrics	Min	Max	Weight
CPU Utilization	45%	100%	1
Memory Usage	45%	100%	1
Latency	20ms	44ms	0.01

In all instances of all six dimensions, the parameter *e* is set to zero, lower and upper bounds of request rate are 1000 and 5000, and the weight of request rate is set to 0.1.

All optimizations were performed on a laptop computer with 11th Gen Intel(R) Core(TM) i5 and 16 GB of RAM. Problems were solved in GLPK modelling environment using combination of four methods: mixed cover cuts, Gomory's mixed integer cuts, MIR (mixed integer rounding) cuts, and clique cuts. The optimization time was limited to 600 seconds.

Table 2 gives the main data about optimization results. In the columns Objective, the average values of the objective, i.e. sum of deviations and their standard deviation are shown. In the column No of optimal solution, the number of instanced which have been solved exactly within time limitation is given. The column Time gives the average optimization time and their standard deviations. The last column (Deviations) contains the average and maximal values of deviations. For instances that could not be solved exactly, the results obtained after 600 seconds are shown.

Table 2: Summarized optimization results

No of services	Objective		No of optimal solution	Time (sec)		Deviations	
	average	SD		average	SD	average	max
5	0.296	0.073	10	8.20	7.01	0.0148	0.056
10	1.369	0.097	10	145.39	107.70	0.0152	0.076
15	3.144	0.325	4	340.53	170.81	0.0150	0.082
20	5.757	0.323	1	442.90		0.0152	0.086
25	8.840	0.648	0			0.0147	0.087
30	13.204	0.878	0			0.0152	0.087

DRAFT VERSION

The first conclusion based on Table 2 is that the possibility of an exact solution decreases rapidly with increasing problem dimensions, i.e. the number of services. Given that the number of services in many software systems may vary depending on the application domain and the level of service granularity, heuristics need to be developed to solve real-world problems. On the other hand, even in cases where an optimal solution was not obtained, the quality of the solutions obtained after 600 seconds is very good. The average and maximum deviation values are similar to those in the optimal solutions. This suggests that in most of these instances an optimal solution was found, but that the search process could not be completed within the given time limit.

As expected, the average value of the objective increases with increasing dimensions, i.e. with increasing the number of services. The relatively small standard deviation of the value of the objective in ten instances indicates robustness, i.e. that the solution is reliable regardless of input data variations.

5. CONCLUSION

Software architecture is one of the key artefacts defined in software design phase. This research has introduced a conceptual mapping between fundamental thermodynamic principles and components of service-oriented systems, leading to the development of a mathematical model for determining the optimal request rate required to maintain thermal equilibrium. These findings enable software architects to design systems in which services are balanced at runtime with respect to their operational load. Additionally, the proposed model supports the detection of imbalance in terms of service temperature, providing a foundation for informed architectural decisions (e.g., allocation of additional resources, deallocation of resources, service composition, service decomposition, etc.).

Future research may explore models where examined software metrics are no longer fixed, allowing dynamic adaptation. In addition, the model can be extended to incorporate a broader set of software metrics to characterize the operational load of a software service. Finally, heuristics will be developed for solving problems of realistic dimensions, given the small dimensions of the problem that can be solved exactly.

DRAFT VERSION

ACKNOWLEDGMENT

This research received no external funding. The authors express their gratitude to the University of Belgrade – Faculty of Organizational Sciences for supporting this research.

LITERATURE

Abeysinghe, A. (2025). Cell-Based Architecture: A Decentralized Reference Architecture for Cloud Native Applications. Available online: https://github.com/wso2/reference-architecture/blob/master/reference-architecture-cell-based.md (Access Date: June 2025).

Alshuqayran, N., Ali, N., & Evans, R. (2016). A systematic mapping study in microservice architecture. *In 2016 IEEE 9th international conference on service-oriented computing and applications (SOCA)*, 44-51. IEEE. https://doi.org/10.1109/SOCA.2016.15.

Cordingly, R., Yu, H., Hoang, V., Sadeghi, Z., Foster, D., Perez, D., Hatchett, R., & Lloyd, W. (2020). The serverless application analytics framework: Enabling design trade-off evaluation for serverless software. In Proceedings of the 2020 Sixth International Workshop on Serverless Computing, 67-72. ACM. https://doi.org/10.1145/3429880.3430103.

de Oliveira Rosa, T., Daniel, J. F. L., Guerra, E. M., & Goldman, A. (2020). A method for architectural trade-off analysis based on patterns: Evaluating microservices structural attributes. *In Proceedings of the European Conference on Pattern Languages of Programs 2020*, 1-8. https://doi.org/10.1145/3424771.3424809.

Ford, N., Richards, M., Sadalage, P., & Dehghani, Z. (2021). *Software Architecture: The Hard Parts*. O'Reilly Media, Inc.

Haigh, M. (2010). Software quality, non-functional software requirements and IT-business alignment. *Software Quality Journal*, 18, 361-385. https://doi.org/10.1007/s11219-010-9098-3.

Hassan, S., & Bahsoon, R. (2016). Microservices and their design trade-offs: A self-adaptive roadmap. In 2016 IEEE International Conference on Services Computing (SCC), 813-818. IEEE. https://doi.org/10.1109/SCC.2016.113.

Hasselbring, W. (2018). Software architecture: Past, present, future. In V. Gruhn, & R. Striemer (Eds.), *The essence of software engineering*, 169-184. Springer. https://doi.org/10.1007/978-3-319-73897-0_10.

ISO/IEC/IEEE 24765:2017. (2017). Systems and software engineering — Vocabulary. *International Organization for Standardization*. Available online: https://www.iso.org (Access Date: June 2025).

ISO/IEC/IEEE 42010:2022. (2022). Software, systems and enterprise — Architecture description. *International Organization for Standardization*. Available online: https://www.iso.org (Access Date: June 2025).

Pacifici, G., Segmuller, W., Spreitzer, M., & Tantawi, A. (2008). CPU demand for web serving: Measurement analysis and dynamic estimation. *Performance Evaluation*, 65(6-7), 531-553. https://doi.org/10.1016/j.peva.2007.12.001.

Papazoglou, M. P., & Van Den Heuvel, W. J. (2007). Service oriented architectures: approaches, technologies and research issues. *The VLDB journal*, 16, 389-415. https://doi.org/10.1007/s00778-007-0044-3.

Shahrad, M., Balkind, J., & Wentzlaff, D. (2019). Architectural implications of function-as-a-service computing. *In Proceedings of the 52nd annual IEEE/ACM international symposium on microarchitecture*, 1063-1075. https://doi.org/10.1145/3352460.3358296.

Shen, K. (2010). Request behavior variations. *ACM SIGARCH Computer Architecture News*, 38(1), 103-116. https://doi.org/10.1145/1735970.1736034.

Walker, J. S. (2017). *Physics (Fifth Edition)*. Pearson, Addison-Wesley.

Washizaki, H. (Ed.) (2024). Guide to the Software Engineering Body of Knowledge v4.0. IEEE Computer Society.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

A HYBRID APPROACH TO DATA MANAGEMENT: INTEGRATING JSON DOCUMENTS INTO RELATIONAL DATABASES

MIHAJLO PAVLOVIĆ[1], TATJANA STOJANOVIĆ[2], SAŠA LAZAREVIĆ[3]

[1] University of Belgrade, Faculty of Organizational Sciences, Belgrade, pav.mihajlo@gmail.com, 0009-0004-3033-8691
[2] University of Belgrade, Faculty of Organizational Sciences, Belgrade, tatjana.stojanovic@fon.bg.ac.rs, 0000-0001-7191-6444
[3] University of Belgrade, Faculty of Organizational Sciences, Belgrade, sasa.lazarevic@fon.bg.ac.rs, 0000-0002-5588-4195

Abstract: *The aim of this paper is to present the theoretical foundation and practical application of the hybrid model using JSON documents in relational databases. The focus is placed on the application of this approach in a Microsoft SQL Server environment, where specific examples illustrate the differences in performance between hybrid, relational and document approaches. A key finding of this paper is the identification of the advantages and challenges associated with the hybrid approach, particularly in terms of query efficiency and modeling flexibility.*

Keywords: *Hybrid database, Relation database, JSON, Data modeling*

1. INTRODUCTION

Throughout the history of software system development, the study of data processing and storage has played a crucial role in the design of software systems. The use of relational databases has represented the standard approach to implementing data storage management. The interpretation of data through tables brought significant innovations to software engineering and enabled easier data management (Codd, 2007).

Despite their widespread use, relational databases exhibit particular drawbacks and structural limitations. The strict schema and table structure in relational databases have proven to be a limitation in systems that require managing flexible data, especially when such data changes dynamically (Kleppmann, 2017).

As a response to these demands, various types of non-relational (NoSQL) databases emerged. The main advantage of these databases is the representation of data using JSON documents (Petković, 2017). The flexible structure of JSON documents has enabled easier data handling, particularly for data with a variable structure.

Relational databases have introduced support for working with JSON documents, thus enabling a new method of data management through a hybrid approach. This approach combines the relational model with JSON documents, allowing the storage of both structured and unstructured data.

2. THEORETICAL BACKGROUND

2.1. Hybrid Approach

Relational databases were originally designed to handle structured data in the form of tables with clearly defined columns and data types. Modern software systems increasingly require processing and storing not only structured but also semi-structured and unstructured data. As a potential architectural solution for such systems, the concept of **Polyglot Persistence** has been developed.

Polyglot persistence is an architectural approach aimed at managing multiple data sources and different types of data storage systems. It involves using various types of databases and technologies within a single system, depending on the nature and requirements of the specific data. For example, relational databases are suitable for transactional data, graph databases can be used for modeling relationships, while document databases are ideal for storing hierarchical data (McMurtry, Oakley, Sharp, Subramanian, & Zhang, 2013). A system designed in this way ensures optimal performance regardless of the volume and variety of data it needs to process.

However, managing and integrating different databases and technologies can be complex and costly. This is where **hybrid databases** come into play, offering a solution that combines relational and non-relational models within the same technology stack.

DRAFT VERSION

Microsoft SQL Server, for instance, supports graph database capabilities, enabling data modeling through nodes and edges, which simplifies the representation of relationships between entities. This feature is particularly useful in applications like social networks and recommendation systems.

Additionally, SQL Server supports working with JSON and XML documents, providing a balance between efficiency and system complexity. Importantly, the hybrid approach is implemented within a single technology, eliminating the need for external NoSQL solutions.

With these enhancements, SQL Server has become a powerful tool for implementing hybrid models, enabling the development of systems that meet the requirements of both traditional and modern architectures without leaving the well-established and mature SQL ecosystem.

2.2. JSON Document

The main advantage of non-relational databases is their ability to represent data using the **JSON structure**. JSON (JavaScript Object Notation) is a lightweight data-interchange format used to represent structured information in a way that is easy to read and write for humans. The fundamental structure of a JSON document consists of key-value pairs (Pezoa, Reutter, Suárez, Ugarte, & Vrgoč, 2016).

This data format resembles the structure of classes and their attributes in object-oriented programming languages. The flexibility of the JSON schema allows the storage of heterogeneous and unstructured data even within relational databases. Using JSON documents in this way serves as a method of adapting strict data schemas to accommodate more flexible and dynamic data structures.

2.3. JSON in Sql Server

Since SQL Server 2016, Microsoft's database system has introduced significant support for working with JSON documents, making it a competitive option for handling semi-structured and unstructured data. A JSON document is stored as plain text within NVARCHAR or VARCHAR columns, enabling seamless integration with existing relational model structures (Microsoft, 2024).

```
DECLARE @json NVARCHAR(MAX);
SET @json = '{"info": {"address"
                                [{"city": "Belgrade"},
                                 {"city": "Paris"},
                                 {"city":"Madrid"}]}}';
```

This version introduced numerous functions that facilitate efficient storage and processing of JSON documents (Microsoft, 2024).

- ***ISJSON –*** the function checks whether the input text is in a valid JSON format.
- ***JSON_VALUE –*** this function extracts the value of a specified attribute from a JSON document. It takes a path parameter that allows direct access to attribute values.
- ***JSON_OBJECT –*** this function takes key-value pairs as input and creates a JSON object***.***
- ***JSON_ARRAY –*** this function creates a JSON array from the provided values, which may be of different data types.
- ***JSON_QUERY –*** this function returns an object or array from a JSON text. The difference between JSON_QUERY and JSON_VALUE lies in their return types: JSON_VALUE returns scalar values, while JSON_QUERY returns full JSON objects or arrays***.***
- ***JSON_MODIFY –*** This function updates the value of an attribute within a JSON document and returns the updated document. It can also add new attributes or remove existing ones. Moreover, it allows changes within JSON arrays.
- ***OPENJSON –*** enables tabular representation of a JSON document. Each object becomes a row, while key-value pairs are represented as columns. This function is especially useful when working with structured or semi-structured data, as it allows the use of standard relational operations on JSON.
- ***FORJSON –*** this clause formats query results as JSON text. It offloads the responsibility of generating JSON from the application layer to the database engine. The clause supports two modes: AUTO for automatic structure and PATH for custom output formatting.

DRAFT VERSION

3. PRACTICAL IMPLEMENTATION

This section of the paper presents the practical implementation of basic operations (SELECT, INSERT, and UPDATE) using three different approaches: relational, hybrid, and document. The experiment uses a domain model with a focus on hierarchical inheritance.

Figure 1: Domain model

In the relational model, the Table per Type (TPT) strategy separates data into related tables to avoid NULLs and represent inheritance. The hybrid model stores parent entities in relational tables and embedding subtypes data as JSON documents. The implementation of the relational and hybrid models was carried out using SQL Server technology. In the document model, each standalone entity is stored as a separate document in MongoDB, simplifying handling of diverse data.

Table 1: Implementaion cross approaches

Relation	Hybrid	Document

CREATE TABLE Employee (
id BIGINT PRIMARY KEY,
firstname NVARCHAR(255) NOT NULL,
lastname NVARCHAR(255) NOT NULL,
email NVARCHAR(255) NOT NULL,
birthday DATE NOT NULL,
title NVARCHAR(255) NOT NULL,
phone NVARCHAR(50) NOT NULL);
CREATE TABLE Developer (
id BIGINT PRIMARY KEY
FOREIGN KEY REFERENCES
employee(id) ON DELETE CASCADE,
seniority NVARCHAR(255) NOT NULL,
yearsOfExperience INT NOT NULL,
isRemote BIT NOT NULL);
CREATE TABLE Manager (
id BIGINT PRIMARY KEY
FOREIGN KEY REFERENCES
employee(id) ON DELETE CASCADE,
department NVARCHAR(255) NOT NULL,
realisedProject INT NOT NULL,
method NVARCHAR(255) NOT NULL);
CREATE TABLE SoftwareDeveloper (
id BIGINT PRIMARY KEY
FOREIGN KEY REFERENCES
developer(id) ON DELETE CASCADE,
programmingLanguage NVARCHAR(255) NOT NULL,
ide NVARCHAR(255) NOT NULL,
isfullstack BIT NOT NULL);
	CREATE TABLE Employee (
	id BIGINT PRIMARY KEY,
	firstname NVARCHAR(255) NOT NULL,
	lastname NVARCHAR(255) NOT NULL,
	email NVARCHAR(255) NOT NULL,
	birthday DATE NOT NULL,
	title NVARCHAR(255) NOT NULL,
	phone NVARCHAR(50) NOT NULL,
	manager NVARCHAR(MAX) NULL);
	
	CREATE TABLE Developer (
	id BIGINT PRIMARY KEY FOREIGN KEY
	REFERENCES employee(id) ON DELETE
	CASCADE,
	seniority NVARCHAR(255) NOT NULL,
	yearsOfExperience INT NOT NULL,
	isRemote BIT NOT NULL,
	softwareDeveloper NVARCHAR(MAX)
	NULL);
		{
		"_id": ObjectId("ManagerId"),
		"firstname": "String",
		"lastname": "String",
		"email": "String",
		"birthday": "Date",
		"title": "String",
		"phone": "String",
		"department": "String",
		"realisedProject": "Number",
		"method": "String"
		}
		
		{
		"_id": ObjectId("SoftwareDeveloperId"),
		"firstname": "String",
		"lastname": "String",
		"email": "String",
		"birthday": "Date",
		"title": "String",
		"phone": "String",
		"seniority": "String",
		"yearsOfExperience": "Number",
		"isRemote": true,
		"programmingLanguage": "String",
		"ide": "String",
		"isFullStack": true
		}

DRAFT VERSION

In the relational approach, data is inserted into two tables—one for general employee attributes and another for manager-specific details. The hybrid approach uses a single table with general attributes as columns and manager-specific data stored as a JSON document. In the document approach, all data is stored together in a single document.

Table 2: Implementation INSERT query

Approach	Query
Relation	INSERT INTO Employee (id, firstname, lastname, email, birthday, title, phone) VALUES (@Id, @FirstName, @LastName, @Email, @Birthday, @Title, @Phone); INSERT INTO manager (id, department, realisedproject, method) VALUES (@Id, @Department, @RealisedProject, @Method);
Hybrid	INSERT INTO Employee (id, firstname, lastname, email, birthday, title, phone,manager) VALUES (@Id, @FirstName, @LastName, @Email, @Birthday, @Title, @Phone,@Manager);

Document	public void InsertManager(ManagerModel manager)
	{
	manager.Role = "Manager";
	_employeesCollection.InsertOne(manager);
	}

The SELECT query retrieves all managers who are younger than thirty years old and use the Agile methodology. The results are sorted by date of birth in ascending order.

Table 3: Implementation SELECT query

Approach	Query
Relation	SELECT e.id, e.firstname, e.lastname, e.email, e.birthday, e.title, e.phone, m.department, m.realisedProject, m.method FROM Employee e INNER JOIN Manager m ON e.id = m.id WHERE DATEDIFF(YEAR, e.birthday, GETDATE()) < 30 AND m.method = 'Agile' ORDER BY e.birthday ASC;
Hybrid	SELECT e.id, e.firstname, e.lastname, e.email, e.birthday, e.title, e.phone, m.department, m.realisedProject, m.method FROM Employee e CROSS APPLY OPENJSON(manager) WITH(Department NVARCHAR(100) '$.Department', RealisedProject INT '$.RealisedProject', Method NVARCHAR(20) '$.Method')AS m WHERE DATEDIFF(YEAR, e.birthday, GETDATE()) < 30 AND m.method = 'Agile' ORDER BY e.birthday ASC;

Document	public List<ManagerModel> GetManagersYoungerThan30AndAgileMethodSorted()
	{
	var filterForManagers = Builders<ManagerModel>.Filter.And(
	Builders<ManagerModel>.Filter.Eq("Role", "Manager"),
	Builders<ManagerModel>.Filter.Gt(m=>m.BirthDay,DateTime.UtcNow.AddYears(-30)),
	Builders<ManagerModel>.Filter.Eq(m=>m.Method,"Agile")
	);
	var foundManagers = ManagersCollection
	.Find(filterForManagers)
	.Project<ManagerModel>(ManagerProjection)
	.ToList();
	return foundManagers;
	}

This query sets the methodology to Lean for every manager who belongs to the IT or Logistics department and is between forty and fifty years old. In the relational approach, the operation is performed across two tables, while in the hybrid approach, the condition applies to both the table and the embedded document.

DRAFT VERSION

Table 4: Implementation UPDATE query

Approach	Query
Relation	UPDATE Manager SET method = 'Lean' WHERE department IN ('IT','Logistics') AND id IN (SELECT e.id FROM Employee e WHERE DATEDIFF(YEAR, e.birthday, GETDATE()) BETWEEN 40 AND 50);
Hybrid	UPDATE Employee SET manager = JSON_MODIFY(manager,'$.Method','Lean') WHERE JSON_VALUE(manager,'$.Department') IN ('IT','Logistics') AND DATEDIFF(YEAR, birthday, GETDATE()) BETWEEN 40 AND 50;

Document	public void UpdateMethodByYearsAndDepartment()
	{
	var filterManager = Builders<ManagerModel>.Filter.And(
	Builders<ManagerModel>.Filter.In(m => m.Department,new []{ "IT", "Logistics" }),
	Builders<ManagerModel>.Filter.Gt(m=>m.BirthDay,DateTime.UtcNow.AddYears(-50)),
	Builders<ManagerModel>.Filter.Lt(m=>m.BirthDay,DateTime.UtcNow.AddYears(-40))
	);
	var updateMethod = Builders<ManagerModel>.Update.Set(m => m.Method, "Lean");
	
	ManagersCollection.UpdateMany(filterManager, updateMethod);
	}

4. ANALYSIS OF EXPERIMENTAL RESULTS

This chapter presents the results of an experimental comparison of three distinct approaches—relational, hybrid, and non-relational—in terms of the execution of fundamental operations: SELECT, INSERT, and UPDATE. Execution time was measured for each operation across varying data volumes (5,000, 50,000, and 500,000 records), using the *BenchmarkDotNet* tool. This benchmarking library, implemented in C#, ensures high-precision and statistically reliable results. All tests were performed under identical hardware conditions, with the following specifications: Intel i7 10750h 2.60GHz processor, 16 GB DDR4 ram memory, and Windows 10 Pro operating system. The execution times are reported in milliseconds.

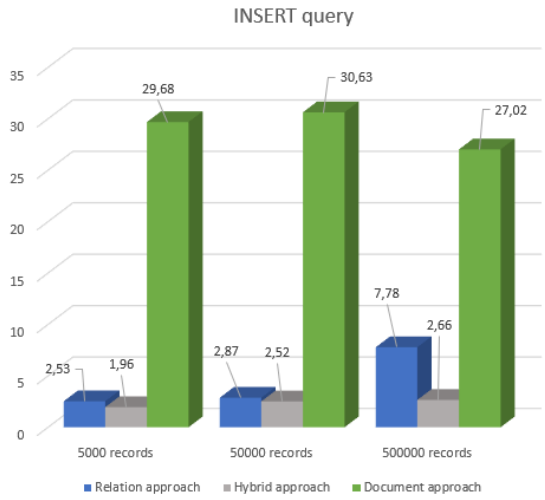


Figure 2: Insert Operation Performance

The results for the INSERT operation show that the Document approach consistently has the slowest data insertion times. Its execution time is significantly higher compared to the relational and hybrid approaches. The hybrid approach performs best, with the relational approach closely following. This can be attributed to MongoDB's document handling and limited optimization for bulk inserts, whereas relational databases often leverage batch inserts and transactional management, resulting in faster data insertion.

DRAFT VERSION

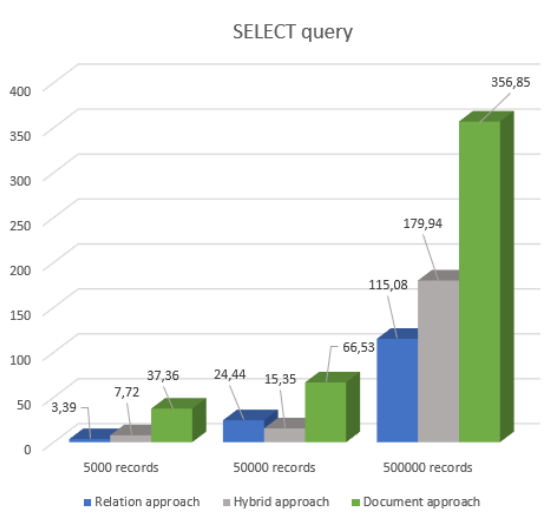


Figure 2: SELECT Operation Performance

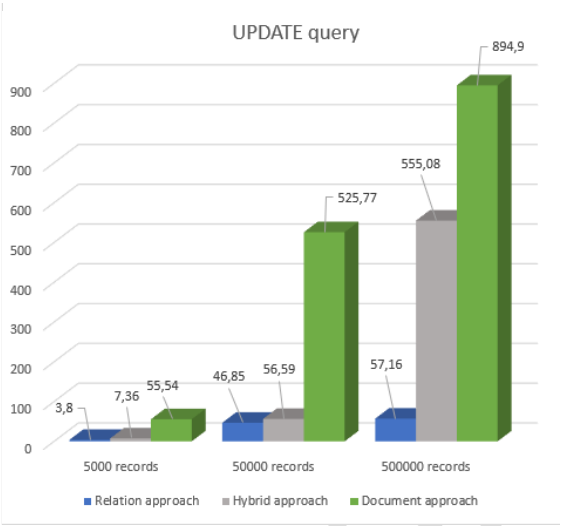


Figure 3: UPDATE Operation Performance

Based on the SELECT query results, the relational approach shows the best performance across all data volumes. Document approach is significantly slower, especially with larger datasets, while the hybrid approach performs moderately but leans closer to the relational model. This is largely due to SQL's efficient indexing and query optimization, unlike MongoDB's document structure, which is less suited for complex queries.

Regarding the UPDATE operation, document model performs the slowest, particularly with larger datasets. The hybrid model is slightly better, while the relational approach is the fastest and most consistent due to its optimized mechanisms for updating specific values, unlike MongoDB which often rewrites entire documents.

5. CONCLUSION

This research demonstrates that relational, non-relational, and hybrid databases each offer distinct advantages and disadvantages depending on query complexity and data structure. The relational model provides stability and efficiency when working with structured data and complex queries, while the document approach allows greater flexibility for dynamic and unstructured data.

The hybrid model, combining features of both systems, proved to be an optimal solution for scenarios requiring fast execution of simple queries and data insertion. However, its performance decreases with more complex operations, especially those involving grouping, sorting, and data updates.

In conclusion, the choice of a suitable database should be guided by the specific needs of the application and the nature of the data. The hybrid approach represents a promising solution that, with further optimization, can significantly contribute to the development of efficient and sustainable software systems.

Future research could investigate the security implications of each approach, focusing on data protection, access control, and compliance with relevant regulations.

LITERATURE

Codd, E. F. (2007). Relational database: A practical foundation for productivity. *Communications of the ACM, 25*(2), 109–117. https://doi.org/10.1145/358396.358400

Kleppmann, M. (2017). *Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems*. O'Reilly Media.

McMurtry, D., Oakley, A., Sharp, J., Subramanian, M., & Zhang, H. (2013). *Data access for highly-scalable solutions: Using SQL, NoSQL, and polyglot persistence*. Microsoft Developer Guidance.

Microsoft. (2024). *JSON data in SQL Server*. https://learn.microsoft.com/en-us/sql/relational-databases/json/json-data-sql-server?view=sql-server-ver16

Pezoa, F., Reutter, J. L., Suárez, F., Ugarte, M., & Vrgoč, D. (2016). Foundations of JSON schema. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)* (pp. 263–273). Association for Computing Machinery. https://doi.org/10.1145/2872427.2883029

Petković, D. (2017). *JSON integration in relational database systems*. https://www.academia.edu/download/73604553/JSON_Integration_in_Relational_Database_Systems.pdf

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Review article

A SYSTEMATIC REVIEW OF VECTOR DATABASE USE IN RETRIEVAL-AUGMENTED GENERATION FOR LLM-BASED EDUCATIONAL PLATFORMS

FILIP STAMENKOVIĆ[1], JELICA STANOJEVIĆ[2], MIROSLAV MINOVIĆ[3]

[1] University of Belgrade, Faculty of Organizational Sciences, Belgrade, filip.stamenkovic@fon.bg.ac.rs, 0009-0008-4575-5892
[2] University of Belgrade, Faculty of Organizational Sciences, Belgrade, jelica.stanojevic@fon.bg.ac.rs, 0000-0002-0570-0167
[3] University of Belgrade, Faculty of Organizational Sciences, Belgrade, miroslav.minovic@fon.bg.ac.rs, 0000-0002-4270-759

Abstract: *This systematic review explores the use of vector databases in Retrieval-Augmented Generation (RAG) for educational platforms based on large language models (LLMs). As RAG becomes a promising approach to enhance the contextual accuracy of LLM outputs by retrieving relevant content, vector databases serve as a core component for storing and retrieving embedded educational materials. This review is comprised of 9 studies from 2023 to 2025, focusing on use cases in higher education, including domain-specific applications and chatbots for student and educator support. Findings show diverse choices of vector stores, such as FAISS, Chroma, Qdrant, Weaviate, Milvus, Vectara, MongoDB and Postgres with pgVector, often combined with orchestration frameworks like LangChain or LlamaIndex. The reporting on embedding models, orchestration frameworks and system architecture is inconsistent, limiting the comparability of studies and reducing confidence in synthesizing performance trends, which impacts the reliability of conclusions drawn from the review. The findings provide a reference point for researchers and developers creating context-aware, LLM-based educational platforms, and suggest future research directions including performance benchmarking, model transparency, and evaluating learning outcomes.*

Keywords: *Retrieval-Augmented Generation, Large Language Models, Vector Databases, LLMs in Education.*

DRAFT VERSION

1. INTRODUCTION

Generative AI, especially large language models (LLMs), can support both students and teachers by making complex material easier to understand (Laato et al., 2023).

One useful method that enhances LLM performance by adding relevant outside information during the response process is Retrieval-Augmented Generation (RAG). RAG retrieves data from a database using vector search and gives it to the model as context (Gao et al., 2024). This makes the model's answers more accurate and useful, especially in education, managing data like lecture notes, textbooks, or slides (Mullins et al., 2024).

Vector databases play an important role in these systems. They store content as numerical embeddings, which helps the system better understand what the user is asking. This leads to better answers and also cuts down on hallucinations, made-up or wrong responses that can happen when LLMs rely only on their training data (Jing et al., 2024).

Different systems use different tools, depending on their goals. Educational systems using RAG vary depending on their purpose. Some are made as chatbots that help students by answering their questions and giving support in learning (Wölfel et al., 2023), while others are aimed at educators, helping them generate exam questions, prepare teaching materials, or analyze course content (Hennekeuser et al., 2024; Nikolovski et al., 2025). In this review, educational platforms are defined as tools and environments primarily aimed at higher education.

Even though the number of educational applications which utilize RAG is increasing, there is limited research that systematically examines how vector databases are integrated and applied in these systems. To better understand how these systems work, this review explores two main questions:

- RQ1: Which vector databases are used in educational RAG-based systems?
- RQ2: How are these databases used in different studies, and what kinds of content and tasks do they support?

By analyzing recent research, this paper provides a clearer picture of how RAG and vector databases are used in education and identifies directions for future research.

In the following sections, the paper describes the methodology used for selecting and analyzing relevant studies (Section 2), presents key findings on the use of vector databases in educational RAG systems (Section 3), discusses the findings (Section 4), and concludes with a summary of main insights and application possibilities (Section 5).

2. METHODS

The literature research was conducted using two major scientific databases: Scopus and Web of Science. These databases were selected due to their wide use in systematic reviews and their extensive coverage of high-quality papers in the chosen fields. The databases feature a reliable advanced search engine based on user-defined criteria.

The goal of the review was to identify studies that explore the use of Retrieval-Augmented Generation (RAG) in combination with large language models (LLMs) within educational platforms, with a particular focus on the selection and use of vector databases that support the RAG technique.

To identify relevant studies, queries were constructed for each database using logical operators (AND, OR) and truncation symbols (*). The search fields were limited to title, abstract, and keywords. The queries targeted papers that mention key concepts related to: Retrieval-Augmented Generation, Large Language Models, Platform or API integration, and educational context. The following search queries were used, tailored to the syntax of each database:

- **Scopus:** TITLE-ABS-KEY("retrieval-augmented generation" OR "RAG") AND TITLE-ABS-KEY("large language model" OR "LLM") AND TITLE-ABS-KEY("LLM-based" OR "LLM-powered" OR "API" OR "application programming interface" OR "chatbot*") AND TITLE-ABS-KEY("education" OR "educational platform*").
- **Web of Science**: TS=("retrieval-augmented generation" OR "RAG") AND TS=("large language model" OR "LLM") AND TS=("LLM-based" OR "LLM-powered" OR "API" OR "application programming interface" OR "chatbot*") AND TS=("education" OR "educational platform*").

Although the focus of this review is on the use of vector databases in RAG LLM-based systems, terms such as “vector database” or “vector store” were not included in the search queries. Preliminary searches including the terms “vector database” or “vector store” were conducted to test their impact on retrieval. When these terms were added to the Web of Science query, only 2 studies were returned, fully overlapping with results already captured by the broader query. In Scopus, adding these terms yielded 6 studies, again fully overlapping with results already captured by the broader query. This confirmed that including these terms unnecessarily narrows the search without contributing additional relevant studies. The search was conducted on June 6, 2025.

The study selection process followed the PRISMA 2020 guidelines (Page et al., 2021). After conducting the database search, duplicate records were removed. The remaining papers were then screened by examining their titles, keywords, and abstracts. Studies that were not relevant were excluded. Then, the full texts of the selected papers were read. More papers were excluded if they did not use RAG, did not include an actual LLM-based system, or did not give out vector database implementation details. Ultimately, only studies that met all inclusion criteria were retained. Two reviewers independently screened all titles, abstracts, and full texts, resolving any disagreements through discussion to reach consensus. Although this dual-review process reduces the risk of selection bias, inter-rater reliability statistics were not calculated, which limits the ability to quantitatively assess the consistency of reviewer judgments.

3. RESULTS

After finding the initial 70 papers and then removing duplicates, a total of 55 unique studies were identified. The titles, keywords and abstracts of these papers were then screened, and 25 studies were excluded for not meeting basic relevance criteria.

Of the 30 remaining studies that were sought for retrieval, 8 of them were not retrieved successfully, having resulted in a reading and review of the 22 retrieved papers. During the reading phase, an additional 12 papers were excluded. The number of excluded papers for each exclusion reasons were: absence of RAG use in the implementation (n = 1), lack of actual implementation of an LLM-based system (n = 3) and absence of vector database implementation details (n = 8). The full selection process is summarized in the PRISMA 2020 flow diagram (Page et al., 2021) shown in Figure 1.

DRAFT VERSION

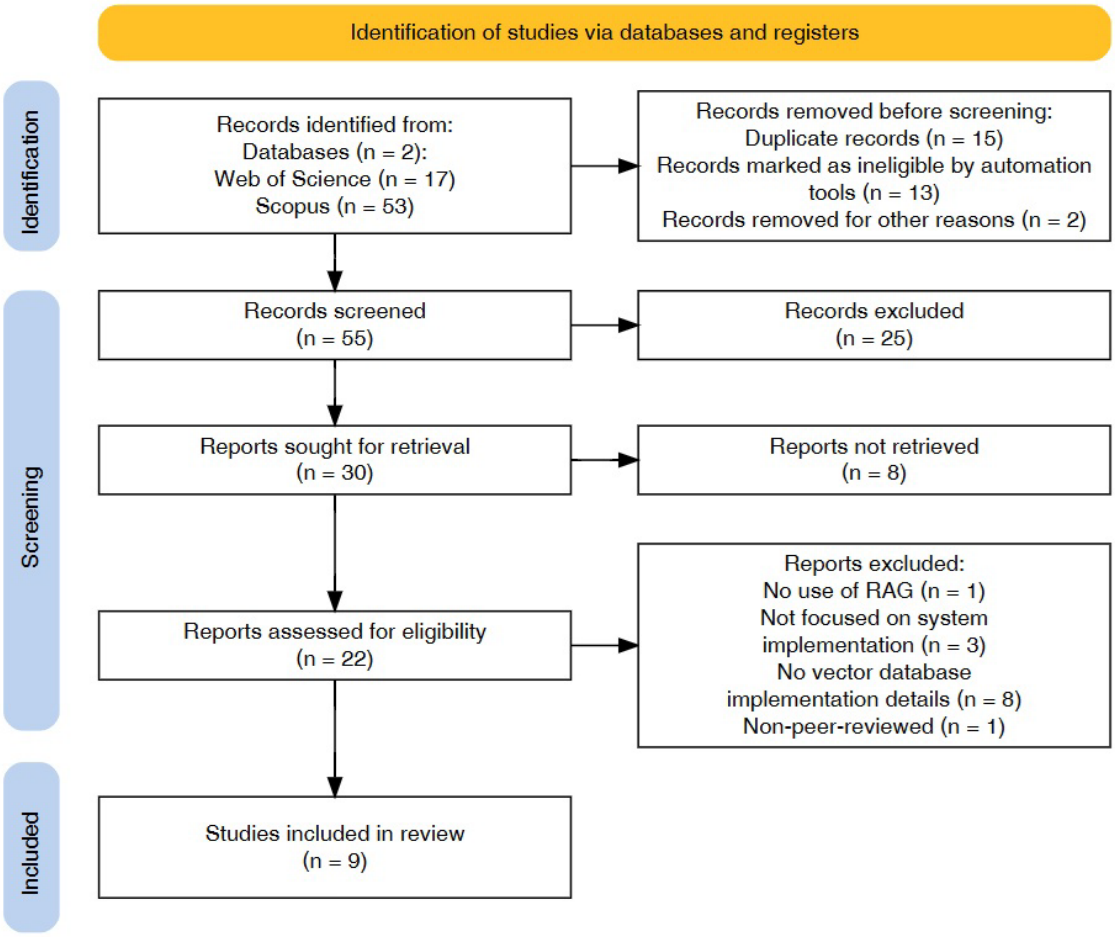


Figure 1: PRISMA 2020 flow diagram

The review included 9 studies published between 2023 and 2025, with all of them focused on the use of Retrieval-Augmented Generation (RAG) with large language models (LLMs) in educational settings. The selected papers mainly describe applications and tools designed to assist students and educators in higher education through improved content retrieval and generation. Below is a summary of the key characteristics of the papers which were included:

Table 1: Summary of the included papers

Paper	Education domain	LLM	Vector database	Additional framework	Embedding model
Advancing AI in Higher Education: A Comparative Study of Large Language Model-Based Agents for Exam Question Generation, Improvement, and Evaluation (Nikolovski et al., 2025)	General education assistance for educators	Gemini 2.0	Qdrant	LangChain	Not specified
Towards an AI tutor for undergraduate geotechnical engineering: a comparative study of evaluating the efficiency of large language model application programming interfaces (Tophel et al., 2024)	Geotechnical engineering	GPT-4 Llama-3	FAISS	LangChain	OpenAI's embedding model
Enlarged Education - Exploring the Use of Generative AI to Support Lecturing in Higher Education (Hennekeuser et al., 2024)	General education assistance for educators	GPT-4	Milvus	Not specified	OpenAI's text-embedding-ada-002 model
ChatPapers: An AI Chatbot for Interacting with Academic Research (Dean et al., 2023)	Academic research	GPT-4	Postgres with pgVector	LangChain	all-MiniLM-L6-V2 model from Hugging Face

DRAFT

An LLM-Driven Chatbot in Higher Education for Databases and Information Systems (Neumann et al., 2025)	Computer science	GPT-4	Weaviate	LangChain LlamaIndex	OpenAI's embedding model
Knowledge-Based and Generative-AI-Driven Pedagogical Conversational Agents: A Comparative Study of Grice's Cooperative Principles and Trust (Wölfel et al., 2023)	General education assistance for students	GPT-4 GPT-3.5	Chroma	LangChain	Not specified
MerryQuery: A Trustworthy LLM-Powered Tool Providing Personalized Support for Educators and Students (Tabarsi et al., 2025)	General education assistance for students and educators	GPT-4o	MongoDB	LangChain	Not specified
Enhancing Engineering Education Through LLM-Driven Adaptive Quiz Generation: A RAG-Based Approach (Gopi et al., 2024)	Engineering education and quiz generation	GPT-4	Vectara	Not specified	Not specified
MEDCO: Medical Education Copilots Based on a Multi-agent Framework (Wei et al., 2024)	Medical education	GPT-3.5 Claude 3.5 Sonnet	Chroma	Not specified	OpenAI's embedding model

The described systems vary mainly in their purpose. Most studies present chatbots that help students learn different subjects, while others focus on tools for generating exam questions or academic explanations. These systems use embedding techniques to turn educational content into vectors, making it easier to find relevant information based on meaning, not just keywords. The reviewed studies used a range of vector databases, with integration methods differing across implementations.

Vector databases in the nine studies were: FAISS (1), Chroma (2), Qdrant (1), Weaviate (1), Milvus (1), Vectara (1), MongoDB with vector search (1), and Postgres with pgVector (1). All but Vectara were open-source. LangChain was used in 6 studies, and LlamaIndex in 1 study (combined with LangChain). Embedding models were inconsistently reported: 2 studies used OpenAI's text-embedding-ada-002, 2 unspecified OpenAI models, 1 used all-MiniLM-L6-v2, and 4 studies did not report the model.

In systems built for educators, Nikolovski et al. (2025) used Qdrant with LangChain to store and access data from lecture slides, notes, code examples, textbook excerpts, and lab guides to help create and check exam questions. The PDFs were turned into markdown files, split into chunks, and embedded for quick search. However, the authors did not provide further explanation about why Qdrant was chosen over other available options. Hennekeuser et al. (2024) built a similar assistant using Milvus and OpenAI's text-embedding-ada-002 model. It helped teachers access content from PDFs for course preparation, like books and scripts. The authors write that Milvus's support for multiple similarity search functions makes it well-suited for efficient retrieval in this educational setting.

For students, vector search changed how chatbots work, fitting them closer to student needs. Wölfel et al. (2023) used Chroma and LangChain to build a system that worked with lecture slides and transcripts, helping students through more responsive conversations. Chroma was used as vector database due to its efficiency in leveraging embeddings to represent documents as vectors and retrieving relevant information based on semantic similarity. However, the authors did not explain why Chroma was chosen over other options. Tabarsi et al. (2025) used MongoDB with vector search to support both students and teachers, letting them retrieve PDFs with text, tables, and images. MongoDB was also used to store interaction logs. This choice simplified system architecture by combining transactional data and vector search in one platform.

In engineering education, vector search was key for retrieving specialized knowledge. Tophel et al. (2024) used FAISS and LangChain to help create a chatbot in geotechnical engineering, working with files in PDF, DOCX, and TXT formats. OpenAI embeddings helped store and find the right information. FAISS was used because of its proven efficiency in large-scale similarity search, being well suited for technical domains like engineering. Gopi et al. (2024) used Vectara and the MathVista dataset to make quiz questions for engineering students, based on subject-specific reasoning. Vectara, an API platform, simplified data ingestion and retrieval with optimized defaults for chunking and preprocessing.

In computer science, Neumann et al. (2025) built a chatbot using Weaviate, LangChain, and LlamaIndex to help students access learning content. PDF materials were embedded with OpenAI's model to get accurate answers. The authors did not provide further details on the selection of Weaviate for the system. Dean et al. (2023) built a research assistant using Postgres with pgVector and LangChain. It used Hugging Face's all-MiniLM-L6-V2 model to store and find papers and metadata for computer science research. The system

DRAFT VERSION

leveraged vector search to quickly retrieve relevant research papers based on user queries, improving the efficiency of literature review processes. PgVector allowed the efficient storage and querying of both vectorized data and associated metadata, such as paper titles, authors, and publication details.

In medical education, Wei et al. (2024) used Chroma for efficient storage and rapid queries in a multi-agent tutoring system, with roles like patient, radiologist, and expert, to deliver content across specialties. Data was embedded using OpenAI's model, and vector search pulled relevant cases from the MVME dataset. This helped create a more reliable learning experience for medical students.

4. DISCUSSION

The reviewed studies show various methods for using vector search in educational RAG-based systems, providing answers to both RQ1 and RQ2. All systems rely on vector-based retrieval, but vary in their choice of vector database, embedding model, and orchestration tools.

Used vector stores include FAISS, Chroma, Qdrant, Weaviate, Milvus, Vectara, MongoDB and Postgres with pgVector. These databases stored different data types including lecture slides, research papers, textbooks, and course materials, depending on the system's domain and goals. They were often combined with orchestration frameworks like LangChain or LlamaIndex. This distribution indicates a clear preference for open-source solutions and highlights a major gap in reporting: 4 out of 9 of studies omitted embedding model details. Such omissions reduce reproducibility and complicate meaningful comparisons of retrieval performance across implementations. While OpenAI's embedding models are most common, alternatives like all-MiniLM-L6-V2 are also shown as viable.

A notable gap in the reviewed studies is the lack of explanation for why specific vector databases were chosen. Most papers only mention which database was used and note that it worked efficiently, without discussing the reasons behind the choice. Some studies, like those using Postgres or MongoDB, stored both vectors and metadata or interaction logs in one system instead of using a dedicated vector database. This simplifies architecture but raises questions about trade-offs compared to specialized solutions. Institutional preference may also influence these decisions. Future work should explain database selection more clearly to enable better comparisons of effectiveness across contexts.

For practitioners, this review suggests that RAG systems in education can be adapted to different institutional needs. Future work should benchmark vector stores and embedding models using metrics like retrieval accuracy, latency, scalability, and cost-performance trade-offs, comparing open-source solutions with cloud-hosted proprietary databases like Vectara. Additionally, studies should assess how retrieval design impacts student learning outcomes, trust in AI, and engagement.

Ethical considerations are essential when discussing AI in education. Huang (2023) highlights the importance of student data privacy, risks of unauthorized data use, and the need for transparency in AI processing of sensitive information. They stress the necessity of accountability mechanisms from institutions and developers. While this review focuses on technical aspects, these ethical concerns are crucial to ensure responsible AI use in education, respecting students' privacy rights.

One methodological limitation of this review is its reliance solely on Scopus and Web of Science for literature retrieval. While these databases offer high-quality content, expanding the search to include other sources, such as Google Scholar, could uncover additional implementations and broaden the understanding of RAG-based systems in education.

5. CONCLUSION

This review explores how vector databases are used in educational RAG systems, emphasizing their adaptability to various educational needs, whether for student support, teacher assistance, or subject-specific focus. This adaptability allows for the creation of RAG systems tailored to different educational goals, although more details are needed to fully understand the potential of these systems.

A major limitation is the lack of detailed technical descriptions, especially regarding embedding models and system architecture. While selected studies mention which vector databases were used, they rarely explain why they were chosen, and several omit embedding model details entirely. This omission is significant, as it limits reproducibility, hinders meaningful comparisons of system performance, and reduces confidence in the broader conclusions of the review. This review highlights that the lack of standardized reporting not only complicates synthesis but also makes it difficult to determine which combinations of databases and models work best for specific educational tasks.

Despite these challenges, the findings provide a useful starting point for designing more context-aware educational systems. Future research could focus on benchmarking different vector stores and models,

DRAFT VERSION

exploring their impact on learning outcomes, and establishing clearer reporting practices to support more stable comparisons.

ACKNOWLEDGMENT

This research received no external funding.

LITERATURE

Dean, M., Bond, R. R., McTear, M. F., & Mulvenna, M. D. (2023). ChatPapers: An AI Chatbot for Interacting with Academic Research. *2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, 1–7. https://doi.org/10.1109/AICS60730.2023.10470521

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey* (No. arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

Gopi, S., Sreekanth, D., & Dehbozorgi, N. (2024). Enhancing Engineering Education Through LLM-Driven Adaptive Quiz Generation: A RAG-Based Approach. *FIE 2024 Conference Proceedings*, 1–8. https://doi.org/10.1109/FIE61694.2024.10893146

Hennekeuser, D., Vaziri, D. D., Golchinfar, D., Schreiber, D., & Stevens, G. (2024). Enlarged Education – Exploring the Use of Generative AI to Support Lecturing in Higher Education. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00424-y

Huang, L. (2023). Ethics of Artificial Intelligence in Education: Student Privacy and Data Protection. *Science Insights Education Frontiers*, *16*(2), 2577–2587. https://doi.org/10.15354/sief.23.re202

Jing, Z., Su, Y., & Han, Y. (2024). *When Large Language Models Meet Vector Databases: A Survey* (No. arXiv:2402.01763). arXiv. https://doi.org/10.48550/arXiv.2402.01763

Laato, S., Morschheuser, B., Hamari, J., & Björne, J. (2023). AI-Assisted Learning with ChatGPT and Large Language Models: Implications for Higher Education. *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, 226–230. https://doi.org/10.1109/ICALT58122.2023.00072

Mullins, E. A., Portillo, A., Ruiz-Rohena, K., & Piplai, A. (2024). *Enhancing classroom teaching with LLMs and RAG* (No. arXiv:2411.04341). arXiv. https://doi.org/10.48550/arXiv.2411.04341

Neumann, A. T., Yin, Y., Sowe, S., Decker, S., & Jarke, M. (2025). An LLM-Driven Chatbot in Higher Education for Databases and Information Systems. *IEEE Transactions on Education*, *68*(1), 103–116. https://doi.org/10.1109/TE.2024.3467912

Nikolovski, V., Trajanov, D., & Chorbev, I. (2025). Advancing AI in Higher Education: A Comparative Study of Large Language Model-Based Agents for Exam Question Generation, Improvement, and Evaluation. *Algorithms*, *18*(3), 144. https://doi.org/10.3390/a18030144

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71

Tabarsi, B., Basarkar, A., Liu, X., Xu, D. (DK), & Barnes, T. (2025). MerryQuery: A Trustworthy LLM-Powered Tool Providing Personalized Support for Educators and Students. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(28), 29700–29702. https://doi.org/10.1609/aaai.v39i28.35372

Tophel, A., Chen, L., Hettiyadura, U., & Kodikara, J. (2024). *Towards an AI Tutor for Undergraduate Geotechnical Engineering: A Comparative Study of Evaluating the Efficiency of Large Language Model Application Programming Interfaces*. In Review. https://doi.org/10.21203/rs.3.rs-4658661/v1

Wei, H., Qiu, J., Yu, H., & Yuan, W. (2024). *MEDCO: Medical Education Copilots Based on A Multi-Agent Framework* (No. arXiv:2408.12496). arXiv. https://doi.org/10.48550/arXiv.2408.12496

Wölfel, M., Shirzad, M. B., Reich, A., & Anderer, K. (2023). Knowledge-Based and Generative-AI-Driven Pedagogical Conversational Agents: A Comparative Study of Grice's Cooperative Principles and Trust. *Big Data and Cognitive Computing*, *8*(1), 2. https://doi.org/10.3390/bdcc8010002

DRAFT VERSION

LOGISTICS AND SUPPLY CHAINS

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

PRODUCT-FORM SOLUTION OF A QUEUEING INVENTORY WITH TWO MODES OF PRODUCTION

ABDUL ROF V[1], ACHYUTHA KRISHNAMOORTHY[2]

[1] Department of Mathematics, Korambayil Ahamed Haji Memorial Unity Women's College, Manjeri, Malappuram, Kerala-676122, India, abdulrof@unitywomenscollege.ac.in, 0000-0002-7479-3651
[2] Centre for Research in Mathematics,CMS College, Kottayam, Kerala-686001, India, achyuthacusat@gmail.com, 0000-0002-5911-1325

Abstract: *The system under consideration is a queueing inventory system, where customers come following a Poisson process with rate λ, and inventory items are served according to an exponential distribution with parameter μ. Production of items commences when the inventory level falls below a specified threshold s, and it continues until the inventory level reaches the maximum capacity S. The time for producing one item follows an exponential distribution with a parameter β. When the inventory level further decreases to a level of r, the production rate increases to $\theta\beta$, where $\theta > 1$. The system returns to normal production mode whenever the inventory level reaches $r+1$ through accelerated production. When the system lacks inventory, it blocks customers in order to reduce holding costs. We study the system in detail and calculate the stationary probabilities. Multiple performance measures are identified. Numerical illustrations determine the optimal triplet (r,s,S). A parameter sensitivity analysis is performed.*

Keywords: *Inventory, Production Inventory, Normal Mode of Production, Accelerated Mode of Production, Product form solution*

1. INTRODUCTION

Queueing theory and inventory control are central in supply chain research. In production-inventory models, items are produced internally, unlike classical queueing-inventory models where they are sourced externally. Production begins once stock hits a lower threshold and halts at the maximum level.

The production-inventory model with positive service time was introduced by Krishnamoorthy and Viswanath (Krishnamoorthy and Viswanath, 2013), and extended under the (s,S) policy by Krishnamoorthy *et al.* (Krishnamoorthy, Nair, and Narayanan, 2013). Otten *et al.* (Otten et al., 2015) studied a two-tier supplier–inventory system, while Rafiei *et al.* (Rafiei et al., 2015) addressed wood remanufacturing under uncertainty. Ghiami and Williams (Ghiami and Williams, 2015) analyzed deterioration in a manufacturer–buyer setup, and Sivashankari and Panayappan (Sivashankari and Panayappan, 2015) modeled variable production for deteriorating items. Further, Patra (Patra, 2018) examined imperfect production; Huang *et al.* (Huang et al., 2018) proposed a Stackelberg model for food supply chains; Ruidas *et al.* (Ruidas et al., 2019) studied demand reduction from substitutes; Karim and Nakade (Karim and Nakade, 2019) analyzed random disruptions; Liu *et al.* (Liu et al., 2020) and He *et al.* (He et al., 2022) investigated co-production; Ruidas *et al.* (Ruidas et al., 2023) considered high-tech products; and Dey *et al.* (Dey et al., 2024) developed a periodic review model for perishables.

This paper analyzes a production-inventory queueing model with positive service time and an emergency production mode triggered below a critical threshold, extending Rof and Jacob (Rof and Jacob, 2024), who assumed negligible service time. Our contributions include product-form solutions, bounds for cycle time distributions, and optimization of the (r,s,S) policy. The remainder of the paper is organized as follows: Section 2 presents the model, Section 3 analyzes production cycles, Section 4 formulates the optimization problem, and Section 5 gives numerical illustrations.

Notations Used:

$exp(\mu)$: Exponential Distribution with parameter μ
$E_k(\mu)$: Erlang Distribution with parameter μ of order k
$GE_k(\lambda,\mu)$: Generalized Erlang of order k with parameters λ and μ
CTMC: Continuous Time Markov Chain

DRAFT VERSION

2. MODEL DESCRIPTION AND MATHEMATICAL FORMULATION

Consider a queueing-inventory system where customers arrive according to a Poisson process with rate λ. If the server is idle, a customer receives inventory and departs; otherwise, he joins an infinite-capacity queue. Service times are exponential with parameter μ. Production starts when inventory falls to s, with item production time exponentially distributed with rate β, continuing until level S is reached. If inventory drops to r ($r < s$), production switches to a faster rate $\theta\beta$ ($\theta > 1$) until level $r+1$, after which it reverts to normal. Customers are not allowed to join when inventory is zero, facilitating a product-form solution while reducing holding costs.

At time t, let $N(t)$ be the number of customers in the system, $I(t)$ the inventory level, and $C(t)$ the production status. Define

$$C(t) = \begin{cases} 0, & \text{if production is off,} \\ 1, & \text{if production is on with normal speed,} \\ 2, & \text{if production is on with accelerated speed.} \end{cases}$$

Then $X(t) = \{(N(t), I(t), C(t)) : t \geq 0\}$ is a CTMC with state space $\Omega = \{(n,i,j) : n \geq 0, 0 \leq i \leq S, j \in \{0,1,2\}\}$.

We enumerate the Markov process states in lexicographic order. The infinitesimal generator matrix Q records all transition rates over $[t, t+h)$. The following theorem is proved by analyzing these transitions.

► **Theorem 1.** *The infinitesimal generator matrix Q of the Markov Chain $\{(N(t), I(t)) : t \geq 0\}$ has the block-tridiagonal structure*

$$Q = \begin{bmatrix} B_1 & A_0 & & & \\ A_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}.$$

$$B_1 = \begin{bmatrix} L_0 & M_0 & & & \\ & L_1 & M_1 & & \\ & & \ddots & \ddots & \\ & & & L_{S-1} & M_{S-1} \\ & & & & L_S \end{bmatrix} \quad A_1 = \begin{bmatrix} K_0 & M_0 & & & \\ & K_1 & M_1 & & \\ & & \ddots & \ddots & \\ & & & K_{S-1} & M_{S-1} \\ & & & & K_S \end{bmatrix} \quad A_0 = \begin{bmatrix} 0 & & & \\ & N_1 & & \\ & & \ddots & \\ & & & N_{S-1} \\ & & & & N_S \end{bmatrix} \quad A_2 = \begin{bmatrix} 0 & & & \\ P_1 & & & \\ & \ddots & \ddots & \\ & & P_{S-1} & \\ & & & P_S \; 0 \end{bmatrix}$$

$$L_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\theta\beta \end{bmatrix}, \quad L_i = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -(\lambda+\theta\beta) \end{bmatrix} \; (1 \leq i \leq r), \quad L_i = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -(\lambda+\beta) & 0 \\ 0 & 0 & 0 \end{bmatrix} \; (r+1 \leq i \leq s),$$

$$L_i = \begin{bmatrix} -\lambda & 0 & 0 \\ 0 & -(\lambda+\beta) & 0 \\ 0 & 0 & 0 \end{bmatrix} \; (s+1 \leq i \leq S-1), \quad L_S = \begin{bmatrix} -\lambda & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad M_i = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \theta\beta \end{bmatrix} \; (0 \leq i \leq r-1),$$

$$M_r = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \theta\beta & 0 \end{bmatrix}, \quad M_i = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & 0 \end{bmatrix} \; (r+1 \leq i \leq S-2), \quad M_{S-1} = \begin{bmatrix} 0 & 0 & 0 \\ \beta & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad K_0 = L_0,$$

$$K_i = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -(\lambda+\mu+\theta\beta) \end{bmatrix} \; (1 \leq i \leq r), \quad K_i = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -(\lambda+\mu+\beta) & 0 \\ 0 & 0 & 0 \end{bmatrix} \; (r+1 \leq i \leq s), \quad K_i = \begin{bmatrix} -(\lambda+\mu) & 0 & 0 \\ 0 & -(\lambda+\mu+\beta) & 0 \\ 0 & 0 & 0 \end{bmatrix} \; (s+1 \leq i \leq S-1),$$

$$K_S = \begin{bmatrix} -(\lambda+\mu) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad N_i = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \lambda \end{bmatrix} \; (1 \leq i \leq r), \quad N_i = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & 0 \end{bmatrix} \; (r+1 \leq i \leq s), \quad N_i = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & 0 \end{bmatrix} \; (s+1 \leq i \leq S-1),$$

$$N_S = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad P_i = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \mu \end{bmatrix} \; (1 \leq i \leq r), \quad P_{r+1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \mu \\ 0 & 0 & 0 \end{bmatrix}, \quad P_i = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & 0 \end{bmatrix} \; (r+2 \leq i \leq s),$$

$$P_{s+1} = \begin{bmatrix} 0 & \mu & 0 \\ 0 & \mu & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad P_i = \begin{bmatrix} \mu & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & 0 \end{bmatrix} \; (s+2 \leq i \leq S-1), \quad P_S = \begin{bmatrix} \mu & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

3. STEADY STATE ANALYSIS OF THE SYSTEM

To find the steady state probabilities, we set $A = A_0 + A_1 + A_2$ and let $\phi = (\phi_0, \phi_1, \ldots, \phi_S)$ with $\phi_i = (\phi_i(0), \phi_i(1), \phi_i(2))$, $i = 0, 1, \ldots, S$. Define $\mu_1 = \mu/\beta$, $f(s) = \frac{\mu_1}{\mu_1 - 1}\left(\mu_1^{S-s+1} + \mu_1^{S-s} + \mu_1^{S-s-1} - 2\mu_1 + 1\right)$, $g(r) = \frac{1}{\theta}\left[f(s)\frac{\mu_1^{s-r}-1}{\mu_1-1} - \frac{\mu_1^{S-s}-1}{\mu_1-1}\frac{\mu_1^{s-r-1}-1}{\mu_1-1}\right]$, $h(r,s) = \frac{\mu_1^{S-s}-1}{\mu_1-1}\frac{\mu_1^{s-r-2}-1}{\mu_1-1}$, $m(r,s) = \frac{\mu_1(\mu_1^{s-r-1}-1)}{\mu_1-1}\left((\mu_1+\theta)^{r-2} - \theta^{r-2}\right)$, $t(i) = \mu_1\left((\mu_1+\theta)^{r-i-1} - \theta^{r-i-1}\right)$.

Solving $\phi A = 0$, $\phi e = 1$, we get

$$\phi_i(j) = \begin{cases} \phi_S(0), & s+1 \leq i \leq S, \; j = 0, \\ \left[f(s)\frac{\mu_1^{s-i}-1}{\mu_1-1} - \frac{\mu_1^{S-s}-1}{\mu_1-1}\frac{\mu_1^{s-i-1}-1}{\mu_1-1}\right]\phi_S(0), & r+1 \leq i \leq s-2, \; j = 1, \\ f(s)\phi_S(0), & i = s-1, \; j = 1, \\ \mu_1\frac{\mu_1^{S-i-1}-1}{\mu_1-1}\phi_S(0), & s \leq i \leq S-1, \; j = 1, \\ g(r)\phi_S(0), & i = r, \; j = 2, \\ \frac{1}{\theta^{r-i}}\left[g(r)\frac{\mu_1^{r-i+1}-\theta^{r-i+1}}{\mu_1-\theta} - t(i)f(s)\frac{\mu_1^{s-r-1}-1}{\mu_1-1} + t(i)h(r,s)\right]\phi_S(0), & 1 \leq i \leq r, \; j = 2, \\ \frac{\mu_1}{\theta^{r-i}}\left[g(r)\frac{\mu_1^r-\theta^r}{\mu_1-\theta} - m(r,s)f(s) + m(r,s)\frac{\mu_1^{S-s-1}-1}{\mu_1-1}\right]\phi_S(0), & i = 0, \; j = 2. \end{cases}$$

DRAFT VERSION

$$\phi_S(0)^{-1} = S-s+f(s)\left[\frac{\mu_1^2(\mu_1^{s-r-2}-1)}{(\mu_1-1)^2} - \frac{s-r-2}{\mu_1-1} - \frac{\mu_1^{s-r-1}-1}{\mu_1-1}\left(\frac{(\mu_1+\theta)^r-\theta^r}{(\mu_1+\theta)\theta^{r-1}} - \frac{r\mu_1}{\theta}\right)\right]$$

$$-\frac{\mu_1}{(\mu_1-1)^3}(\mu_1^{S-s}-1)\left[\mu_1^{s-r-2}-1-(\mu_1-1)(s-r-2)\right] + \frac{\mu_1}{(\mu_1-1)^2}\left[\mu_1(\mu_1^{S-s}-1)-(S-s)(\mu_1-1)\right]$$

$$+g(r)\left[\frac{\mu_1^r}{(\mu_1-\theta)^2\theta^{r-1}}\left(\mu_1^r-\theta^r-\theta r(\mu_1-\theta)\right) + \frac{\mu_1(\mu_1^r-\theta^r)}{\theta^r(\mu_1-\theta)}\right].$$

3.1. Stability Condition

► **Theorem 2.** *The system is stable if and only if* $\lambda < \mu$

Proof. The system is stable if and only if $\phi A_0 e < \phi A_2 e$
which implies, $\lambda < \mu$

◄

4. SYSTEM STATE PROBABILITY; PRODUCT-FORM SOLUTION

To find the state probability vector of $\{(I(t),C(t)) : t \geq 0\}$, we consider a production-inventory system with negligible service time and no backlog. The state space is $\{(i,j) : 0 \leq i \leq S,\ j \in \{0,1,2\}\}$. From steady-state analysis, the probability vector is $\pi = (\pi_0, \pi_1, \ldots, \pi_S)$, where $\pi_i = (\pi_i(0), \pi_i(1), \pi_i(2))$.

$$\lambda_1 = \tfrac{\lambda}{\beta}, \quad u(s) = \left(\tfrac{\lambda_1}{\lambda_1-1}\right)(\lambda_1^{S-s+1} + \lambda_1^{S-s} + \lambda_1^{S-s-1} - 2\lambda_1 + 1),$$
$$v(r) = \tfrac{1}{\theta}\left[u(s)\tfrac{\lambda_1^{s-r}-1}{\lambda_1-1} - \tfrac{\lambda_1^{S-s}-1}{\lambda_1-1}\cdot\tfrac{\lambda_1^{s-r-1}-1}{\lambda_1-1}\right],$$
$$w(r,s) = \tfrac{\lambda_1^{S-s}-1}{\lambda_1-1}\cdot\tfrac{\lambda_1^{s-r-2}-1}{\lambda_1-1}, \quad l(r,s) = \left(\lambda_1\tfrac{\lambda_1^{s-r-1}-1}{\lambda_1-1}\right)((\lambda_1+\theta)^{r-2}-\theta^{r-2}),$$
$$z(i) = \lambda_1\left((\lambda_1+\theta)^{r-i-1}-\theta^{r-i-1}\right).$$

Solving $\pi A = 0$ and $\pi e = 1$, we obtain

$$\pi_i(j) = \begin{cases} \pi_S(0), & s+1 \leq i \leq S,\ j=0, \\ \left[u(s)\tfrac{\lambda_1^{s-i}-1}{\lambda_1-1} - \tfrac{\lambda_1^{S-s}-1}{\lambda_1-1}\tfrac{\lambda_1^{s-i-1}-1}{\lambda_1-1}\right]\pi_S(0), & r+1 \leq i \leq s-2,\ j=1, \\ u(s)\pi_S(0), & i=s-1,\ j=1, \\ \lambda_1\tfrac{\lambda_1^{S-i-1}-1}{\lambda_1-1}\pi_S(0), & s \leq i \leq S-1,\ j=1, \\ v(r)\pi_S(0), & i=r,\ j=2, \\ \tfrac{1}{\theta^{r-i}}\left[v(r)\tfrac{\lambda_1^{r-i+1}-\theta^{r-i+1}}{\lambda_1-\theta} - z(i)u(s)\tfrac{\lambda_1^{s-r-1}-1}{\lambda_1-1} + z(i)w(r,s)\right]\pi_S(0), & 1 \leq i \leq r,\ j=2, \\ \tfrac{\lambda_1}{\theta^{r-i}}\left[v(r)\tfrac{\lambda_1^r-\theta^r}{\lambda_1-\theta} - l(r,s)u(s) + l(r,s)\tfrac{\lambda_1^{S-s-1}-1}{\lambda_1-1}\right]\pi_S(0), & i=0,\ j=2. \end{cases}$$

Normalization gives

$$\pi_S(0)^{-1} = S-s+u(s)\left[\frac{\lambda_1^2(\lambda_1^{s-r-2}-1)}{(\lambda_1-1)^2} - \frac{s-r-2}{\lambda_1-1} - \frac{\lambda_1^{s-r-1}-1}{\lambda_1-1}\left(\frac{(\lambda_1+\theta)^r-\theta^r}{(\lambda_1+\theta)\theta^{r-1}} - \frac{r\lambda_1}{\theta}\right)\right]$$

$$-\frac{\lambda_1}{(\lambda_1-1)^3}(\lambda_1^{S-s}-1)\left[\lambda_1^{s-r-2}-1-(\lambda_1-1)(s-r-2)\right] + \frac{\lambda_1}{(\lambda_1-1)^2}\left[\lambda_1(\lambda_1^{S-s}-1)-(S-s)(\lambda_1-1)\right]$$

$$+v(r)\left[\frac{\lambda_1^r}{(\lambda_1-\theta)^2\theta^{r-1}}\left(\lambda_1^r-\theta^r-\theta r(\lambda_1-\theta)\right) + \frac{\lambda_1(\lambda_1^r-\theta^r)}{\theta^r(\lambda_1-\theta)}\right].$$

Using the component probability vector π, the steady-state vector of $\{(N(t),I(t),C(t)) : t \geq 0\}$ is derived. Let $x = (x_0, x_1, \ldots)$ with $x_i = (x_i(0), x_i(1), \ldots, x_i(S))$, $i \geq 1$. The balance equations are

$$x_0B_1 + x_1A_2 = 0, \qquad x_iA_0 + x_{i+1}A_1 + x_{i+2}A_2 = 0,\ i \geq 0,$$

with

$$x_0 = \zeta\pi, \qquad x_i = \zeta\left(\frac{\lambda}{\mu}\right)^i\pi,\ i \geq 1,$$

for some constant ζ.

► **Theorem 3.** *If* $\rho = \frac{\lambda}{\mu} < 1$*, then the steady-state vector is*

$$x_0 = (1-\rho)\pi, \qquad x_i = \rho^i(1-\rho)\pi,\ i \geq 1,$$

where π *is the steady-state vector of* $\{(I(t),C(t)) : t \geq 0\}$.

DRAFT VERSION

4.1. Performance Measures

- Average Number of Customers in the System: $L_s = \frac{\lambda}{\mu-\lambda}$.
- Average Number of Customers in the Queue: $L_q = \frac{\lambda^2}{\mu(\mu-\lambda)}$.
- Expected Number of Inventory in the System: $E_{inv} = \sum_{j=1}^{S} j\pi_j = \sum_{j=1}^{r} j\pi(j,2) + \sum_{j=r+1}^{s} j\pi(j,1) + \sum_{j=s+1}^{S} j\pi(j,0)$.
- Expected Rate of Production: $ERP = \sum_{j=0}^{r} \pi(j,2)\theta\beta + \sum_{j=r+1}^{S-1} \pi(j,1)\beta$.
- Expected Number of Customers Lost: $E_{loss} = \frac{\lambda}{\theta\beta}$.
- Expected Rate at which Production is Switched Off: $E_{off} = \sum_{j=s+1}^{S} \pi(j,0) = (S-s)\pi(S,0)$.
- Expected Rate of Production in Normal Mode: $ERP_{norm} = \left(\sum_{j=r+1}^{S-1} \pi(j,1)\right)\beta$.
- Expected Rate of Production in Accelerated Mode: $ERP_{acc} = \sum_{j=0}^{r} \pi(j,2)\theta\beta$.
- Expected Rate at which Normal Production is Switched On: $E_{normON} = \sum_{i=0}^{\infty} x_i(s+1,0)\mu = \sum_{i=0}^{\infty}(1-\rho)\rho^i\pi(S,0)\mu = \rho\mu\pi(S,0) = \lambda\pi(S,0)$.
- Expected Rate at which Accelerated Production is Switched On: $E_{accON} = \sum_{i=0}^{\infty} x_i(r+1,1)\mu = \sum_{i=0}^{\infty}(1-\rho)\rho^i\pi(r+1,1)\mu = \rho\mu\pi(r+1,1) = \lambda\pi(r+1,1)$.

5. BOUND FOR THE PRODUCTION CYCLE DISTRIBUTION

The production cycle is the time between two successive visits to level s. Exact cycle time distribution is intractable, but a greatest lower bound (GLB) can be obtained. Consider an epoch when inventory drops to s from $s+1$ with production off. Production starts and continues until level S, switching off at time T_1.

We compute the distribution in two cases:

i) **No customer waiting when level drops to s:** Production begins after an exponential(β) delay and continues until $S-s$ items are produced. The duration is Erlang$(S-s,\beta)$, denoted $E_{S-s}(\beta)$. The probability of no arrivals during this period is

$$\Pr\{E_{S-s}(\beta) < E_1(\lambda)\} = \int_0^{\infty} \frac{e^{-\beta t}\beta^{S-s}t^{S-s-1}}{(S-s-1)!} e^{-\lambda t}\, dt.$$

ii) **One or more customers waiting at s:** Here, we compute the probability of no service completion before production ends:

$$\Pr\{E_{S-s}(\beta) < E_1(\mu)\} = \int_0^{\infty} \frac{e^{-\beta t}\beta^{S-s}t^{S-s-1}}{(S-s-1)!} e^{-\mu t}\, dt.$$

Combining these two cases yields the probability distribution of the cycle time.

$$\begin{aligned} f(t) &= x_0 e \int_0^{\infty} \frac{e^{-\beta t}\beta^{S-s}t^{S-s-1}}{(S-s-1)!} e^{-\lambda t}\, dt \quad + (1-x_0 e)\int_0^{\infty} \frac{e^{-\beta t}\beta^{S-s}t^{S-s-1}}{(S-s-1)!} e^{-\mu t}\, dt \\ &= \int_0^{\infty} \frac{e^{-\beta t}\beta^{S-s}t^{S-s-1}}{(S-s-1)!} \left[x_0 e e^{-\lambda t} + (1-x_0 e)e^{-\mu t}\right] dt \\ &= \int_0^{\infty} \frac{e^{-\beta t}\beta^{S-s}t^{S-s-1}}{(S-s-1)!} \left[(1-\rho)e^{-\lambda t} + \rho e^{-\mu t}\right] dt \end{aligned}$$

6. OPTIMIZATION PROBLEM

Based on the performance measures discussed above, we construct the following cost function:

$$C(r,s,S) = E_{inv}H_c + E_{loss}L_c + ERP_{normal}N_c + ERP_{acc}A_c + E_{normON}N_{on} + E_{accON}A_{on}$$

Let H_c be the holding cost per unit, L_c the cost of lost customers, N_c the normal production cost per unit time, and A_c the accelerated production cost. The fixed costs for starting normal and accelerated production are N_{on} and A_{on}, respectively. Substituting the performance measures, we obtain

$$\begin{aligned} C(r,s,S) = H_c &\sum_{j=1}^{S} j\pi_j + L_c\frac{\lambda}{\theta\beta} + N_c \sum_{j=r+1}^{S-1} \pi(j,1)\beta \\ &+ A_c \sum_{j=0}^{r} \pi(j,2)\theta\beta + N_{on}\lambda\pi(S,0) + A_{on}\lambda\pi(r+1,1) \end{aligned}$$

Since the expression is complex, verifying convexity is difficult; hence, we determine the optimal values numerically.

7. NUMERICAL ILLUSTRATION

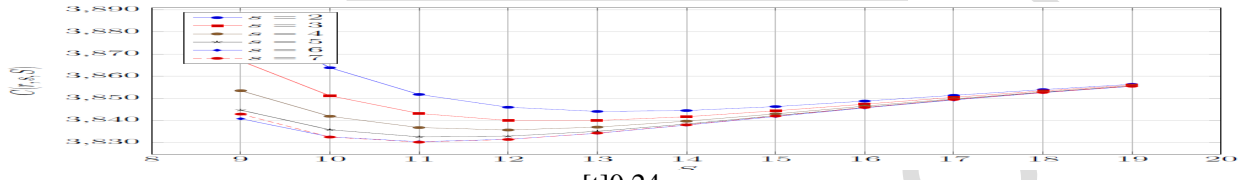
For numerical analysis, parameter values are set under the stability condition. We study the effect of varying one parameter at a time on performance measures and identify the optimal triplet (r,s,S) minimizing system cost. The cost function is evaluated for different r,s,S values, with results shown in tables and (s,S) graphs for varying r. We find the optimal triplet (r^*,s^*,S^*) such that $C(r^*,s^*,S^*) = \min\ C(r,s,S)$. From *Table 4 to Table 7*, the optimal triplet is $C(r^*,s^*,S^*) = (1,7,11)$ with a minimum cost of \$3830.1.

DRAFT VERSION

[t]0.24

Table 1: $r = 1$

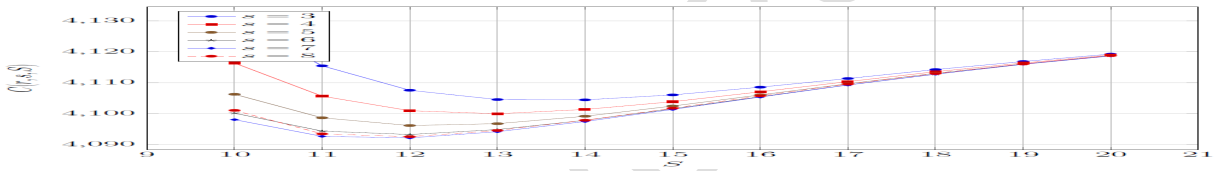
S/s	2	3	4	5	6	7
9	3885.4	3866.9	3853.4	3844.6	3840.7	3842.8
10	3863.8	3851.1	3841.8	3835.7	3832.5	3832.4
11	3851.7	3843.1	3836.7	**3832.5**	**3830.3**	**3830.1**
12	3845.9	**3839.9**	**3835.6**	3832.8	3831.4	3831.3
13	**3844.0**	3839.9	3836.9	3835.0	3834.1	3834.2



[t]0.24

Table 2: $r = 2$

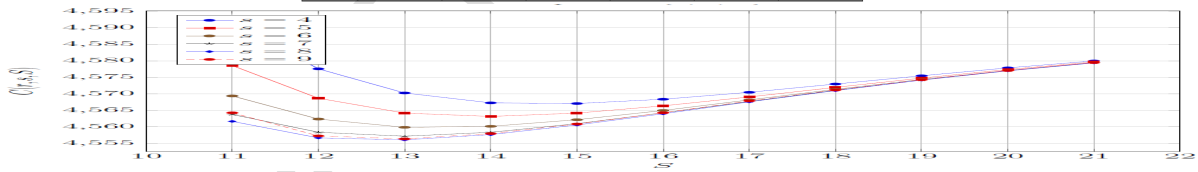
S/s	3	4	5	6	7	8
10	4130.7	4116.4	4106.3	4100.2	4098.1	4101.1
11	4115.5	4105.7	4098.7	4094.4	4092.7	4093.5
12	4107.6	4101.0	**4096.2**	**4093.3**	**4092.2**	**4092.6**
13	4104.6	**4100.0**	4096.8	4094.9	4094.2	4094.6
14	**4104.5**	4101.4	4099.2	4097.9	4097.5	4097.9



[t]0.24

Table 3: $r = 3$

S/s	4	5	6	7	8	9
11	4591.6	4578.6	4569.4	4563.8	4561.7	4564.3
12	4577.6	4568.7	4562.4	4558.4	4556.7	4557.3
13	4570.3	4564.2	**4559.9**	**4557.2**	**4556.1**	**4556.4**
14	4567.3	**4563.2**	4560.2	4558.4	4557.7	4558.0
15	**4567.1**	4564.2	4562.2	4561.0	4560.6	4560.9



[t]0.24

Table 4: $r = 4$

S/s	5	6	7	8	9	10
12	5193.0	5180.0	5262.7	5164.2	5161.4	5162.9
13	5177.8	5168.9	5212.9	5157.9	5155.6	5155.5
14	5169.3	5163.1	5185.8	**5155.6**	**5154.0**	**5153.8**
15	5165.1	5160.9	5170.4	5155.7	5154.7	5154.6
16	5163.8	**5160.8**	5162.2	5157.4	5156.7	5156.7

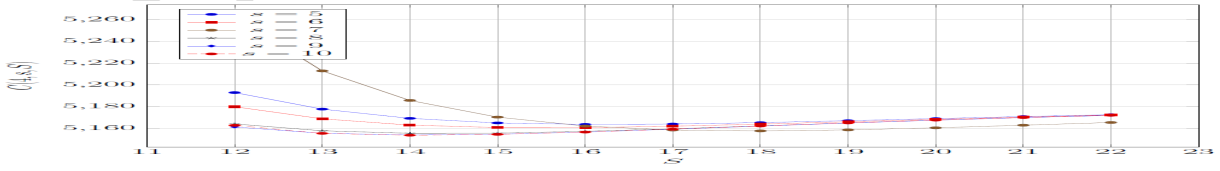


Figure 1: Comparison of $C(r,s,S)$ values and line graphs for $r = 1,2,3,4$

DRAFT VERSION

CONCLUSION

In this article, we discuss a queueing inventory model with an alternative mode of production. Normally, production begins when the inventory level reaches s and continues until it reaches S. The production time is exponentially distributed with parameter β, and shifts to a faster mode when the inventory level falls to r, with a rate $\theta\beta$ ($\theta > 1$). Customer arrivals follow a Poisson process with rate λ, and service is exponential with rate μ. The chapter provides a product-form solution for the system state probabilities and discusses performance measures, including the least upper bound for cycle time duration and a cost function. Numerical studies are performed to find the optimal cost, and the convex nature of the cost function is verified. Future work could involve exploring more general distributions for the model's parameters.

ACKNOWLEDGMENT

This work has no funding and has no contribution other than authors.

LITERATURE

Dey, D., Kar, S., & Bhattacharya, D. K. (2024). Periodic review policy in the production management of a single-item inventory with deterioration. *Operational Research*, *34*.

Ghiami, Y., & Williams, T. (2015). A two-echelon production-inventory model for deteriorating items with multiple buyers. *International Journal of Production Economics*, *159*, 233–240.

He, S., Zhang, J., Zhang, J., & Cheng, T. C. E. (2022). Production/inventory competition between firms with fixed-proportions co-production systems. *European Journal of Operational Research*, *299*(2), 497–509.

Huang, H., He, Y., & Li, D. (2018). Pricing and inventory decisions in the food supply chain with production disruption and controllable deterioration. *Journal of Cleaner Production*, *180*, 280–296.

Karim, R., & Nakade, K. (2019). A stochastic model of a production-inventory system with consideration of production disruption. *International Journal of Advanced Operations Management*, *11*(4), 287–316.

Krishnamoorthy, A., Nair, S. S., & Narayanan, V. C. (2013). Production inventory with service time and interruptions. *International Journal of Systems Science*, *46*(10), 1800–1816.

Krishnamoorthy, A., & Viswanath, N. C. (2013). Stochastic decomposition in production inventory with service time. *European Journal of Operational Research*, *228*(2), 358–366.

Liu, H., Zhang, J., Cheng, T. C. E., & Ru, Y. (2020). Optimal production-inventory policy for the multi-period fixed proportions co-production system. *European Journal of Operational Research*, *280*(2), 469–478.

Otten, S., Krenzler, R., & Daduna, H. (2015). Models for integrated production-inventory systems: Steady state and cost analysis. *International Journal of Production Research*, *54*(20), 6174–6191.

Patra, K. (2018). A production inventory model with imperfect production and risk. *International Journal of Applied and Computational Mathematics*, *4*(3), 91.

Rafiei, R., Nourelfath, M., Gaudreault, J., Eulalia, L., & Bouchard, M. (2015). Dynamic safety stock in co-production demand-driven wood remanufacturing mills: A case study. *International Journal of Production Economics*, *165*, 90–99.

Rof, A., & Jacob, J. (2024). Queueing inventory with an additional mode of production. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(21s), 4414–4420.

Ruidas, S., Seikh, M. R., & Nayak, P. K. (2019). A single period production inventory model in interval environment with price revision. *International Journal of Applied and Computational Mathematics*, *5*, 7.

Ruidas, S., Seikh, M. R., & Nayak, P. K. (2023). A production inventory model for high-tech products involving two production runs and a product variation. *Journal of Industrial and Management Optimization*, *19*(3).

Sivashankari, C. K., & Panayappan, S. (2015). Production inventory model for two-level production with deteriorative items and shortages. *The International Journal of Advanced Manufacturing Technology*, *76*, 2003–2014.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

QUADRATIC MODEL FOR MULTI-DEPOT VEHICLE ROUTING PROBLEM WITH TIME WINDOWS IN DISASTER MANAGEMENT

NEMANJA GOLUBOVIĆ[1], BRANISLAVA RATKOVIĆ, NENAD BJELIĆ[3]

[1] University of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, n.golubovic@sf.bg.ac.rs, 0009-0009-4336-9039
[2] University of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, b.ratkovic@sf.bg.ac.rs, 0000-0002-7868-7899
[3] University of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, n.bjelic@sf.bg.ac.rs, 0000-0002-0376-5215

Abstract: *Disasters, natural or man-made, represent a serious threat to communities and the environment, with an increasing number of affected people worldwide. Efficient disaster management, particularly in the post-disaster phase, plays a vital role in reducing the consequences of such events through the timely distribution of essential aid. Logistics is recognized as a critical component of disaster management, encompassing facility location, resource allocation, and vehicle routing. This paper addresses the Multi-Depot Vehicle Routing Problem with Time Windows (MDVRPTW), which integrates depot locations and vehicle routing, while considering the time windows that are introduced to ensure that aid reaches affected population within critical time frames. A mathematical model is proposed that improves existing mathematical formulations by incorporating realistic post-disaster constraints such as vehicle capacity, depot limitations, and time-critical service requirements. The model is validated through a case study focusing on two urban neighborhoods in Belgrade, demonstrating the effectiveness of the approach in supporting aid distribution under constrained conditions. The findings emphasize the importance of logistics operations and provide a foundation for future improvements in disaster relief planning.*

Keywords: *location-routing, disaster management, MDVRPTW*

DRAFT VERSION

1. INTRODUCTION

Disasters, whether they be natural or man-made, represent a serious threat to the community and the environment. In recent years, we have witnessed a significant increase in the number and expansion of disasters and their consequences. According to the research by Delforge et al. (2025), disasters in 2023 affected approximately 93.1 million people worldwide, which resulted in 86,473 fatalities, around 20,000 more than the average over the past two decades. Due to the growing trends of the negative effects of disasters, it is necessary to constantly research scientific disciplines that minimize them. Effective disaster management plays a crucial role in prevention and mitigation, offering a structured approach to preparedness, response, and recovery. Earthquakes, floods and other types of natural disasters are always accompanied by a series of consequences for the population, infrastructure, and functionality of the community. Due to disruptions in daily activities, there is often an urgent need to provide assistance to the endangered population. That assistance may include the delivery of essential supplies, evacuation or medical assistance. In such situations, it is necessary to establish a stable logistic network. Considering this, we can say that logistics is an integral component of effective disaster management.

Disaster management and logistics are recognized as an essential for properly dealing with disasters. Almost always, they are followed by uncertainty that is reflected in disruptions in transportation networks, demands for resources, severity of damage, etc. which complicates planning and decision-making. Through various studies and researches, models and methods for managing logistics activities in disasters have been developed. These activities are related to the problems of inventory management, location of facilities, vehicle routing, etc. All these activities can be classified based on the time they are carried out in relation to the occurrence of the disaster itself. Accordingly, we distinguish between pre-disaster and post-disaster phases. The pre-disaster phase includes all preparedness activities aimed at mitigating the impact of the disaster, while the post-disaster phase encompasses response and recovery efforts focused on providing assistance to the affected population. Location models can be applied in both the pre-disaster and post-disaster phases. In flood evacuation planning, location models have been used to determine optimal shelter placements (Kongsomsaksakul et al., 2005). Verma and Gaukler (2015) proposed a deterministic location model for

disaster response facilities to pre-position emergency supplies. Similarly, other authors utilize such models to address these issues in the pre-disaster phase. The majority of authors, focusing on the pre-disaster phase, deal with warehouse location and emergency supplies, as it is demonstrated by Yu (2021). While location models in pre-disaster phase are used for locating warehouses and supplies, similar models are applied in the post-disaster phase to locate temporary hospitals and distribution centers. For example, Hu et al. (2013) proposed a model combined with reliability to select the emergency distribution centers. A robust location-allocation framework for healthcare facilities was introduced by Alinaghian et al. (2021). Along with distribution centers and medical centers, location models can be used to determine the optimal locations for warehouses, shelters, and debris removal sites (Boonmee et al., 2017).

In addition to facility location models, vehicle routing models play a significant role in disaster management. These models help optimize the movement of resources, personnel and supplies to affected areas, ensuring efficient and timely delivery. Vahdani et al. (2018) address location and routing problems separately in two stages. In the first stage, facility location is established, while in the second, vehicle routing is performed. It is common for location and routing problems to be solved simultaneously so that overall optimal solution can be achieved. In that sense, mathematical model used in this research, MDVRP (eng. Multi-Depot Vehicle Routing Problem) is developed with the same goal. MDVRP model is proposed for locating local depots and distributing relief supplies by Uslu et al. (2017). Similar, Wei et al. (2020) suggested location-routing model for post-disaster relief distribution. Location-routing models can be used regardless of the type of disaster, for instance, Vieira et al. (2021) applied MDVRP to emergency water trucking. Same or relatively similar MDVRP models can be found in many publications, but there are still differences. Those differences can be reflected in the objective functions, which might aim to maximize the quality of service (Beiki et al., 2020), minimize distribution time and costs (Vahdani et al., 2018) or minimize delays (Wei et al., 2020). Other differences are reflected in concepts such as urgency of delivery or reliability of routes. Although these differences are evident and noticeable, the ultimate goal is the optimization of logistics activities in disaster management.

Logistics plays a critical role in post-disaster operations, directly impacting the ability to save lives and reduce the devastating effects of disasters. In such scenarios, time becomes one of the most crucial factors, requiring constant attention and optimization. Timely delivery of essential supplies in post-disaster operations is crucial, as delays can have severe consequences for the lives and conditions of the affected population. Any delay in delivery can have long-lasting effects and consequences. Due to the urgency of certain types of aid, there are moments and time windows in which the delivery must be completed. The aim of this paper is to highlight the importance of time windows, their introduction, and their application in post-disaster logistics operations.

The remaining of the paper is organized as follows. In the second section we presented problem description along with the MDVRPTW (eng. Multi Depot Vehicle Routing Problem with Time Windows) mathematical model, which enables optimal solving of the described problem. In the third section, the MDVRPTW is applied to a case study to demonstrate the functionality of the model. Finally, the conclusion remarks are given in the last section, along with suggestions for further research to enhance the effectiveness of post-disaster logistics operations.

2. PROBLEM DESCRIPTION AND MODEL FORMULATION

After a disaster, due to the probable infrastructure damage, it is necessary to establish an adequate logistics network for aid distribution which includes a set of distribution facilities on an affected region. In conditions after a disaster every user j from the set $J = \{1, 2,...,|J|\}$ who requires relief, has a specific time window $[T1_j, T2_j]$ during which user must receive an assistance. Time windows are defined by the moment when the need for aid arises ($T1$) and the latest moment a user can wait without an assistance ($T2$). It is crucial to respect time windows because it reduces consequences of the disaster. Accordingly, it is necessary to use a certain number of local depots from the set $I = \{|J|+1, |J|+2,....,|J+I|\}$ which represents a set of potential locations of local depots. Besides the relief goods, the available fleet of distribution vehicles is another key resource. To efficiently distribute and satisfy user requirements it is necessary to define the routes of available delivery vehicles from the set $V = \{1, 2,....,|V|\}$ with limited capacity q through the network of nodes $N = \{1,2,..n,.., |J|,|J|+ 1, |J|+2,..m,..,|J+I|\}$. The problem was addressed using the MDVRPTW model, which was developed as follows. Based on the constraints introduced in the works (Wei et al., 2020) and (Xu et al., 2023), the model outlined in study (Golubović et al., 2024) has been further improved in this research. In this way, new constraints were added to the already existing model, which more closely describe the problems that can be encountered in real situations. Proposed model is used for locating local depots and routing delivery vehicles while respecting time constraints imposed by the users.

DRAFT VERSION

The notation used in this model is presented below:

Sets

$J = \{1,2,\ldots, |J|\}$ – a set of nodes on a network representing users requesting service
$I = \{|J|+1, |J|+2,\ldots, |J+I|\}$ – a set of nodes on a network representing potential locations of local depots
$N = J \cup I = \{1,2,..n,.., |J|,|J|+1, |J|+2,..m,..,|J+I|\}$ – a set of all nodes on a network
$V = \{1,2,\ldots, |V|\}$ – set of distribution vehicles
$K = \{1,2,\ldots, |K|\}$ – a set of tours

Parameters

fc – costs of opening a local depot
Q – local depot capacity
q – vehicle capacity
p – the cost of unsatisfied customer requirements per unit of goods
d_j – demand of user *j*
st – service time
t_{nm} – the delivery time from node *n* to node *m*
$T1_j$ – the moment when the need for help by user *j* arises *(*left hand side of a time window for user *j)*
$T2_j$ – the latest moment for the delivery of relief goods at user *j (*right hand side of a time window for user *j)*

Variables

δ_i – 1 if local depot *i* is opened; 0 otherwise
φ_{ik} – 1 if tour *k* is assigned to local depot *i*; 0 otherwise
x_{nmk} – 1 if node *m* is visited immediately after node *n* within tour *k*; 0 otherwise
y_{jk} – 1 if user *j* is served in tour *k*; 0 otherwise
u_v – 1 if vehicle *v* is used; 0 otherwise
l_{kv} – 1 if vehicle *v* is assigned to tour *k*; 0 otherwise
λ_{iv} – 1 if vehicle *v* is assigned to local depot *i*; 0 otherwise
T_n – moment of arrival at node *n*
μ_j – quantity of goods sent to user *j*
γ – the amount of unsatisfied demands

The mathematical formulation of the proposed model is presented below:

$$\textbf{min } z = \sum_{n\in N}\sum_{m\in N}\sum_{k\in K} t_{nm}x_{nmk} + p\gamma + fc\sum_{i\in I}\delta_i \quad (1)$$

$$\sum_{i\in I}\varphi_{ik} \le 1 \quad (\forall k \in K) \quad (2)$$

$$\varphi_{ik} \le \delta_i \ (\forall k \in K, i \in I) \quad (3)$$

$$\sum_{j\in J} x_{ijk} = \varphi_{ik} \quad (\forall k \in K, i \in I) \quad (4)$$

$$\sum_{j\in J} x_{jik} = \varphi_{ik} \quad (\forall k \in K, i \in I) \quad (5)$$

$$\sum_{m\in N} x_{nmk} = \sum_{m\in N} x_{mnk} \ (\forall n \in N, k \in K) \quad (6)$$

$$\sum_{i\in I}\sum_{j\in J} x_{jik} \le 1 \ (\forall k \in K) \quad (7)$$

$$\sum_{i\in I}\sum_{e\in I}\sum_{k\in K} x_{iek} = 0 \quad (8)$$

$$\sum_{n\in N} x_{njk} = y_{jk} \quad (\forall k \in K, j \in J) \quad (9)$$

$$\sum_{k\in K} y_{jk} \le 1 \quad (\forall j \in J) \quad (10)$$

$$\sum_{v\in V} l_{kv} \le 1 \quad (\forall k \in K) \quad (11)$$

$$\sum_{v\in V} l_{kv} \ge y_{jk} \quad (\forall k \in K, j \in J) \quad (12)$$

$$\sum_{k\in K} l_{kv} \le |K|u_v \quad (\forall v \in V) \quad (13)$$

$$\sum_{k\in K} l_{kv} \ge u_v \quad (\forall v \in V) \quad (14)$$

$$\mu_j = d_j \sum_{k\in K} y_{jk} \ (\forall j \in J) \quad (15)$$

DRAFT VERSION

$$d_j \sum_{n\in N} \sum_{k\in K} x_{njk} \geq \mu_j \ (\forall j \in J) \quad (16)$$

$$\sum_{j\in J} d_j - \sum_{h\in J} \mu_h = \gamma \quad (17)$$

$$\varphi_{ik} y_{jk} = Z_{ijk} \ (\forall k \in K, j \in J, i \in I) \quad (18)$$

$$\sum_{j\in J} \mu_j \sum_{k\in K} Z_{ijk} \leq Q \ \ (\forall i \in I) \quad (19)$$

$$\sum_{j\in J} y_{jk} \mu_j \leq q \ \ (\forall k \in K) \quad (20)$$

$$\sum_{k\in K} \varphi_{ik} l_{kv} \leq |K| \lambda_{iv} \quad (\forall i \in I, v \in V) \quad (21)$$

$$\sum_{i\in I} \lambda_{iv} \leq 1 \quad (\forall v \in V) \quad (22)$$

$$T1_j \leq T_j \leq T2_j \quad (\forall j \in J) \quad (23)$$

$$T_m \geq T_n + (st + t_{nm} * x_{nmk}) - M(1 - x_{nmk}) \quad (\forall n, m \in J) \quad (24)$$

$$T_m \geq T_n + (t_{nm} * x_{nmk}) - M(1 - x_{nmk}) \quad (\forall n \in \mathrm{I}, m \in J) \quad (25)$$

$$\delta_i, \varphi_{ik}, x_{nmk}, y_{jk}, u_v, l_{kv}, \lambda_{iv} \ \in \{0,1\} \ (\forall k \in K, j \in J, i \in I, n, m \in N) \quad (26)$$

$$\mu_j, \gamma, \beta_{nk} \geq 0 \ (\forall n \in I, j \in J, k \in K) \quad (27)$$

The objective function (1) aims to minimize costs associated with opening depots, unsatisfied requests and transportation through the network. Each tour *k* can only be assigned to one depot (2), and any depot that is assigned to a tour must be located (3). Constraint sets (4), (5) and (6) ensure that each tour begins and ends at its corresponding depot and that every vehicle must leave the node it enters. Additionally, each vehicle can only return to a located depot once within a single tour (7). Constraint (8) prevents direct movement between two local depots. Constraint sets (9) and (10) guarantee that each user within a tour is served exactly once. Each tour must be completed by only one vehicle, as it is defined by restriction (11). Every tour must be assigned to a service vehicle (12), with the possibility that each vehicle can do multiple tours (13) and (14). Sets of constraints (15), (16), and (17) describe the quantities delivered to users, as well as any unsatisfied requests. Constraint sets (18) and (19) ensure that local depots do not exceed their capacity, while vehicle capacity is limited by constraints (20). Expression (21) ensures that a vehicle can only be assigned to a tour if the tour and the vehicle are assigned to the same depot. Additionally, each vehicle can be assigned to only one depot (22). Constraint sets (23), (24), and (25) define the time windows within which users must be served. The nature of variables is provided by expressions (26) and (27).

3. NUMERICAL EXAMPLE

To demonstrate the practical application of the proposed MDVRPTW model, this numerical example utilizes an illustrative example that reflects real challenges in aid distribution after the disaster. Two municipalities in the city of Belgrade were observed: Karaburma and Mirijevo. These areas municipalities are home to a population of 150000 residents. Due to their characteristics, these neighborhoods were considered as a part of the central urban area, which is reflected in heavy traffic congestion and slower vehicle movement. Accordingly, the delivery vehicle speed in this example was set at 30 km/h. For users, i.e. points where aid needs to be delivered, 20 locations were randomly selected. In addition to the users and demand, potential locations for local depots also had to be selected and state institutions such as health care centers were considered for several reasons. First, these facilities already provide the necessary infrastructure, reducing the costs and time required for adaptation or potential expansion. Additionally, these facilities are typically strategically located, ensuring good connectivity with other key points in the region. Furthermore, by using existing facilities, the need to build new facilities is eliminated, which reduces the time of implementation of all planned activities. Accordingly, in the considered area, 5 potential locations for local depots were identified (Figure 1).

As previously mentioned, each user has a specific time window within which this need must be fulfilled. These time constraints, along with the corresponding demand quantities, were considered in the model formulation. In accordance with the example, the vehicle capacity was set to 100 units. Implementation of the given model requires the distances between relevant locations. In this case, the key points include existing healthcare facilities selected as potential locations and the users requiring services. All users are numbered from 1 to 20, while the potential depot locations are numbered from 21 to 25. Time windows and demands of every user are presented in Table 1.

DRAFT VERSION

Table 1: Users demands and time windows

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
T1*j*	2.4	0.5	2	2.3	2.1	0	2.5	1.3	1.4	2.5	2.7	0.5	1	2.5	0.4	2	2.3	1.1	0.2	1.2
T2*j*	2.5	0.6	.2.1	2.4	2.2	0.1	2.6	1.4	1.5	2.6	2.8	0.6	1.1	2.6	0.5	2.1	2.4	1.2	0.3	1.3
d*j*	25	18	37	31	30	23	34	14	43	23	49	12	24	40	24	43	38	34	46	15

The presented model found the optimal solution for the observed problem. The demands of all 20 users were fully satisfied through the deployment of 5 vehicles operating across 9 routes, with the strategic location of 2 local depots. This configuration ensured an efficient allocation of resources and effective coverage of all service areas. The graphical presentation of the locations and vehicle routes is shown in Figure 1.

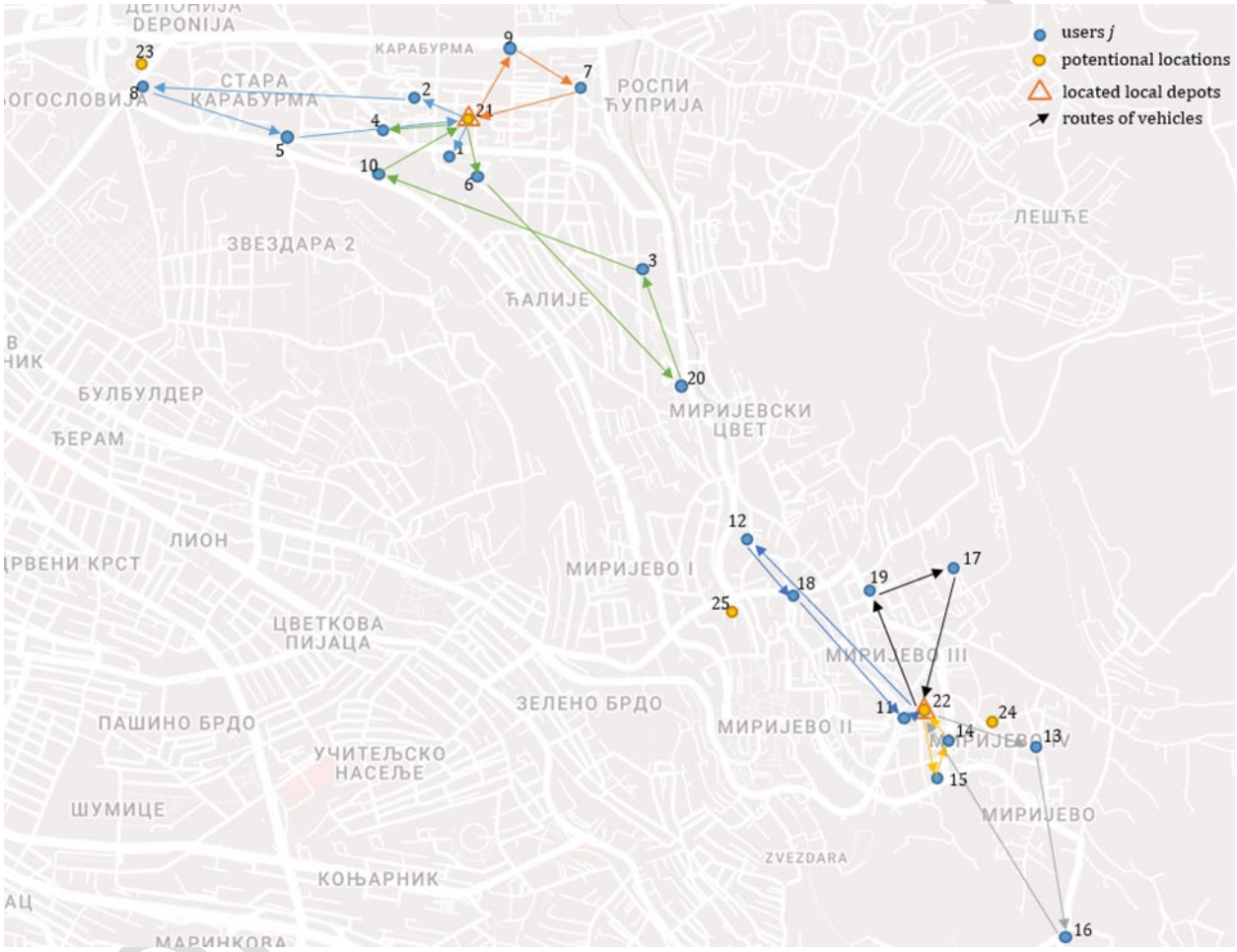


Figure 1: Users, local-depots and routes of vehicles

4. CONCLUSION

This paper highlights the crucial role of optimized logistics in effective disaster response and recovery efforts, which include various operations such as aid distribution, evacuation, and health care. Optimization in logistics not only influence resources reduction, but also significantly mitigates the consequences of disasters. This study specifically focuses on the implementation of MDVRPTW model as a mean of optimizing logistics operations. The proposed model is a modification of existing ones, tailored as a practical solution for locating local depots and distributing relief. It considers the dynamic nature of relief requests, which arise at different times, emphasizing the importance of delivering aid within specific, urgent time windows. This adaptability enables more reliable planning under the unpredictable conditions typical of disaster zones, ensuring faster deployment and more effective resource utilization. Ultimately, this improves response capabilities and supports faster recovery in affected regions. Beyond disaster management, the principles of this model could be extended to other time-sensitive logistics applications, such as medical supply distribution or emergency

response in urban areas. Future research could explore integrating real-time data feeds, such as weather updates or road conditions, or leveraging artificial intelligence and machine learning for adaptive route optimization. Additionally, practical challenges in real-world implementation, such as infrastructure limitations or coordination between multiple agencies, should be further investigated.

LITERATURE

Alinaghian, M., Hejazi, S. R., Bajoul, N., & Sadeghi Velni, K. (2021). A Novel Robust Model for Health Care Facilities Location-Allocation Considering Pre Disaster and Post Disaster Characteristics. Scientia Iranica, 0(0), 0–0. https://doi.org/10.24200/sci.2021.5743.1459

Beiki, H., Seyedhosseini, M., & Ghezavati, V. (2020). A Location-Routing Model for Assessment of the Injured People and Relief Distribution under Uncertainty. International Journal of Engineering, 33(7). https://doi.org/10.5829/ije.2020.33.07a.14

Boonmee, C., Arimura, M., & Asada, T. (2017). Facility location optimization model for emergency humanitarian logistics. International Journal of Disaster Risk Reduction, 24, 485–498. https://doi.org/10.1016/j.ijdrr.2017.01.017

Delforge, D., Wathelet, V., Below, R., Sofia, C. L., Tonnelier, M., Van Loenhout, J. A. F., & Speybroeck, N. (2025). EM-DAT: The Emergency Events Database. International Journal of Disaster Risk Reduction, 105509. https://doi.org/10.1016/j.ijdrr.2025.105509

Golubović, N., Bjelić, N., & Ratković, B. (2024). MIQP MODEL RUNNING TIME ANALYSIS IN DISASTER MANAGEMENT. CONFERENCE PROCEEDINGS. 51ST SYMPOSIUM ON OPERATIONAL RESEARCH. https://doi.org/10.24867/SYMOPIS-2024-51-039

Hu, T.-Y., Ho, W.-M., Ko, Y.-N., & Chen, L.-W. (2013). Emergency Distribution Center Selection with Network Reliability Consideration. Proceedings of the Eastern Asia Society for Transportation Studies.

Kongsomsaksakul, S., Yang, C., & Chen, A. (2005). SHELTER LOCATION-ALLOCATION MODEL FOR FLOOD EVACUATION PLANNING. Journal of the Eastern Asia Society for Transportation Studies, 6, 4237–4252.

Uslu, A., ÇetiNkaya, C., & Kürşat, S. (2017). Research Article VEHICLE ROUTING PROBLEM IN POST-DISASTER HUMANITARIAN.

Vahdani, B., Veysmoradi, D., Noori, F., & Mansour, F. (2018). Two-stage multi-objective location-routing-inventory model for humanitarian logistics network design under uncertainty. International Journal of Disaster Risk Reduction, 27, 290–306. https://doi.org/10.1016/j.ijdrr.2017.10.015

Vahdani, B., Veysmoradi, D., Shekari, N., & Mousavi, S. M. (2018). Multi-objective, multi-period location-routing model to distribute relief after earthquake by considering emergency roadway repair. Neural Computing and Applications, 30(3), 835–854. https://doi.org/10.1007/s00521-016-2696-7

Verma, A., & Gaukler, G. M. (2015). Pre-positioning disaster response facilities at safe locations: An evaluation of deterministic and stochastic modeling approaches. Computers & Operations Research, 62, 197–209. https://doi.org/10.1016/j.cor.2014.10.006

Vieira, Y. E. M., De Mello Bandeira, R. A., & Silva Júnior, O. S. D. (2021). Multi-depot vehicle routing problem for large scale disaster relief in drought scenarios: The case of the Brazilian northeast region. International Journal of Disaster Risk Reduction, 58, 102193. https://doi.org/10.1016/j.ijdrr.2021.102193

Wei, X., Qiu, H., Wang, D., Duan, J., Wang, Y., & Cheng, T. C. E. (2020). An integrated location-routing problem with post-disaster relief distribution. Computers & Industrial Engineering, 147, 106632. https://doi.org/10.1016/j.cie.2020.106632

Xu, P., Liu, Q., & Wu, Y. (2023). Energy Saving-Oriented Multi-Depot Vehicle Routing Problem with Time Windows in Disaster Relief. Energies, 16(4), 1992. https://doi.org/10.3390/en16041992

Yu, W. (2021). Pre-disaster location and storage model for emergency commodities considering both randomness and uncertainty. Safety Science, 141, 105330. https://doi.org/10.1016/j.ssci.2021.105330

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

Original article

OPTIMIZING THE LOCATION OF MOBILE MAMMOGRAPHY UNITS IN RURAL COMMUNITIES

PREDRAG GROZDANOVIĆ[1], MILICA ŠELMIĆ[2], MILOŠ NIKOLIĆ[3], DUŠAN TEODOROVIĆ[4]

[1] University of Belgrade, Faculty of Transport and Traffic Engineering, p.grozdanovic@sf.bg.ac.rs, 0009-0002-3401-5799
[2] University of Belgrade, Faculty of Transport and Traffic Engineering, m.selmic@sf.bg.ac.rs 0000-0003-2507-3663
[3] University of Belgrade, Faculty of Transport and Traffic Engineering, m.nikolic@sf.bg.ac.rs 0000-0001-5892-8248
[4] Serbian Academy of Sciences and Arts, duteodor@gmail.com 0000-0002-7535-2727

Abstract: *Breast cancer is the most common malignancy affecting women worldwide. Regular mammographic screening plays a crucial role in both prevention and early detection. One of the key strategies in the fight against this disease is to improve access to diagnostic services, especially in underserved areas. Mobile mammography units present an effective solution by enabling service provision in rural and remote regions, thereby reaching women who might otherwise lack access. However, the deployment of mobile units raises the important question of how to determine their optimal locations. In this paper, a bi-objective model is proposed for locating mobile mammography units. The first objective aims to maximize coverage of high-priority settlements, while the second seeks to minimize the total travel distance for all patients. To address the inherent trade-off between these goals, compromise programming was applied as a solution approach. The model was tested on a real-world case study in the Vojvodina region, focusing on settlements with fewer than 10,000 inhabitants.*

Keywords: *Location problem, compromised programming, mobile mammography unit.*

1. INTRODUCTION

Breast cancer remains the most prevalent cancer among women worldwide. In 2022, approximately 2.3 million new cases were diagnosed globally, accounting for nearly one in four cancer cases among women (International Agency for Research on Cancer, 2022). That same year, breast cancer led to approximately 670,000 deaths, underscoring its significant impact on global health (World Health Organization 2024, Zhang et al, 2025).

In Serbia, breast cancer is also a major public health concern. In 2022, there were 4,900 new breast cancer diagnoses among women, making it one of the most common cancers in the country. The age-standardized incidence rate was 86.8 per 100,000 women, placing Serbia among countries with a high incidence rate of breast cancer (International Agency for Research on Cancer, 2022; Zhang et al, 2025)

Despite the implementation of population-based screening programs since 2012, Serbia continues to face challenges in reducing breast cancer mortality. Barriers such as limited access to early detection and treatment services contribute to the high mortality rates observed (Todorovic et al, 2023)

These statistics highlight the urgent need for improved breast cancer prevention, early detection, and treatment strategies both globally and within Serbia. Access to timely breast cancer screening is a critical component in improving early detection rates and reducing mortality. However, healthcare infrastructure and transportation barriers limit access to fixed-location diagnostic facilities in many regions, particularly rural and underserved areas. Mobile mammography units (MMUs) present a flexible and scalable solution to bridge this gap by bringing services directly to communities. This study addresses the Mobile Mammography Location Problem on a transportation network, aiming to optimize the deployment of MMUs in a way that balances healthcare accessibility and operational efficiency.

The objective function is twofold: maximize coverage, measured by the priority of serving patients within an acceptable travel distance, and minimize the travel distance patients must cover to reach an MMU. We aggregate demand at the settlement level and introduce priority scores for each location. These priorities are computed based on demographic factors (e.g., a larger number of older patients or a higher number of women already diagnosed with breast cancer implies a higher urgency for screening).

DRAFT VERSION

To solve this multi-objective combinatorial optimization problem, we develop a mathematical formulation that incorporates spatial constraints, coverage thresholds, and demand priorities. The solution approach is based on compromise programming. The proposed methodology will be validated using real-world data from Serbia (Vojvodina), ensuring that the results are grounded in realistic geographic, demographic, and healthcare conditions. Through this work, we aim to support evidence-based planning of mobile health services and contribute to reducing healthcare disparities in breast cancer screening.

2. LITERATURE REVIEW

Souza et al. (2019) addressed the problem of locating fixed mammography units to maximize total demand while respecting the constraint that women should not travel more than 60 km to reach a facility. Only cities with existing hospital infrastructure were considered as potential locations for the mammography units. The authors developed a mathematical formulation and applied a Variable Neighborhood Search (VNS) algorithm. The proposed approach was tested on real data collected from the Minas Gerais State, Brazil.

De Assis et al. (2025) demonstrate how combining fixed and MMUs can be used to meet demand. A sequential solution approach is proposed, where the location-allocation problem for fixed mammography units is first solved using an exact method, followed by solving the routing problem for mobile units using a heuristic algorithm in municipalities not covered by fixed units. The proposed approach was tested on three different scenarios in the Minas Gerais State, Brazil. The results show that significantly more mobile units are required to fully meet the demand when the existing fixed unit locations must be maintained and users are restricted to their administrative zones, compared to a scenario where administrative boundaries can be crossed and fixed units can be relocated.

De Araujo et al. (2023) developed a hybrid NSGA-II (Non-dominated Sorting Genetic Algorithm II) for the mobile mammography routing problem with two objectives. The problem is categorized as an open vehicle routing problem with multiple depots. It involves a fixed number of depots, each with a limited number of MMUs that can be stationed there. Each mobile unit has a defined capacity and a set of candidate cities it can visit. The objective was to determine the sequence of city visits in a way that maximizes the demand for screening while minimizing the total distance traveled by the mobile units. The Minas Gerais State, Brazil was also used as a real-world case study to test the proposed approach.

Jewett et al. (2018) showed that driving time to mammography units and the number of facilities within a 10 km radius of women's homes influence the frequency of mammography screening among women. The authors emphasize that in rural areas, mammography underutilization is higher due to the lower availability of fixed facilities and longer travel times to reach them. These findings suggest that strategically locating MMUs in rural areas could increase the frequency of screenings.

Komaie et al. (2018) demonstrated that MMU implementation improves screening rates in medically underserved and rural areas. A MMU traveled through the eastern part of Missouri, serving both the urban region of St. Louis and rural areas in the “Bootheel” region. The study showed that MMUs enhance access to screening for individuals in rural locations and those without health insurance, and economically disadvantaged women.

Vang et al. (2018) provided a literature review on mammography covering the period from 2010 to 2018. They emphasize that MMUs can deliver screening services to users in rural areas and socially disadvantaged populations. To increase the effectiveness of mobile mammography, the authors highlight the application of outreach and promotion programs targeting rural communities and women from ethnic minority groups, low-income backgrounds, and those without health insurance.

Young et al. (2020) used optimization models within a Geographic Information System (GIS) to allocate women to nearby facilities where mammography screenings could be performed. In solving this problem, they considered facility capacities and patients' travel times. According to relevant models, screening capacity was insufficient to meet the theoretical demand, given the travel constraints. The case study was conducted in Arkansas, where the majority of women (80%) live within a 30-minute drive of a screening facility. However, most of the unmet demand was located in rural areas. This is yet another study highlighting the reduced accessibility to mammography for women in rural regions. Young et al. (2020) point out that the MMU's deployment could potentially address this issue.

The literature review identified a gap for future research related to the location of MMUs in rural areas, aimed at providing screening services to women living in these regions. The main criterion for selecting the location of mobile units that could be highlighted is the distance to users, either in terms of travel time or geographical proximity.

DRAFT VERSION

3. PROBLEM DESCRIPTION AND MATHEMATICAL FORMULATION

Women in rural areas are less likely to visit hospitals for mammographic screening. To improve accessibility and encourage more women from these regions to undergo examinations, MMUs are utilized. However, deploying MMUs introduces the challenge of identifying optimal locations for their placement, as well as determining how to allocate users to each unit. In the allocation process, each user must be assigned to exactly one MMU. Selecting the best locations for MMUs involves multiple criteria. In this paper, two main objectives are considered: coverage and travel distance. The first objective aims to maximize service coverage by reaching as many patients as possible within an acceptable distance. This is done by maximizing the overall service priority, which is based on the number of potential patients in each settlement. Settlements with more potential patients are given higher priority when determining MMU locations. The second objective focuses on minimizing the total travel distance between users and the MMUs. The number of MMUs to be deployed, denoted as p, is predetermined.

The mathematical formulation of the described problem is in the text to follow.

Sets:

I — a set of settlements where patients reside, $i \in I$

J — a set of potential locations for MMU, $j \in J$

D_{max} – acceptable travel distance for patients,

N_i – set of potential locations ($j \in J$) in which a patient from node i could be served, while traveling a smaller distance than D_{max}: $N_i = \{j \mid d_{ij} \leq D_{max}\}$,

d_{ij} — distance between settlement i and potential mammography site j

p — number of MMUs that have to be deployed

$prio_i \in [0,1]$ — priority score for settlement i, reflecting the urgency of service based on demographic and health data (e.g., settlements with more estimated new cases per year imply higher priority).

Decision variables:

$$x_{ij} = \begin{cases} 1, if\ patients\ from\ settlement\ i\ are\ served\ at\ location\ j \\ 0, otherwise \end{cases}$$

$$y_j = \begin{cases} 1, if\ a\ MMU\ is\ deployed\ a\ location\ j \\ 0, otherwise \end{cases}$$

Objective functions:

$$\text{Max } Z_1 = \sum_{i \in I} \sum_{j \in N_i} prio_i x_{ij} \quad (1)$$

$$\text{Min } Z_2 = \sum_{i \in I} \sum_{j \in J} d_{ij} x_{ij} \quad (2)$$

Constraints:

$$\sum_{j \in J} x_{ij} = 1 \qquad \forall i \in I \quad (3)$$

$$x_{ij} \leq y_j \qquad \forall i \in I, j \in J \quad (4)$$

$$\sum_{j \in J} y_j = p \quad (5)$$

$$x_{ij} \in \{0,1\} \qquad \forall\, i \in I, j \in J \quad (6)$$

$$y_j \in \{0,1\} \qquad \forall\, j \in J \quad (7)$$

There are two objective functions. The first (Z_1) represents the weighted sum of the covered priority scores, encouraging the placement of units near high-priority settlements (at a distance less than or equal to D_{max}). The second objective function (Z_2) represents the total distance traveled by patients to be served. All demand nodes (settlements) must be assigned to exactly one MMU, constraint (3). Constraint (4) indicates that users can only be assigned to a location where an MMU is located. Constraint (5) defines the number of MMUs that should be located. Finally, the definition of the variables is given by constraints (6) and (7).

To handle the multi-objective nature of the problem, namely, maximizing the coverage of high-priority settlements (patients) and minimizing the total travel distance, we used compromise programming to solve the created model by applying the following formula (Coelo et al., 2007):

$$L_p(\vec{x}) = \left[\sum_{i=1}^{k} w_i^p \left(\frac{|f_i(\vec{x}) - f_i^0|}{f_i^{worst} - f_i^0} \right)^p \right]^{1/p} \quad (8)$$

DRAFT VERSION

Where:
$\vec{x}$ – decision vector (solution),
k – number of criteria
w_i – weight for the criterion *i,*
$f_i(\vec{x})$ – the value of the i-th criterion for the solution $\vec{x}$
f_i^0 – ideal value for the criterion *i,*
f_i^{worst} – the worst possible value for the criterion *i,*
$p \in [1, \infty)$ – parameter that determines the type of distance (eg p = 1: Manhattan, p = 2: Euclidean)

By tuning w_1 and w_2, decision-makers can balance the compromise between accessibility for high-priority populations and overall efficiency in patient travel distance. For example, setting w_1=0 and w_2=1 results in a pure coverage-maximization model, while setting w_1=1 and w_2=0 results in a pure *p*-median model. Setting w_1=0.5 and w_2=0.5 make balances both objectives. This flexible formulation allows the model to adapt to different public health policies or operational constraints based on the region or strategic goals.

4. CASE STUDY

To test the created model, the territory of Vojvodina was analyzed. According to data from the Institute for Public Health of Serbia "Dr. Milan Jovanović Batut", the incidence rate of breast cancer in women in Vojvodina was 78.9 per 100,000 women, while the death rate was 22.8 per 100,000 women (Institute of Public Health of Serbia, 2023). In Vojvodina, organized mammography screening is carried out for women aged 50 to 69 (Ministry of Health of the Republic of Serbia, n.d.). Also, MMUs provide free examinations in rural areas. For example, within the "First Mammography" project, MMUs examined 18,800 women, of whom 147 women were diagnosed with breast cancer (Ministry of Health of the Republic of Serbia, 2021). In this paper, we considered all settlements from Vojvodina with up to 10,000 inhabitants. There are 11 settlements as shown in Table 1. Those 11 nodes of demand (settlements) represent, at the same time, potential locations of MMUs. All women residing in a given settlement and belonging to the target screening population were aggregated into one demand point. An approximate estimate of the number of women who suffer from breast cancer annually in these 11 settlements in Vojvodina was made as follows: Population of women in towns (according to the 2022 census) taken as half of the total population (Statistical Office of the Republic of Serbia, 2023) (Table 1); The incidence rate in Vojvodina was used: 78.9 new cases per 100,000 women per year (Dugandžija et al, 2020) (Table 1). The distance between urban settlements was taken from Google Maps.

Table 1: Input data

Settlement	Population	Woman (~50%)	Incidence	Estimated new cases per year *ENC*	$prio=\frac{ENC}{\sum ENC}$
Crvenka	7 232	3 616	78,9	~2,85	0.095
Odžaci	8 811	4 405	78,9	~3,48	0.116
Palić	7 771	3 885	78,9	~3,06	0.102
Ada	9 564	4 782	78,9	~3,77	0.125
Mol	6 009	3 005	78,9	~2,37	0.079
Novi Kneževac	6 960	3 480	78,9	~2,74	0.091
Čoka	4 028	2 014	78,9	~1,59	0.053
Titel	5 294	2 647	78,9	~2,09	0.070
Žitište	2 903	1 452	78,9	~1,15	0.038
Sremski Karlovci	8 839	4 420	78,9	~3,49	0.116
Beočin	8 839	4 420	78,9	~3,49	0.116

To solve the proposed model, we used ILOG CPLEX 12.1 on a 64-bit ACER computer with an Intel(R) Core (TM) i5 2.50 GHz processor and 8 GB of RAM. The model was tested on the example of Vojvodina for a different number of MMU locations. A detailed analysis of the results is presented for one example (location 5 MMUs).When locating MMU, one should aim for the MMU to be less than D_{max}=25 km away from users from urban settlements with a higher priority. The test results are given in Table 2. The model was tested for different values of the weighting coefficients. The first example represents the borderline case (w_1=1; w_2=0) in which the highest value of the objective function (Z_1=0.893) is obtained, which refers to the priority of patient service, but also the highest total distance of users from MMUs (Z_2=311.6 km). For two demand nodes, that is, patients from them, the condition of the maximum acceptable distance (D_{max}=25 km) that users are

DRAFT VERSION

willing to travel to be served is not met. Patients from those nodes have to travel more than 25km to the nearest MMU to be served. In the second borderline situation (w_1=0; w_2=1), the smallest total distance of patients to MMUs is obtained (Z_2=149.8km), but also the lowest patients' service priority (Z_1=0.766). In this case, for three demand nodes, the nearest MMU is at a distance greater than D_{max}. To achieve a compromise between these two conflicting criteria, compromise programming was used. In the case when the criteria are of the same importance (w_1=0.5; w_2=0.5), a solution was generated that differs by only 2.8 % from the ideal solution in terms of service priority, i.e., 2.3 % from the ideal solution in terms of the distance criterion.

Table 2: Test results in case of locating 5 MMUs

Br.	w_1	w_2	Mobile mammography locations	Number of urban settlements (D_{max}>25km)	Z_1	Z_2	L
1	1	0	Crvenka, Odžaci, Palić, Čoka, Beočin	2	0.893	311.6	0.000
2	0	1	Odžaci, Ada, Novi Kneževac, Titel, Beočin	3	0.766	149.8	0.001
3	0.5	0.5	Odžaci, Palić, Čoka, Titel, Sremski Karlovci	2	0.868	153.3	0.109

Spatial distribution of demand nodes (potential locations of mammograms), as well as test results for values w_1=0.5 and w_2=0.5, are shown in Figure 1. MMUs are located in 5 locations (Palić, Čoka, Odžaci, Sremski Karlovci, and Titetl). Patients who do not have an MMU in their node are allocated to the MMU closest to them (Figure 1). Figure 1 shows that patients from two demand nodes must travel a distance greater than D_{max} to receive the service (Crvenka to Odžaci and Žitište to Titel).

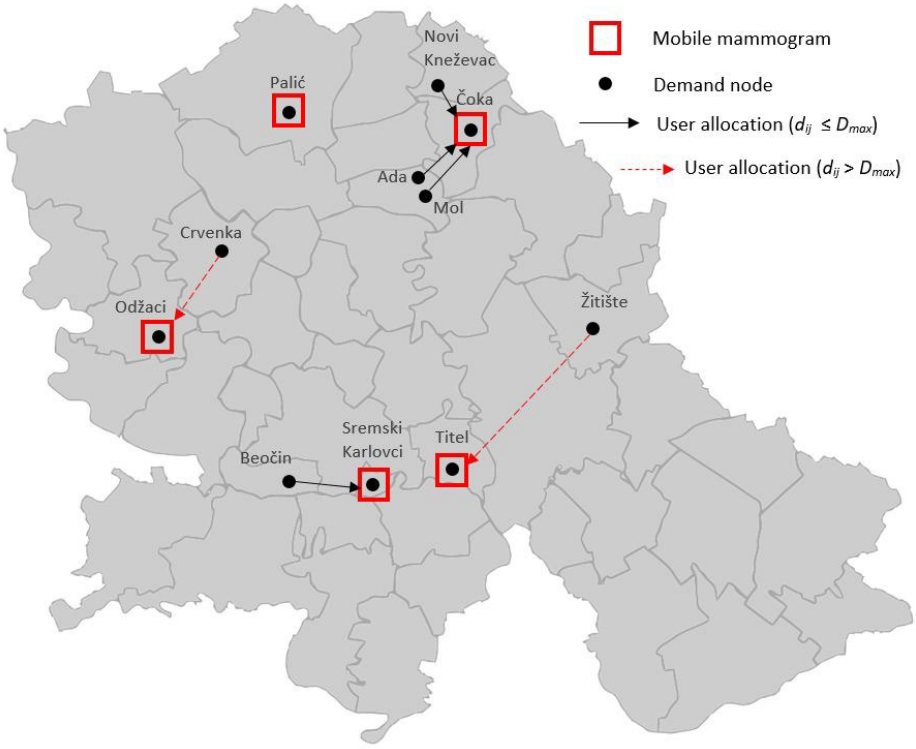


Figure 1: Example of locating 5 mobile mammograms and patient allocation

5. CONCLUSION

In recent years, examples of various forms of cancer have become more frequent. The chance of curing cancer increases if it is detected in time. That is why it is very important to go for regular check-ups. The distance of the facilities where the control examination can be performed is one of the factors that affects the regularity of visits to the doctor. Bearing this in mind, examples of the application of mobile facilities that provide various medical services in rural areas, such as taking samples, various forms of imaging, providing therapies, etc., are becoming more and more common. The use of mobile medical facilities necessitated the creation of a model that will deal with determine their location.

This paper presents a novel location model for determining the optimal placement of MMUs, applied within the healthcare system of Vojvodina. The proposed model offers several key advantages: it ensures short computational times, achieves a balanced trade-off between maximizing the coverage of high-priority urban settlements and minimizing the travel distances for patients, and is highly practical and easy to implement, as demonstrated in the case study conducted on urban settlements in Vojvodina.

Future research directions may include the following: (1) extending the application of the model to the entire territory of the Republic of Serbia and analyzing the obtained results; (2) enhancing the model by introducing capacity constraints, which would allow for a more realistic allocation of users; (3) incorporating a broader set of criteria when determining service priorities, such as patient age, history of breast cancer in the settlement, proportion of women with health insurance, and others; (4) applying metaheuristic algorithms to efficiently solve larger problem instances.

LITERATURE

Coello, C. A. C., Van Veldhuizen, D. A., & Lamont, G. B. (2007). Evolutionary algorithms for solving multi-objective problems. Springer.

De Araujo, T. G., Penna, P. H. V., & Souza, M. J. F. (2023). A hybrid NSGA-II algorithm for the bi-objective mobile mammography unit routing problem. In International Conference on Artificial Intelligence and Soft Computing (pp. 339–351). Cham: Springer Nature Switzerland.

De Assis, R. D. S., de Campos, M. V., Souza, M. J., Souza, M. A., de Siqueira, E. C., Wanner, E. F., & de Souza, S. R. A. (2025). Solution procedure for fixed mammography unit location-allocation and mobile mammography unit routing problems.

Dugandžija T., Sekerija, M., Hinic, N., Rajcevic, S., & Kusturica, M. P. Trend analyses of breast cancer incidence and mortality in Vojvodina. JBUON. 2020; 25(2):655–661.

Google Maps. (2025, June 25). Retrieved from https://www.google.com/maps/@44.7698133,20.491812,14z?hl=sr [07.20.2025]

International Agency for Research on Cancer. (2022). *Global Cancer Observatory: Cancer Today – Serbia Fact Sheet*. World Health Organization. Retrieved from: https://gco.iarc.fr/today/data/factsheets/populations/688-serbia-fact-sheets.pdf [07.20.2025]

International Agency for Research on Cancer. (2022). *Global Cancer Observatory: Cancer today*. World Health Organization. Retrieved from: https://gco.iarc.fr/today [07.20.2025]

Jewett, P. I., Gangnon, R. E., Elkin, E., ... & Trentham-Dietz, A. (2018). Geographic access to mammography facilities and frequency of mammography screening. Annals of Epidemiology, 28(2), 65–71.

Komaie, G., Lyons, S., Appleton, C., Glover-Collins, K., Marx, C., & Colditz, G. (2018). Abstract C73: A mobile mammography outreach program to increase screening in medically underserved areas in Missouri. Cancer Epidemiology, Biomarkers & Prevention, 27(7_Supplement), C73–C73.

Ministry of Health of the Republic of Serbia. (2021). Report on the implementation of the "First Mammography" project. Belgrade: Ministry of Health.

Ministry of Health of the Republic of Serbia. (n.d.). National program for early detection of breast cancer. Retrieved from http://www.zdravlje.gov.rs

Souza, M. J. F., Penna, P. H. V., Moreira de Sá, M. V. S., & Rosa, P. M. (2019). A VNS-based algorithm for the mammography unit location problem. In International Conference on Variable Neighborhood Search (pp. 37–52). Cham: Springer International Publishing.

Statistical Office of the Republic of Serbia. (2023). Census of population, households and dwellings in 2022 – Book 1: National or ethnic affiliation by settlements. Belgrade: Statistical Office of the Republic of Serbia.

Todorovic, J., Stamenkovic, Z., Stevanovic, A., Terzic, N., Kissimova-Skarbek, K., Tozija, F., et al. (2023). The burden of breast, cervical, and colon and rectum cancer in the Balkan countries, 1990–2019 and forecast to 2030. Archives of Public Health, 81(1), 156.

Vang, S., Margolies, L. R., & Jandorf, L. (2018). Mobile mammography participation among medically underserved women: A systematic review. Preventing Chronic Disease, 15, E140.

World Health Organization. (2024). Breast cancer: prevention and control. Retrieved from https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Young, S. G., Ayers, M., & Malak, S. F. (2020). Mapping mammography in Arkansas: Locating areas with poor spatial access to breast cancer screening using optimization models and geographic information systems. Journal of Clinical and Translational Science, 4(5), 437–442.

Zhang, Y., Ji, Y., Liu, S., Li, J., Wu, J., Jin, Q., et al. (2025). Global burden of female breast cancer: New estimates in 2022, temporal trend and future projections up to 2050 based on the latest release from GLOBOCAN. Journal of the National Cancer Center.

DRAFT VERSION

MANAGEMENT

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

ANALYSES OF THE CONVERGENT TELCO-INSURANCE BUSINESS MODEL: MANAGERIAL PERSPECTIVES IN NORTH MACEDONIA

METODIJA MIRCHEV[1], VIOLETA CVETKOSKA[2], IGOR IVANOVSKI[3]

[1] A1 Macedonia, Skopje/Ss Cyril and Methodius University in Skopje, Faculty of economics-Skopje Ph.D. candidate, mircev.metodija@gmail.com
[2] Ss Cyril and Methodius University in Skopje, Faculty of economics-Skopje, violeta.cvetkoska@eccf.ukim.edu.mk, 0000-0002-0472-5522
[3] Ss Cyril and Methodius University in Skopje, Faculty of economics-Skopje, ivanovski@eccf.ukim.edu.mk, 0009-0003-4856-3312

Abstract: *The aim of the paper is to offer contribution to the national and regional theoretical and empirical framework and literature, by analyzing the aspects and factors for creation, implementation and assesment of the convergent, cross-industries joint business model, based on the open innovation model. The paper particularly analyzes the factors of the convergent Telco-insurance business model, in general, as well as for the possibilities at the national market. We focus on evolving opportunities, critical factors and managerial challenges and perspectives offered by a national survey among Telco, the insurance and the banking industry's top managers in the country. We asses them as a valuable input for current and further research focused on the perception of the changes, new value creation, managing the risks and the factors influencing the creation, initialization and performance of such a model. The paper is a part of the ongoing and broader first of a kind national academic work and empirical, qualitative and quantitative, analyses of the open innovation and synergi based convergent model.*

Keywords: *Telco industry, insurance industry, open innovation, business model, managerial perspective, survey.*

DRAFT VERSION

1. INTRODUCTION

The world, in an economic and socio-cultural way, has changed drastically. The only axiomatic conclusion is that the changes are intense, irreversible and continuous. All the other economic theories are under tests, thus defining the current approach in theoretical and empirical evolution assessment as dynamic and highly transformational. As no industry is left intact, the changes should be evaluated interdisciplinary and in a convergent manner. In that regard, we particularly focus on the disruptive changes and development of the convergent business model between the Telco and insurance companies, analyzing the managerial perspectives of the top managers in the telecommunication, insurance and banking industry in North Macedonia. The work is a part of the broader qualitative and quantitative analyses of the future of the open innovation based model rooted in the initial and first synergy of the Telco and insurance companies in the country. The moments of the close past linked to the disruptive consequences of the COVID-19 pandemic, related to the exponential increase of the usage of voice, data and digital platforms by the customers, are now absorbed by the new perspectives of the ongoing AI revolution. If this was previously understood as normal in the world of the Telco industry, nowadays all the rest of the industries really do not have a choice anymore. Insurers started to seize the unique and sole opportunity to capitalize on this shift. Main actions underline developement of new innovative business models with platform providers, expanding their reach and delivering value propositions to the platform customer. Amongst strong partner candidates for insurers, Telcos have always been a valuable and logical option. It is a result of their large subscriber base, strong channels, strong approach in customer profiling & analytical capabilities and digital assets with an ability to integrate customer journeys with insurance offers. The synergy gains are twofold. The Telco's continue expanding their core asset as the customers and the digital portfolios by providing financial and insurance services, throughout the proprietary platforms and strategic partnerships (Research and Markets, 2024) and the insurance companies are increasing the cost effective distribution and enhance the most important factor of selling the products and services and increasing the insurance pool in highly competitive market by nature. From the perspective of Telco's, a

partnership with insurance players helps create a strong revenue stream and improve customer loyalty with a convergence approach with insurance services supply. These evolving models are effective and are interspined by the forms of cooperation and competition. The key opportunity is the process of the digital transformation which offers variety of broad ecosystems of diverse players that co-evolve constantly via collaboration and competition in offering industries core and additional financial services (Baumann et. al., 2023).

Both insurers and Telco's are under strong pressure to reinvent and reposition their digital businesses by possible new partnership models. The recipe for a successful insurer-telco partnership involves two key ingredients: value propositions and operating model.

Joint value propositions can be offered based on the customer's digital journeys with the Telco's, as well as by targeting specific customer segments using their behavior on the Telco's platforms. In terms of operating model, Telcos should consider being a part of a fully integrated "true partnership" model with insurers. Careful consideration is required in setting up a joint coop model and the commercial structure of a new strong partnership. A partnership will require both parties to develop new capabilities and integrate but also develop into future sustainable and open innovation business model (Chesbrough, 2006).

The paper is structured in a concise way, excluding the literature overview as there isn't adequate and vast national academic literature input for the particular open innovation model of Telco-insurance model. Among the rest, it is our motivation to fill in the gap. The first chapter is focused on the analysis of the evolution, nature and implications of the Telco and insurance industry rapid changes, as well as with the key elements of the transformation and the current and future possibilities. In the following chapter, we position the results of the conducted survey among the top managers of the Telco, insurance and banking industry in North Macedonia, followed by conclusion and future considerations.

2. TRANSFORMATIONAL CHALLENGES OF THE TELCO AND INSURANCE COMPANIES: THE JOINT MODEL FORWARD?

Telco had and are still facing the strong disruptive power of the start-ups threatening their very core. Just as free messaging platforms like WhatsApp and Instagram has killed the usage of sms as proprietary telco platform, the free calls market have been diminishing the international and roaming calls and allows evolving dominant role of the streaming platforms on tv/video market. That was the final call for the companies to open and evolve, by new approach and speed in new product development and delivery, investment efficiency and cross-industry supply. Again it always comes to the main issue: who will win the customer (PWC, 2021). Which service provider will be the one having direct interaction, contractual obligations and benefits, thus the main goal for long-term customer ownership (Figure 1).

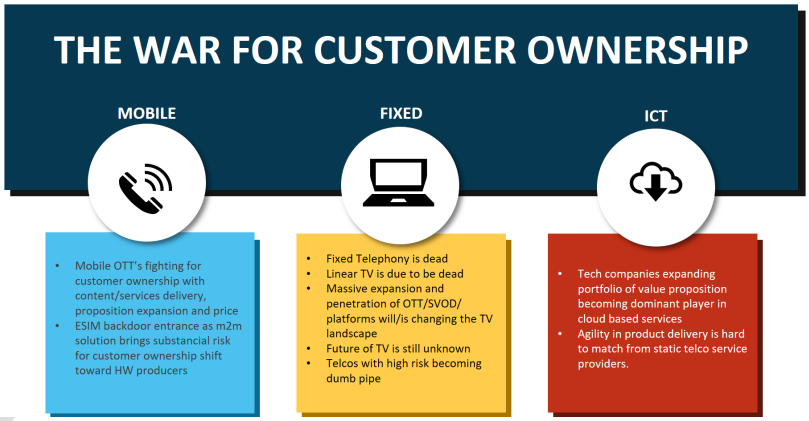


Figure 1. Customer ownership in the Telco's industry

To win customers loyalty and strong long-term relationship, Telco must expand their ecosystems of value propositions. Moving beyond core products and services is inevitable. It is not just a matter of competitivity, but it is also important for organic growth which is endangered (Artur D. Little, 2019). The Telco stocks' returns have been failing as the digital transformation, introduction of 5G, the process of M&A, market saturation, hypercompetitive and regulatory policies are putting high pressure on the Capex and Opex (Figure 2).

DRAFT VERSION

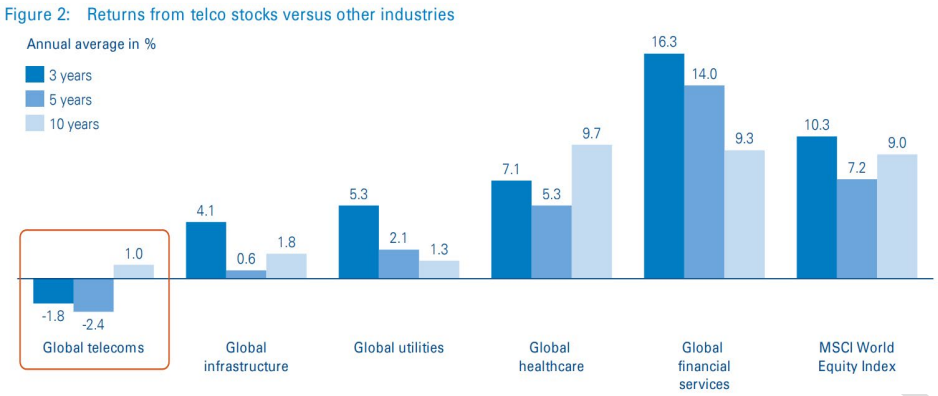


Figure 2. Comparative industries stock returns

That defines the business strategy of new markets entrance and creating new and open business models, via strong partnership and strong collaboration with financial companies and with StartUp's, based on the Open Innovation Model.

In parallel, the insurance industry has finally started a massive digital reorganization and adaptation to the new market environment evolution, as an increasing number of consumers are moving online. The established trends for health and life products changed the consumers demand needs, as well as the per-need-based insurance linked to the new, digitally supplied, products or service, the new risks and increased claims and the introduction of the AI in the business models, are redefining the insurers past static position. All mentioned is transforming them into the new stage of "predict and prevent" compared to the old "detect and repair" (McKinsey, 2022). Does that define safe ground for a sustainable and joint convergent Telco-insurance model? Does the new time, exponentially enhanced with AI and robotics, merge the insurance companies as the oldest big data entities and the Telcos as the economy's and people's highways? And how to cope with the newly expected disruptors? Clearly, there are more pros, even empirically proved, that can support the win-win game for the two industries and confirm the success of the Christensen's theory of disruptive innovation (see Christensen, 1997) and evolution to a productive disruption. Both industries are based on the fundamentals of managing risk and leveraging data to deliver value. They have large customer base ownership. They need, posses and convert large data into structured, personalized value of products and services. They require advanced analytical systems, for predictive analytics and market churn, as the retention of the customers is of paramount importance. The issue of transferring the risk in relation to the business processes and customer cycles need's further clarification from the time, cost and satisfaction perspective in the joint model. Since the customers are centered (Synpulse, 2024), the synergy model, in various applications, could add the new impetus to the insurance companies insufficient proactive role and support Telco's customers risk protection in return (Figure 3).

Figure 3. Customer centered cross-industry relations

The convergence isn't highly current model, as it isn't a merely easy job done as someone might expect. The issues of national markets conditions, regulation and behavioral factors must be taken into consideration. However, the best possible starting model could be defined by using the ecosystem strategy (Intellias,Makarchuk, 2021). Namely, the model in which the Telcos act as orchestrators by the advantage of owning the customer relationship, digital interface, and data infrastructure, and in which the Insurers act as capability partners, providing the risk expertise, regulatory compliance, and claims management, could work

DRAFT VERSION

effectively. This focused, structured and joint genuine capacities and strengths, enables the development of modular, targeted and scalable services that can be competitive across the communication and insurance markets and customer segments (McKinsey, 2022).

3. MANAGERIAL PERSPECTIVES FOR THE INNOVATIVE ECOSYSTEMS AND CONVERGENT BUSINESS MODELS: THE CASE OF NORTH MACEDONIA

The following analyzes is a result of the ongoing qualitative and quantitative research for the innovative ecosystems between the Telco, insurance and banking industry, with focus on North Macedonia. The pioneering business model was established in 2022 between the largest Telco operator A1 Macedonia (part of Telekom Austria Group) and Triglav non-life insurance, a leading non-life insurance company at that time. In parallel to the rest of the vast and data quantitative analyses, at the end of 2024, we have conducted a survey, with a structured questionnaire using using Likert-type scale, with 5 level (1 = Strongly Disagree and 5 = Strongly Agree) with 60 respondents, part of the top-management structures of the Telco operators, non-life and life insurance companies and commercial banks in North Macedonia. The overall form (Q&A) is open by the authors and available upon request, as due to the lenght limitations of this paper, it could not be positioned as an appendix. Our goal was to examine top managers positions, opinions, expectations, beliefs and attitudes towards innovative behavior, convergent models, risks and crucial factors of creation, implementation and sustainability of an innovative and convergent cross-industry business model. The respondents were adequatelly distributed across the industries, insurance (40%), Telco (30%) and banking (30%), with high levels of education (bachelor, master, exectutive education, Ph.D) and with long term work expirience of 15+ years (70%). Their positions are within the managing and supervisory Boards, as well as at the managing positions at sales, NPD, IT, risk, marketing and strategy departments.

The questionary was divided into 4 main sections and consisted of 30 questions. In the first section of measuring the stands towards the intensity of changes, the respondents find the current changes in the respective industries as revolutionary. 69% of the respondents believe that the current transformations represent a *revolution* in the way industries operate, while only 31**%** view them as part of a slower, evolutionary process. This emphasizes the process of fundamental reconfiguration, not just incremental improvement. The vast majority of almost 90% find that the changes permanently transform the industries, that they are globally introduced rather than locally, that the main factors come both from the consumers perspectives and structural changes. They are claiming that witness and recognize the current creation of joint models and strategic partnerships between the three industries.

In the second section, dedicated to the issues of value creation, the responses are almost unanimous. 90% of the respondents believe in the business success of the joint models and strategic partnerships, as well as they find ii beneficiary for the consumers, for the growth of the industries in relation to the higher income/profit and for the competitiveness of their companies. The latter is challenging to analyze more deeply, as for the Macedonian small market, with limited Telco competition, the issues of relation between competitivity and profitability imposes careful approach. In any case, they perceive the evolution of the current business models as convergent ones as irreversible and moreover as sustainable on the market.

In the third section, related to the perception and awareness of the risks in general, the distribution is slightly different from the first two sections. 66% of the respondents find the development and implementation of the business models burdened with high risks. In a contrary, 25% of respondents partially don't agree with that. The main risks for the respondents are: 72,9% as a result of the demand factors susch as higher product complexity, lack of trust in the other industry's partner, complex process of new product acceptance/identification etc; 83,4% stresses the regulatory and supervision framework and administrative barriers; 80,9% emphisize the managerial and internal procedures allignment for the new product launch (commercialization, different organizational cultures) and 72,9% points the internal employees' resistance to change, esspeacially to a complex new product development.

The final section determines the top managers attitudes towards the key factors for initialization, preparation and launch of the convergent business model. 58,3% think that the new model should be developed internally, in comparasion to the 29,2% that find it more sucssesfull if imported from outside, one of the key issues in the model of open innovation. Evaluating the most critical factor for the introduction of the new model, the managers disperse the imporatance among the top management (33,3%), regulatory bodies (31,3%) and employees (25%). Finnally, the respondents were asked to rank the importance of offered additional 5 factors for the convergent model succsess, by grading each of them from 1=lowest to 5=highest importance. They put the highest importance to the vision, strategy and leadership (95,4%), followed by mutual cross-

DRAFT VERSION

industry companies'trust (81.3%), proffitabillity and sustainability of the model (80,3%), human capital factor by 73% and lastly the investment with 52,1%.

As a wrap, the respondents, by 92%, expect that the convergent models would be created and operational in the following period of 1-5 years, namely 2025-2030. In that relation, the respondents find the national Telco industry as most favorable for convergent model launch (54,2%), followed by the insurance industry (25%) and banking industry (18,8%).

4. CONCLUSION

The paper aimed to contribute to the challenging academic and expert debate on the development of convergent business models based on open innovation with primary focus on the model of Telco and insurance companies, which are still rarely introduced. The motivation is high for the developing economy in North Macedonia. The theoretical findings comply to contemporary economic science. The used quantative methodology is part of a broader and ongoing research using the primary data from the first convergent model between Telco and insurance non-life company in thenational market. The paper confirms the great potential for joint business model between the two analyzed industries that can create new value propositions to the customers and enhance the classic business model. The Telcos disruptors and insurance companies' long-lasting challenges of proffitabilitty and product development, could be overcome by synergies and the adequate use of the new technologies in a single innovative model. The managerial perception and perspective for the positive outcome models of such kind are high. The perception of the key factors for such model is evolving. Currently, the managers prioritize the organizational and regulatory factors in comparison to the classic economic factors.

The limitations are present. The survey was orientated towards the top managers, excluding the human capital in the middle management that is substantial for the convergent model operation and development. The national academic participation in the process is marginal contrary to the positive cases of open innovation models across the world. The managerial perception defines the demand side as a primary one and sets less focus on responsibility and actions on the supply side in the business model development. The results of the survey are understood as a starting point for futher research, on a national level, with an inter-cross industry approach, that would include the consumers and the members of the different regulatory bodies. Furthermore, the more enhanced data series of the ongoing business model will allow profound and in-depth data and business analytics, thus offering first conclusions for the pioneering Telco-insurance model in the country.

REFERENCE

Artur D. Little. (2019). Embracing the future- How can operators embrace telecom disruption? Telecommunications industry report

Baumann, N., Renier, J., Tang, M., French, C., & Sanchez Soriano, G. (2023). The ecosystem imperative: Digital transformation of financial services and moving from Open Banking to Open Data. Institute of International Finance (IIF) and Deloitte.

Chesbrough, H. W. (2006). Open Innovation: The New Imperative for Creating and Profiting from Technology. Boston, MA: Harvard Business School Press.

Christensen, C. M. (1997). The Innovator's Dilemma

Intellias,Roman Makarchuk.(2021). Reinventing Telcos into Data-Driven, Customer-Centric Ecosystems

McKinsey (2022). Insurance 2030: The Impact of AI on the Future of Insurance

PWC.(2021). Perspectives from the Global Telecom Outlook; Dr.Florian Grone, Wilson Chow, Russell Taylor

Research and Markets. (2024). Telco, Fintech, and Insurtech Portfolio Evolution and Positioning Strategies (Report No. 6029446).

Globe Newswire Press Release: The convergence of telecommunications, fintech, and insurtech as key driver for digital services portfolio expansion

Synpulse. (2024). Embedded Insurance:Unlocking Opportunities for Telecom

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Review article

CROWDFUNDING ART PROJECTS: A SYSTEMATIC LITERATURE REVIEW

DRAGANA PANTIĆ[1], DRAGAN BJELICA[1]

[1] University of Belgrade, Faculty of Organizational Sciences, Belgrade, dragana.pantic@fon.bg.ac.rs, 0009-0000-8965-5254,
[2] University of Belgrade, Faculty of Organizational Sciences, Belgrade, dragan.bjelica@fon.bg.ac.rs, 0000-0003-0203-2877

Abstract: *Crowdfunding is becoming an increasingly prevalent method for financing projects in arts and culture. This phenomenon involves gathering financial resources via the Internet from various donors to support the work of organizations or individuals, enabling the diversification of funding sources and greater community involvement. The research aims to understand the challenges, opportunities, and trends identified through a systematic review of academic literature on financing art projects using the crowdfunding model. This study examined 20 scholarly papers indexed in the Web of Science (WoS) database, including 2 conference papers and 18 papers published in journals. The results indicate that the academic literature on this phenomenon is recent, showing a growing trend in the volume of research. The analyzed studies employed various methodologies, with case studies and both qualitative and quantitative analyses being the most dominant. The contribution of this work can be seen in its systematic review and in-depth analysis of the relevant literature closely related to the crowdfunding model in art. The study has identified the need for further research into specific contexts and local practices in less-researched environments.*

Keywords: *Crowdfunding, art projects, literature review.*

DRAFT VERSION

1. INTRODUCTION

Financing creative projects, especially when they are not part of an institutional framework, poses numerous challenges and risks (Rijanto, 2022). Technological advancement, rising digitalization trends, and frequent financial issues require that artists go beyond traditional financing methods (Foà, 2019; Guimarães & Maehle, 2025). By analyzing the positions of artists in times of social and economic instability, such as wars and pandemics, some authors (Kovpak & Lebid, 2022) have identified several principal problems, among which one may single out artist emigration, significant drops of income in the creative sector, and the discontinuation of financial activities. These findings suggest that artistic activity is vulnerable in times of crisis and underline the need to develop sustainable models of art support. Crowdfunding, the practice of gathering financial resources from a number of individuals or organizations (Chalmeta et al., 2025; Hossain & Oparaocha, 2017) is able to transform the way in which art projects are produced (Marlett, 2015). Creative artists present their projects and ideas on platforms, using photographs, video materials and text, so as to invite the audience to support their initiatives. In other words, with the help of technology, crowdfunding puts together creative artists and the community, based on common values (Hossain & Oparaocha, 2017). Crowdfunding does not only represent a financing model but it also supports deeper audience involvement, by means of direct interaction between the artist and the community. The process results in fostering dialogue and cherishing common values. Some authors (Dalla Chiesa & Dekker, 2021) think that crowdfunded financing has a positive influence on the visibility of users, i.e. artists, and on their reputation.

The goal of the present paper is to review the most common models in artistic practice, their advantages and limitations, with a particular focus on the importance of the community for the success of the project. Relying on the overview of existing materials and empirical studies available so far, we consider how artists perceive the crowdfunding model and its long-term sustainability in an interdisciplinary context.

2. METHODOLOGY

In accordance with the research goals set above, PRISMA methodology has been used (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) as it enables a transparent and systematic approach to reviewing the relevant literature (Page et al., 2021). The analysis encompasses papers indexed in Web of Science (Core Collection) published between 2014 and 2025, looking into crowdfunded financing of

art projects. Key words used to search titles, abstracts, and full texts were *crowdfunding* and *art project*, which resulted in a list of 326 papers. This number was further reduced through the incorporation in our study of only those articles published in journals and conference proceedings, based on the assumption that these papers had undergone a stricter peer review process, resulting in their added scientific validity. Likewise, only literature in the English language was considered. After these additional criteria were applied, we ended up with 272 papers to analyze. In the next step we used a screening tool, aimed at the removal of duplicate texts, and also of papers that, in terms of the combination of titles, abstracts, and full texts, were not relevant to our research topic. The research sample comprised 20 papers, of which 18 were journal articles and 2 were contributions from conference proceedings.

3. RESULTS

In the studies that we have analyzed, authors consider various organizational and financial aspects of artistic activity, such as music, theatre, animation, and film, within a broader cultural and geographic context. The most commonly used platforms are Kickstarter, Tumblbug, and Voordekunst. These platforms can be viewed as financial tools, but also as instruments in a broader social context, i.e. as important elements in community buildup (Foà, 2019). Studies have encompassed various aspects of crowdfunding in art projects – from types of campaign creation and factors influencing their success, to financing models and the attitude of artists to crowdfunding strategies, all the way to a specific focus on the relationship between the creative artist and the audience donating funds.

Artists and donors do not start working together solely based on market principles. Rather, their cooperation is largely grounded in common values, such as social engagement, art promotion, and the buildup of collective identity (van den Hoogen, 2020).

3.1. Peculiarities of crowdfunding art projects

Our analysis of studies presented in Table 1 has enabled us to identify the main advantages and limitations of the crowdfunding model. The literature overview suggests that the authors have primarily studied the suitability of crowdfunding within the nonprofit sector.

Table 1: Papers identified in the systematic literature review

Author/year	Research goal	Methodology	Key findings
(Galuszka & Bystrov, 2014)	An analysis of crowdfunding as an alternative to the traditional financing of music production	A case study (the MegaTotal platform)	Project success largely depends on the number and frequency of donations and on active communication with the audience. Investors are often motivated by emotional connection with the participants, rather than profit.
(Foà, 2019)	A research of crowdfunding of art projects which specifically accentuate social and cultural values, as well as of marketing strategies and business models	A multiple case study	Crowdfunding campaigns pose risks and challenges, but also provide an opportunity for a specific model based on values, rather than product functionality.
(Wang, 2016)	A study of the potentials for a sustainable financing model based on the Kickstarter platform in music industry	Combined approach	Campaign success to a large extent depends on the already developed contact network. Social networks represent the primary channel for communicating with the audience.
(Alexiou et al., 2020)	A study of how the success of crowdfunding campaigns influences the overall fundraising for nonprofit organizations in performing arts	A quantitative data analysis (the Kickstarter platform)	Younger art organizations are more vulnerable to risks posed by crowdfunding, while for older organizations this financing method provides an additional fundraising opportunity.
(Villén Higueras et al., 2020)	An analysis of the influence of crowdfunding on project implementation in Chinese animated cartoon industry	A qualitative analysis (exploratory and descriptive methodology); case studies	Crowdfunding does not aim to fully cover production costs, but it represents an important project promotion strategy.
(Loriguillo-López, 2017)	Testing the possibilities for financing beyond the traditional framework, with a focus on crowdfunding Japanese animated cartoons	Multiple case studies	Although it cannot replace traditional investment, this model offers an opportunity to test the market and ideas; it also allows more creative freedom and more dynamic communication with the audience.

DRAFT VERSION

(Fanea-Ivanovici & Baber, 2021)	Suggesting a model to finance films and web TV shows that can be used in practice .	A literature overview	The authors have developed a reward-based model consisting of nine phases which, in addition to involving the community to raise the funds, allows that the project idea be tested in a marked-based environment.
(D'Amato & Cassella, 2021)	A study of how various hidden mechanisms on crowdfunding platforms influence the success of the campaign.	A qualitative analysis (interview, discourse analysis)	The authors report that the platform inheres a selection process, which prioritizes projects that already exert social influence.
(Galuszka & Brzozowska, 2016)	Analysis of the concept of gift economy in the context of crowdfunding	A qualitative analysis (interview)	Artists offer gifts that require an amount of effort while donors' gifts are commercial in nature. The platform is not a mere market mechanism, but also provides room for enhancing the bond between the artist and the community.
(Guimarães & Maehle, 2025)	Analysis of crowdfunding aspects beyond the mere gathering of financial resources	A multiple case study	Crowdfunding is particularly relevant to young artists who aspire to position themselves in the market and win recognition, while more experienced artists use this tool in a better thought-out, strategic manner

3.2. Perceptions of participants in crowdfunded art projects

Table 2 presents the most important results of prior studies, with a particular focus on the attitudes of creative artists on the one hand, and donors (i.e. the audience) on the other. In their research, the authors have strived to understand the motivation for participating in crowdfunding campaigns. However, their results also suggest some limitations of this model, due to which the model itself is often perceived as just a temporary solution.

Table 2: Papers identified in the systematic literature review

Author/year	Research goal	Methodology	Key findings
(Dalla Chiesa & Dekker, 2021)	A study of the role and influence of crowdfunding on artist projects and careers	Inductive methodology and descriptive analyses (the data were collected by interviewing artists who had used the Voordekunst platform)	The artists cannot reach new audiences easily, which is why they mostly rely on the current network of acquaintances. Platforms enable them to establish professional distance.
(Galuszka & Brzozowska, 2017)	An analysis of the relationship between artists and donors within an equity/royalty based model on a crowdfunding platform.	A qualitative analysis (the data were obtained by interviewing platform users)	In parallel, crowdfunding helps empower artists, but may also result in their additional exploitation. Some negative effects have been identified: the model prioritizes early investors; in spite of the possibility of participation, creative cooperation often fails to materialize. However, the model still offers transparent communication.
(Fanea-Ivanovici, 2019)	Testing the attitudes of Romanian film industry workers on crowdfunding as a possible means of financing film projects	Combined approach (a survey and an interview)	Most film producers in Romania are not sufficiently familiar with this financing model, at least not enough to use it in practice. Some potentials of crowdfunding have been identified yet they require changes in the regulations.
(Dalla Chiesa, 2022)	A study of how artists perceive crowdfunding, with particular attention paid to the analysis of advantages and limitations	A qualitative analysis (an interview)	Most artists view crowdfunding as a temporary rather than long-term mechanism, to be used when other forms of financial support are unavailable. The model's lack of sustainability and insufficient distance from the audience have a negative influence on the artists' readiness to repeatedly participate in this practice. The model also poses additional risks to artists, since they often need to invest their own funds in campaigns.
(Davidson & Poor, 2015)	A research of the connection between social and psychological traits of artists and the crowdfunding process	A quantitative approach (questionnaire)	Results of the study suggest that more extroverted artists are more willing to use crowdfunding. Likewise, artists who already have a well-developed support network assess this model more positively.
(Kossecki & Świerczyńska-Kaczor, 2016)	An attempt to understand how crowdfunding works in Polish film industry, with a specific emphasis on the analysis of projects and attitudes of film professionals	Combined approach	Over 85% of participants positively assess the inclusion of sponsors in the production process. Identified obstacles involve an insufficient trust in crowdfunding and the lack of regulations. Participants who think that it is acceptable to share the profit with the donors comprise only one third of the sample.

DRAFT VERSION

(van den Hoogen, 2020)	A study of how various sets of values influence crowdfunding as a social phenomenon, on the example of the Netherlands	Combined approach	Personal connection with the artists and the broader societal impact of the project play an important part in attracting the audience. The majority of funds come from the already-existing, narrow circles of supporters. The market model most commonly works on the reward basis.
(Mollick & Nanda, 2016)	A comparison between the perception of theatre project financing by the audience and art experts	A quantitative method	Results suggest high correlation between the decisions made by the audience and by the experts. Some differences in terms of presentation have also been noted – materials based on multimedia and involving various rewards tend to better attract the audience, while the essence and content of the project tend to be better valued by the experts. However, there is no evidence that expert assessment positively influences project success – very often, it was the audience-selected projects that ended up being the most successful.
(Shneor et al., 2024)	An attempt to identify and provide deeper insight into artists' key decision-making factors regarding crowdfunding	A quantitative method (an online survey)	Artists already experienced in crowdfunding positively view the chance to use it again. If the environment encourages and supports the artist to undertake such action, the probably increases that the creative artist shall indeed take such a step.
(Boeuf et al., 2014)	An analysis of factors of success in a crowdfunding campaign involving theater projects, with a specific emphasis on the financial and symbolic rewards motivating donors	A quantitative analysis	Donors are not motivated by financial, but rather by symbolic rewards, such as their positive image and affiliation with the community. A need to broaden the networks has been identified, due to audience fatigue.

4. DISCUSSION

Crowdfunding is an innovative model which potentially offers several advantages, foremost the opportunity to diversify financing sources, which is crucially important for alternative projects, invisible to traditional donors (Fanea-Ivanovici & Baber, 2021; Galuszka & Bystrov, 2014; Loriguillo-López, 2017). The tool is essential for the visibility and promotion of as yet unestablished artists (Guimarães & Maehle, 2025). Further, crowdfunding provides an opportunity for direct artist-audience communication, fostering active audience participation (Foà, 2019; Galuszka & Bystrov, 2014; Villén Higueras et al., 2020). The equity-based and reward-based crowdfunding model additionally stimulates the stakeholders to participate in promotion and thus contribute to the project success, due to the fact that they either have a financial interest or symbolically feel affiliated with the project (Fanea-Ivanovici & Baber, 2021). Crowdfunding also provides a tool for preliminary market research, where the success of a campaign may signal more general audience interest. In this way, creative artists are in a position to test their ideas prior to more substantial investment (D'Amato & Cassella, 2021; Fanea-Ivanovici & Baber, 2021; Loriguillo-López, 2017). Research results also point out some limitations of the model, primarily the unpredictability of the project's outcomes (Galuszka & Bystrov, 2014) and impossibility to finance more ambitious or demanding endeavors (Galuszka & Bystrov, 2014; Loriguillo-López, 2017; Villén Higueras et al., 2020).

Dominant concepts in the analyzed literature revolve around two research methods: an analysis of the specifics of the practice of crowdfunding, with particular emphasis put on its organizational and financial aspects, and tests of creative artists' and donors' perceptions. For most artists, commercial success in this process goes hand in hand with numerous obstacles, which often limit the crowdfunding endeavors to a narrower, specialized audience (Dalla Chiesa & Dekker, 2021; Foà, 2019; Galuszka & Brzozowska, 2017). Artists claim that this model, which entails continued communication and constant involvement, often results in them being emotionally overwhelmed and exhausted. For this reason, they are not always ready to undertake new crowdfunding efforts (Dalla Chiesa, 2022; Fanea-Ivanovici, 2019; Kossecki & Świerczyńska-Kaczor, 2016). The lack of clear legislation and legal frameworks for donors often poses an additional obstacle, making their interest in investing in crowdfunded campaigns low (Fanea-Ivanovici, 2019; Kossecki & Świerczyńska-Kaczor, 2016). Studies comparing the audience and art experts have shown that the audience is ready to finance projects which experts would not prioritize, which does not mean that such projects are less successful (Mollick & Nanda, 2016). This opens up some room for defining crowdfunding as a mechanism of cultural diversification and correction of institutional mistakes.

DRAFT VERSION

5. CONCLUSION

The present research considers the phenomenon of crowdfunding in the context of art projects, with a specific emphasis on the predominant concepts and trends in the research literature. The goal of the study has been to provide a deeper insight into the most important theoretical and practical implications. The findings suggest that crowdfunding can significantly assist both early-career artists and alternative projects, in terms of fundraising and gathering support, which increases visibility and ensures independence from traditional institutional frameworks. In numerous studies, democratization is seen as an important element of crowdfunding (Foà, 2019; Guimarães & Maehle, 2025; Mollick & Nanda, 2016; van den Hoogen, 2020). It opens up room for social inclusion, pluralism in art, equal access to artistic materials, and participation in art productions. In spite of the increasing presence of the crowdfunding model, some studies (Fanea-Ivanovici, 2019; Kossecki & Świerczyńska-Kaczor, 2016) stress the lack of knowledge of its mechanisms and limited trust in its practical applicability.

The contribution of the paper can be found in its identification of dominant thematic wholes and key insights given in current literature, which can serve as the basis for future research. Since crowdfunding in the creative sector is a topic that has not yet been studied in Serbia, this paper can serve as additional inspiration for researchers to start working on the problem and enrich the relevant literature, but also for artists to potentially initiate the first models based on this practice. Our recommendation for future research is that empirical studies should be conducted in the Republic of Serbia, so as to better understand the attitudes to, familiarity with, and perception of the crowdfunding phenomenon among creative workers. Implications of such research can help support and develop crowdfunding platforms in Serbia, primarily oriented towards projects in the arts.

Limitations of the study pertain to the fact that we have considered only papers in the English language and that we have searched just one research database (WoS), which may have left out some relevant papers not detectable by our initial search criteria. Although the paper provides insight into the most important implications stemming from the relevant literature, the fact that we relied solely on secondary sources provides solid room for verifying our findings, especially in the local context.

DRAFT VERSION

ACKNOWLEDGMENT

This research received no external funding.

LITERATURE

Alexiou, K., Wiggins, J., & Preece, S. B. (2020). Crowdfunding Acts as a Funding Substitute and a Legitimating Signal for Nonprofit Performing Arts Organizations. *Nonprofit and Voluntary Sector Quarterly*, *49*(4), 827–848. https://doi.org/10.1177/0899764020908338

Boeuf, B., Darveau, J., & Legoux, R. (2014). Financing creativity: crowdfunding as a new approach for theatre projects. *International Journal of Arts Management*, *16*(3), 33–48.

Chalmeta, R., Cabezas-Hernando, D., & dos Santos Silva, J. (2025). Framework for crowdfunding sustainable development goals projects using blockchain. *Sustainable Development*, *33*(2), 2740–2758. https://doi.org/10.1002/sd.3260

Dalla Chiesa, C. (2022). The Artists' Critique on Crowdfunding and Online Gift-Giving. *The Journal of Arts Management, Law, and Society*, *52*(1), 20–36. https://doi.org/10.1080/10632921.2021.1997848

Dalla Chiesa, C., & Dekker, E. (2021). Crowdfunding artists: beyond match-making on platforms. *Socio-Economic Review*, *19*(4), 1265–1290. https://doi.org/10.1093/ser/mwab006

D'Amato, F., & Cassella, M. (2021). Cultural production and platform mediation: A case in music crowdfunding. *New Media & Society*, *23*(9), 2575–2592. https://doi.org/10.1177/1461444820932555

Davidson, R., & Poor, N. (2015). The barriers facing artists' use of crowdfunding platforms: Personality, emotional labor, and going to the well one too many times. *New Media & Society*, *17*(2), 289–307. https://doi.org/10.1177/1461444814558916

Fanea-Ivanovici, M. (2019). Filmmaking and Crowdfunding: A Right Match? *Sustainability*, *11*(3), 799. https://doi.org/10.3390/su11030799

Fanea-Ivanovici, M., & Baber, H. (2021). Crowdfunding model for financing movies and web series. *International Journal of Innovation Studies*, *5*(2), 99–105. https://doi.org/10.1016/j.ijis.2021.06.001

Foà, C. (2019). Crowdfunding cultural projects and networking the value creation. *Arts and the Market*, *9*(2), 235–254. https://doi.org/10.1108/AAM-05-2019-0017

Galuszka, P., & Brzozowska, B. (2016). Early career artists and the exchange of gifts on a crowdfunding platform. *Continuum*, *30*(6), 744–753. https://doi.org/10.1080/10304312.2016.1231790

Galuszka, P., & Brzozowska, B. (2017). Crowdfunding: Towards a redefinition of the artist's role – the case of MegaTotal. *International Journal of Cultural Studies*, *20*(1), 83–99. https://doi.org/10.1177/1367877915586304

Galuszka, P., & Bystrov, V. (2014). Crowdfunding: A Case Study of a New Model of Financing Music Production. *Journal of Internet Commerce*, *13*(3–4), 233–252. https://doi.org/10.1080/15332861.2014.961349

Guimarães, A. D., & Maehle, N. (2025). More than raising money: the role of crowdfunding in the development of cultural projects. *Journal of Cultural Economy*, 1–26. https://doi.org/10.1080/17530350.2025.2451262

Hossain, M., & Oparaocha, G. O. (2017). Crowdfunding: Motives, Definitions, Typology and Ethical Challenges. *Entrepreneurship Research Journal*, *7*(2). https://doi.org/10.1515/erj-2015-0045

Kossecki, P., & Świerczyńska-Kaczor, U. (2016). *The Role of Polish Crowdfunding Platforms in Film Productions – an Exploratory Study*. 1289–1296. https://doi.org/10.15439/2016F236

Kovpak, V., & Lebid, N. (2022). CREATIVE INDUSTRIES AS A MECHANISM OF CREATIVE ECONOMY AND STRATEGIC COMMUNICATIONS. *Baltic Journal of Economic Studies*, *8*(4), 102–109. https://doi.org/10.30525/2256-0742/2022-8-4-102-109

Loriguillo-López, A. (2017). Crowdfunding Japanese Commercial Animation: Collective Financing Experiences in Anime. *International Journal on Media Management*, *19*(2), 182–195. https://doi.org/10.1080/14241277.2017.1298112

Marlett, D. (2015). Crowdfunding Art, Science and Technology: A Quick Survey of the Burgeoning New Landscape. *Leonardo*, *48*(1), 104–105. https://doi.org/10.1162/LEON_a_00813

Mollick, E., & Nanda, R. (2016). Wisdom or Madness? Comparing Crowds with Expert Evaluation in Funding the Arts. *Management Science*, *62*(6), 1533–1553. https://doi.org/10.1287/mnsc.2015.2207

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71

Rijanto, A. (2022). Creative industries project financing through crowdfunding: the roles of fund target & backers. *Creative Industries Journal*, *15*(1), 79–96. https://doi.org/10.1080/17510694.2021.1899489

Shneor, R., Maehle, N., Munim, Z. H., & Rykkja, A. (2024). What drives crowdfunding intentions of artist-entrepreneurs? – Evidence from Norway. *The International Journal of Entrepreneurship and Innovation*. https://doi.org/10.1177/14657503241282939

van den Hoogen, Q. L. (2020). Values in crowdfunding in the Netherlands. *International Journal of Cultural Policy*, *26*(1), 109–127. https://doi.org/10.1080/10286632.2018.1433666

Villén Higueras, S. J., Ma, X., & Ruiz-del-Olmo, F. J. (2020). Crowdfunding as a Catalyst for Contemporary Chinese Animation. *Animation*, *15*(2), 130–144. https://doi.org/10.1177/1746847720933792

Wang, C. (2016). The Promise of Kickstarter: Extents to Which Social Networks Enable Alternate Avenues of Economic Viability for Independent Musicians Through Crowdfunding. *Social Media + Society*, *2*(3). https://doi.org/10.1177/2056305116662394

DRAFT VERSION

MATHEMATICAL PROGRAMMING

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

THE CAPACITATED SINGLE ALLOCATION HUB MAXIMAL COVERING PROBLEM WITH HUB INSTALLATION COSTS

ZORICA STANIMIROVIĆ[1], OLIVERA STANČIĆ[2]

[1] University of Belgrade, Faculty of Mathematics, Belgrade, zorica.stanimirovic@matf.bg.ac.rs, 0000-0001-5658-4111
[2] University of Kragujevac, Faculty of Economics, Kragujevac, olivera.stancic@kg.ac.rs, 0000-0001-7219-9029

Abstract: *This study considers a variant of the hub maximal covering problem that assumes single allocation scheme, hub capacities and hub installation costs. The objective of the problem is to find optimal locations for opening hubs and optimal allocations of each non-hub node to one of the installed hubs with sufficient capacity such that the sum of uncovered demands for all origin-destination pairs and the hub installation costs is minimized. In the hub location literature, this problem is denoted as the capacitated single allocation hub maximal covering problem. Two four-index integer linear mathematical formulations of the considered problem are presented, together with reformulations into a two-index and a three-index mixed integer linear program. Each of the four presented mathematical formulations is used within the framework of an exact solver to find solutions for the set of modified Australian Post hub instances. The results obtained are compared in respect to the number of optimal solutions and the quality of the upper bounds obtained by the exact solver, as well as the computational times required when using the considered mathematical formulations.*

Keywords: *Hub maximal covering problem, Hub capacities, Hub installation costs, Mathematical programming, Integer linear programming.*

DRAFT VERSION

1. INTRODUCTION

Hub covering problems were introduced in (Campbell, 1994) as extensions of the classical covering location problem. In the classical hub covering problem, it is necessary to choose locations for opening hubs and to allocate non-hub nodes to opened hubs, so that each origin-destination (O-D) pair of nodes is covered (Campbell, 1994). An O-D pair is considered "covered" if the distance of the path established via hub nodes is not greater than the specified maximum service distance. The distance between an O-D pair can also be observed as the transportation cost or travel time along the path between an O-D pair via hub nodes.

From the early work by (Campbell, 1994), many variants of hub covering problems have been proposed in the literature. Generally, hub covering problems developed in two main directions: hub set covering and hub maximal covering problems. The goal of the hub set covering problem is to locate hub nodes such that all demands are covered while the cost of opening hub facilities is minimized. The hub maximal covering problem maximizes the demand covered with a fixed number of hubs to be located. An overview of hub covering problems, mathematical formulations, and solution methods can be found in (Alumur et al., 2021; Contreras and O'Kelly, 2019; Karimi and Bashiri, 2011; Wagner, 2008).

Capacitated variants of hub covering problems impose limits on the volume of flow that can be collected at a hub, routed through it, or transported via inter-hub connections. To the best of our knowledge, such capacity constraints have been considered in only a few studies within the existing literature. In (Mohammadi et al., 2011), the capacitated single allocation hub covering problem was considered and a new imperialist competitive algorithm was proposed as a solution method. The capacitated multi-modal multi-product p-hub covering problem under a queuing approach was studied in (Sedehzadeh et al., 2015) and solved by a multi-objective parallel simulated annealing algorithm. In (Hossein Karimi et al., 2016), the authors proposed the multi-modal single allocation capacitated p-hub covering problem over fully interconnected hub networks and used a tabu-search solution approach. The capacitated hub maximal covering problem with transfer flow and hub establishment costs was presented in (Alizadeh et al., 2016) and solved using two multi-objective evolutionary algorithms. The capacitated multiple allocation hub covering flow problem was first introduced in (Sener and Feyzioğlu, 2021).

In this study, we consider a variant of the capacitated single allocation hub maximal covering problem (CSAHMCP), which involves capacity restrictions on incoming flow collected at hubs and fixed costs of opening

hubs. This study extends our previous research on multiple hub covering problem variants (Janković, Mišković, et al., 2017; Janković and Stanimirović, 2017; Todosijević et al., 2024). The considered CSAHMCP arises from the design of various service networks, such as postal delivery, express delivery, logistics, telecommunications, public transportation, etc. In all these networks, it is assumed that a hub node can receive and process a limited amount of flow, and that there are certain expenses when setting up hubs. The number of hubs to be established is not fixed in advance. Each non-hub node sends and receives the flow via exactly one, previously opened hub (single allocation scheme). When choosing hub locations and allocations of non-hub nodes to hubs, the travel time of transporting flow from an origin to destination node through the resulting network must not exceed a certain constant (coverage radius). Only in that case, the transported flow for an O-D pair is considered to be covered. The objective of the CSAHMCP is to minimize the company's total losses, which include the losses (expressed in monetary units) for uncovered flows and the costs of setting up hubs.

The CSAHMCP under consideration is formulated as four equivalent integer or mixed integer linear programs, which are further used within the framework of the commercial CPLEX solver to solve modified AP instances from the literature. The results obtained by the four CSAHMCP formulations are compared in the sense of the quality of obtained solutions and running times required by CPLEX to return either an optimal or a feasible solution. The presented results are analyzed in terms of the number of established hubs in solutions obtained for AP instances with the same dimension but different fixed costs and capacity types.

The remainder of the paper is organized as follows. In Section 2, the four equivalent mathematical formulations of the CSAHMCP are presented. Section 3 presents the results of computational experiments conducted on modified hub instances from the literature, together with the analysis of the presented results. A summary of the study and suggestions for future research directions are provided in Section 4.

2. MATHEMATICAL FORMULATIONS

Let $G(N,E)$ be a complete undirected graph, where N is a set with n nodes, $E = \{(i,j) : i,j \in N\}$ is set of edges, while $H \subseteq N$ is the set of potential hubs. The travel time from origin node i to destination node j via hubs $k,m \in H$ is calculated as $t_{ij}^{km} = \chi t_{ik} + \alpha t_{km} + \delta t_{mj}$, where t_{ij} is the travel time along the edge $(i,j) \in E$. The parameters χ, δ and α multiply travel times from an origin to a hub, from a hub to a destination, and for the inter-hub connections, respectively, and $\alpha < \chi, \delta$ holds. If the travel time t_{ij}^{km} is not greater than a specific value β, the O–D pair $i-j$ is considered to be covered by hubs k and m. The flow of demand T_{ij} from node $i \in N$ to node $j \in N$ is given, and it is not necessary that $T_{ij} = T_{ji}$ holds. For each potential hub $k \in H$, there is capacity limit $g_k \geq 0$ on the flow collected at hub and fixed costs $F_k \geq 0$ for hub installation. Each non-hub node is assigned to exactly one hub (single allocation scheme), while hubs communicate directly with each other. Due to the triangle inequality imposed on t_{ij}, the flow for each O-D pair is routed via at most two hubs. The goal of CSAHMCP is to choose locations for opening hubs from the set $H \subseteq N$ and to allocate each non-hub node to exactly one opened hub, so as to minimize the sum of the total uncovered flow and the costs of hub installations.

The four-index formulation for CSAHMCP uses a binary parameter a_{ij}^{km}, $i,j \in N$, $k,m \in H$ that takes the value of 1 if $t_{ij}^{km} \leq \beta$, and 0 otherwise. Binary variable x_{ik} takes the value of 1 if node $i \in N$ is allocated to hub $k \in H$, and 0 otherwise. Note that $x_{kk} = 1$ implies that the node k is hub and is assigned to itself. Binary variable y_{ijkm} is equal to 1 if there is traffic from origin node $i \in N$ to destination node $j \in N$ that is routed via hubs $k \in H$ and $m \in H$, respectively, and 0 otherwise. By using this notation, the CSAHMCP may be formulated as a four-index Integer Linear Program (ILP):

$$\min \sum_{i\in N}\sum_{j\in N}\sum_{k\in H}\sum_{m\in H}(1-a_{ij}^{km})T_{ij}y_{ijkm} + \sum_{k\in N}F_k x_{kk} \tag{1}$$

subject to

$$x_{ik} \leq x_{kk} \quad \forall i \in N \quad \forall k \in H, \tag{2}$$

$$\sum_{k\in H} x_{ik} = 1 \quad \forall i \in N, \tag{3}$$

$$\sum_{i\in N}\sum_{j\in N} T_{ij}x_{ik} \leq g_k x_{kk} \quad \forall k \in H, \tag{4}$$

$$\sum_{k\in H}\sum_{m\in H} y_{ijkm} = 1 \quad \forall i,j \in N, \tag{5}$$

$$\sum_{m\in H} y_{ijkm} \leq x_{ik} \quad \forall i,j \in N \quad \forall k \in H, \tag{6}$$

DRAFT VERSION

$$\sum_{k\in H} y_{ijkm} \le x_{jm} \quad \forall i,j \in N \quad \forall m \in H, \tag{7}$$

$$x_{ik} \in \{0,1\} \quad \forall i \in N \quad \forall k \in H, \tag{8}$$

$$y_{ijkm} \in \{0,1\} \quad \forall i,j \in N \quad \forall k,m \in H. \tag{9}$$

The objective (1) is to minimize the sum of uncovered demand for all O-D pairs and the cost of hub installations. Constraints (2) and (3) ensure that each non-hub node is assigned to exactly one established hub. The amount of flow collected in each hub is limited by constraints (4). Constraint (5) provides that the flow for each O-D pair $i-j$ is completely routed through a pair of hubs k and m. In this route, the origin i must be assigned to hub k, while destination j must be allocated to hub m, which is ensured by constraints (6) and (7). Finally, constraints (8) and (9) define binary variables x_{ik} and y_{ijkm}, respectively.

The presented four-index formulation for CSAHMCP can be modified as follows. For every $i,j \in H$ and $k,m \in H$, the sets $N_{ij} = \{(k,m) \in H^2 : t_{ij}^{km} \le \beta\}$, $N'_{ijk} = \{m \in H : t_{ij}^{km} \le \beta\}$, and $N''_{ijm} = \{k \in H : t_{ij}^{km} \le \beta\}$ are defined. Moreover, additional binary variables w_{ij}, $i,j \in N$ are introduced, where each w_{ij} equals 1 if the origin-destination pair $i-j$ is uncovered, and 0 otherwise. The resulting modified four-index ILP formulation for CSAHMCP is as follows:

$$\min \sum_{i\in N}\sum_{j\in N} T_{ij} w_{ij} + \sum_{k\in N} F_k x_{kk} \tag{10}$$

subject to (2), (3), (4), (8), and

$$\sum_{(k,m)\in N_{ij}} y_{ijkm} + w_{ij} = 1, \quad \forall i,j \in N, \tag{11}$$

$$\sum_{m\in N'_{ijk}} y_{ijkm} \le x_{ik} \quad \forall i,j \in N \quad \forall k \in H, \tag{12}$$

$$\sum_{k\in N''_{ijm}} y_{ijkm} \le x_{jm} \quad \forall i,j \in N \quad \forall m \in H, \tag{13}$$

$$y_{ijkm} \in \{0,1\} \quad \forall i,j \in N \quad \forall (k,m) \in H^2. \tag{14}$$

$$w_{ij} \in \{0,1\} \quad \forall i,j \in N. \tag{15}$$

In an O-D pair $i-j$ is not covered, there is no traffic from origin i to destination node j, which is indicated by constraints (11). Constraints (12) and (13) have the same role as constraints (6) and (7). Variables y_{ijkm} and w_{ij} are of binary type, which is denoted by constraints (14) and (15).

In order to present a two-index formulation of CSAHMCP, we use parameter $\lambda_{ij} = \max\{a_{ij}^{km} : k,m \in H\}$ from (Peker and Kara, 2015). For each covered O-D pair $i-j$, we introduce a non-negative variable Z_{ij} representing the fraction of flow routed from origin i to destination j. By using this newly introduced notation, and the notation from the four-index formulation (1)–(9), the CSAHMCP may be modeled as the following two-index Mixed Integer Linear Program (MILP):

$$\min \sum_{i\in N}\sum_{j\in N} (1-Z_{ij}) T_{ij} + \sum_{k\in N} F_k x_{kk} \tag{16}$$

subject to (2), (3), (4), (8), and

$$Z_{ij} \le \sum_{k\in H} a_{ij}^{km} x_{ik} + \lambda_{ij}(1-x_{jm}), \quad \forall i,j \in N, m \in H \tag{17}$$

$$Z_{ij} \ge 0, \quad \forall i,j \in N. \tag{18}$$

The objective function (16) carries the same interpretation as (1). Note that the newly introduced variables Z_{ij} are now incorporated in the first sum of the objective function in (16). The new constraints (17) provides that the fraction of flow Z_{ij} for a covered O-D pair $i-j$ corresponds to appropriate values of allocation variables x_{ik} and x_{jm}. Coverage variables Z_{ij} are non-negative, which is indicated by constraints (18).

DRAFT VERSION

Finally, we present a three-index CSAHMCP formulation that involves the set $G_{ij}^k = \{m \in H : a_{ij}^{km} = 1\}$ for each O–D pair $i-j$ and each hub k. This set contains all hubs $m \in H$ such that the established path $i \to k \to m \to j$ is covered. We also introduce a set of binary variables w_{ijk}, where $i, j \in N, k \in H$. Each variable w_{ijk} corresponds to the set G_{ij}^k and takes the value of 1 if the origin node i is allocated to the hub k and the destination node j is assigned to a hub from the set G_{ij}^k. By using this new set of variables w_{ijk}, together with the set of variables x_{ik}, x_{kk}, Z_{ij}, and the previously defined parameters, the CSAHMCP is formulated as the following three-index MILP: (16) subject to (2)-(4), (8), (18), and

$$w_{ijk} \leq \sum_{m \in G_{ij}^k} x_{jm}, \quad \forall i, j \in N, k \in H \tag{19}$$

$$w_{ijk} \leq x_{ik}, \quad \forall i, j \in N, k \in H \tag{20}$$

$$\sum_{k \in H} w_{ijk} \leq 1, \quad \forall i, j \in N \tag{21}$$

$$Z_{ij} \leq \sum_{k \in H} w_{ijk}, \quad \forall i, j \in N \tag{22}$$

$$w_{ijk} \in \{0, 1\}, \quad \forall i \in N, k \in H \tag{23}$$

Constraints (19)–(20) ensure that the variable w_{ijk} takes the value of 1 only if the node i is assigned to hub k and the node j is assigned to some hub m from the set G_{ij}^k. Constraints (21) guarantee that for each O–D pair $i-j$, there is at most one covering path established, while constraints (22) define Z_{ij} as the fraction of the flow between a covered O-D pair $i-j$. Finally, constraints (23) indicate the binary type of variables w_{ijk}.

3. COMPUTATIONAL RESULTS

For our numerical experiments, we employed Australian Post (AP) benchmark instances, which incorporate both node-specific fixed costs and capacity constraints, as introduced by (Ernst and Krishnamoorthy, 1998). Each problem size n was expanded into four scenario types by combining tight (T) and loose (L) values for both fixed costs and capacities, resulting in the following configurations: nTT, nTL, nLT, and nLL. The flow cost parameters were fixed at $\delta = \chi = 1$ and $\alpha = 0.75$. Notably, the AP flow matrix is asymmetric, with $T_{ij} \neq T_{ji}$ and $T_{ii} \neq 0$. n all considered AP instances for CSAHMCP, the values of the binary coverage parameter a_{ij}^{km} were adopted as proposed by (Peker and Kara, 2015):

$$a_{ij}^{km} = \begin{cases} 1, & \text{if } \ t_{ij}^{km} \leq \beta, \ \forall i, j \in N, \ \forall k, m \in H \\ 0, & \text{otherwise.} \end{cases}$$

In case that the two parts of the objective function (the total uncovered demand and the total cost of setting up hubs) are not well balanced, it may lead to a preference for minimizing one part of the objective function. To resolve this issue, we scaled the input data in the objective function, which results in scaling of the two components of the objective function. More precisely, let $T_{total} = \sum_{i,j \in N} T_{ij}$ be the total flow between all O-D pairs, and $Cost_{total} = \sum_{k \in N} F_k$ the total costs of establishing hubs on each potential hub location. The input data T_{ij} and F_k are scaled as follows: $T_{ij} := 100 \cdot T_{ij}/T_{total}$, for each $i, j \in N$ and $F_k := 100 \cdot F_k/Cost_{total}$, for each $k \in N$. As the original amounts of flow T_{ij} between O–D pairs $i-j$ are now scaled, the given capacities of each potential hub must also be scaled: $g_k := 100 \cdot g_k/T_{total}$, for every $k \in N$. The results of preliminary computational experiments have shown that the scaling of T_{ij}, F_k, and g_k leads to a balance of the impact of the two parts in the objective function. The values of the coverage parameter β are set to the corresponding solution values for the uncapacitated multiple allocation p-hub center problem (UMApHCP) from (Brimberg et al., 2017) with $p = 5$ hubs (for AP instances with $n \leq 50$) and $p = 10$ hubs (for AP instances with $n = 100$).

Table 1 presents the results obtained by the CPLEX solver using each of the four CSAHMCP formulations on the modified AP instances with up to 100 nodes. Note that all computational experiments are performed on an Intel(R) Core(TM) i7-7820HQ CPU @ 2.90GHz with OS Windows 11 Pro. The first two columns of Table 1 contain instance's name and the value of parameter β.For each mathematical formulation analyzed, the remaining columns of Table 1 provide detailed information. These include the number of hubs that were opened in the solutions found by the CPLEX solver, which is displayed under the column labeled *hubs*. The column titled *value* reports either the optimal solutions identified by CPLEX or, if the solver did not manage to find an optimal solution within the one-hour time limit, the corresponding upper bounds are shown and marked with

DRAFT VERSION

Table 1: Computational results on AP instances with up to 100 nodes

		four-index form.				modified four-index form.				three-index form.				two-index form.			
Inst.	β	hubs	value	gap	time	hubs	value	gap	time	hubs	value	gap	time	hubs	value	gap	time
10TT	30371.32	4	**39.43**	0.00	1.09	4	**39.43**	0.00	0.34	4	**39.43**	0.00	0.20	4	**39.43**	0.00	0.20
10TL	30371.32	2	**46.07**	0.00	1.14	2	**46.07**	0.00	0.37	2	**46.07**	0.00	0.16	2	**46.07**	0.00	0.20
10LL	30371.32	2	**38.16**	0.00	2.23	2	**38.16**	0.00	0.63	2	**38.16**	0.00	0.17	2	**38.16**	0.00	0.17
10LT	30371.32	2	**34.44**	0.00	2.08	2	**34.44**	0.00	0.42	2	**34.44**	0.00	0.17	2	**34.44**	0.00	0.18
20TT	37868.15	3	**15.29**	0.00	148.48	3	**15.29**	0.00	52.44	3	**15.29**	0.00	3.75	3	**15.29**	0.00	2.90
20TL	37868.15	3	**18.81**	0.00	224.72	3	**18.81**	0.00	45.67	3	**18.81**	0.00	5.99	3	**18.81**	0.00	7.20
20LL	37868.15	1	**13.03**	0.00	240.22	1	**13.03**	0.00	38.25	1	**13.03**	0.00	2.13	1	**13.03**	0.00	1.76
20LT	37868.15	1	**11.51**	0.00	242.66	1	**11.51**	0.00	42.07	1	**11.51**	0.00	3.74	1	**11.51**	0.00	1.86
25TT	45552.50	3	**9.07**	0.00	682.15	3	**9.07**	0.00	317.82	3	**9.07**	0.00	47.01	3	**9.07**	0.00	8.66
25TL	45552.50	3	**12.06**	0.00	908.21	3	**12.06**	0.00	271.39	3	**12.06**	0.00	41.79	3	**12.06**	0.00	13.86
25LL	45552.50	1	**9.54**	0.00	1488.10	1	**9.54**	0.00	589.96	1	**9.54**	0.00	44.20	1	**9.54**	0.00	26.24
25LT	45552.50	2	**7.77**	0.00	1615.01	2	**7.77**	0.00	587.43	2	**7.77**	0.00	58.68	2	**7.77**	0.00	15.48
40TT	49741.20	14	60.56*	>100.00	3600.00	-	-	-	-	5	**6.10**	0.00	1058.93	5	**6.10**	0.00	748.03
40TL	49741.20	14	61.73*	>100.00	3600.00	-	-	-	-	3	8.82*	0.00	3600.00	3	**8.82**	0.00	855.22
40LL	49741.20	3	56.21*	>100.00	3600.00	3	56.21*	>100.00	3600.00	2	**5.89**	0.00	929.79	2	**5.89**	0.00	242.27
40LT	49741.20	3	56.80*	>100.00	3600.00	3	56.80*	>100.00	3600.00	2	**3.98**	0.00	669.75	2	**3.98**	0.00	192.40
50TT	50707.87	-	-	-	-	5	93.37*	>100.00	3600.00	-	-	-	-	4	**5.44**	0.00	1472.04
50TL	50707.87	-	-	-	-	5	83.10*	>100.00	3600.00	-	-	-	-	3	**7.46**	0.00	2235.23
50LL	50707.87	-	-	-	-	-	-	-	-	-	-	-	-	2	**4.70**	0.00	1211.02
50LT	50707.87	-	-	-	-	-	-	-	-	-	-	-	-	3	**3.42**	0.00	1430.38
100TT	51860.03	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
100TL	51860.03	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
100LL	51860.03	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
100LT	51860.03	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

an asterisk (*). The best or optimal values are highlighted in bold. In instances where CPLEX was unable to produce any feasible solution due to memory or time restrictions, a dash ("-") is used. Furthermore, the table presents the optimality gap expressed as a percentage, indicating the difference between the best known solution and the bound returned by CPLEX. Finally, the CPU time in seconds is reported, indicating how long CPLEX took to find an optimal solution or to confirm the lower and upper bounds.

Table 1 demonstrates that the CPLEX solver, utilizing the two-index formulation, efficiently found optimal solutions for all AP instances with up to 50 nodes, with an average computation time of 423.27 seconds. However, it failed to identify any feasible solutions for instances with 100 nodes within the one-hour time limit. When employing the three-index formulation, CPLEX was unable to solve one instance (40TL) among the sixteen tested with up to 40 nodes. Furthermore, for instances with 50 and 100 nodes, the solver exhausted memory resources before generating feasible solutions. The four-index formulations exhibited the poorest performance, with no optimal solutions for instances exceeding 25 nodes. While the first four-index approach yielded feasible solutions for 40-node instances, it failed on larger problems. The modified four-index formulation produced feasible solutions for select instances (40LL, 40LT, 50TT, 50TL), yet was unsuccessful for others, including 40TT and 40TL. Analysis of the reported CPU times confirms that the two-index formulation outperforms the three-index and both four-index formulations in computational efficiency.

A comparative analysis of AP instance groups that share the same number of nodes but differ in the combination of fixed costs and capacity settings (*n*TT, *n*TL, *n*LL, and *n*LT) reveals consistent trends in the obtained solutions. Instances characterized by both tight fixed costs and tight capacities (TT) tend to produce solutions with the highest number of opened hubs. This outcome is expected, as nodes with high flow volumes, otherwise suitable for hub locations, are discouraged by restrictive costs and limited capacity, leading the model to distribute hub openings more broadly. In TL-type instances, where capacities are relaxed but fixed costs remain tight, the number of hubs generally decreases. However, this simplification results in a higher objective function value compared to TT cases. A comparable effect is observed in LT-type instances, where fixed costs are reduced but tight capacities are maintained. Overall, instances of type LL, characterized by both relaxed fixed costs and capacities, tend to yield the fewest established hubs, but also exhibit the highest values of the objective function. This tendency intensifies as the number of nodes increases.

4. CONCLUSION

This study presents four mathematical formulations for the capacitated single allocation maximal covering location problem (CSAHMCP), including two four-index integer linear programming formulations, one two-index, and one three-index mixed-integer linear programming (MILP) formulation. The problem is classified as NP-hard, given that it generalizes its uncapacitated counterpart, which has already been proven to be NP-hard.

DRAFT VERSION

The experimental analysis was carried out on a set of Australian Post (AP) benchmark instances containing up to 100 nodes, each defined by specific capacity limits and installation costs for potential hub locations. Among the formulations examined, the two-index MILP formulation contains the smallest number of variables and constraints. As expected, the computational results confirmed that the two-index formulation outperforms the other three in both solution quality and average CPLEX runtime. In the case of 100-node instances, all four CSAHMCP formulations failed to produce a feasible solution within the imposed computational limits using the CPLEX solver. Future work will focus on considering different allocation strategies and incorporating the concept of partial coverage into the hub covering formulation.

ACKNOWLEDGMENT

The research conducted in this paper was partially supported by the Ministry of Science, Technological Development, and Innovation of the Republic of Serbia through the institutional grants for the year 2025 to University of Belgrade, Faculty of Mathematics, and University of Kragujevac, Faculty of Economics.

LITERATURE

Alizadeh, Y., Tavakkoli-Moghaddam, R., & Ebrahimnejad, S. (2016). A new multi-objective model for a capacitated hub covering problem solving by two multi-objective evolutionary algorithms. *International Journal of Mathematics in Operational Research*, *9*(1), 99–124. https://doi.org/10.1504/IJMOR.2016.077559

Alumur, S. A., Campbell, J. F., Contreras, I., Kara, B. Y., Marianov, V., & O'Kelly, M. E. (2021). Perspectives on modeling hub location problems. *European Journal of Operational Research*, *291*(1), 1–17. https://doi.org/10.1016/j.ejor.2020.09.039

Brimberg, J., Mladenović, N., Todosijević, R., & Urošević, D. (2017). A basic variable neighborhood search heuristic for the uncapacitated multiple allocation p-hub center problem. *Optimization Letters*, *11*, 313–327. https://doi.org/10.1007/s11590-015-0973-5

Campbell, J. F. (1994). Integer programming formulations of discrete hub location problems. *European Journal of Operational Research*, *72*(2), 387–405. https://doi.org/10.1016/0377-2217(94)90318-2

Contreras, I., & O'Kelly, M. (2019). Location science. In G. Laporte, S. Nickel, & F. Saldanha da Gama (Eds.). Springer.

Ernst, A., & Krishnamoorthy, M. (1998). Exact and heuristic algorithms for the uncapacitated multiple allocation p-hub median problem. *European Journal of Operational Research*, *104*(1), 100–112.

Janković, O., Mišković, S., Stanimirović, Z., & Todosijević, R. (2017). Novel formulations and vns-based heuristics for single and multiple allocation p-hub maximal covering problems. *Annals of Operations Research*, *259*(1), 191–216. https://doi.org/10.1007/s10479-017-2508-1

Janković, O., & Stanimirović, Z. (2017). A general variable neighborhood search for solving the uncapacitated r-allocation p-hub maximal covering problem. *Electronic Notes in Discrete Mathematics*, *58*, 23–30. https://doi.org/10.1016/j.endm.2017.03.004

Karimi, H., & Bashiri, M. (2011). Hub covering location problems with different coverage types. *Scientia iranica*, *18*(6), 1571–1578. https://doi.org/10.1016/j.scient.2011.09.018

Karimi, H. [Hossein], Bashiri, M., & Nickel, S. (2016). Capacitated single allocation p-hub covering problem in multi-modal network using tabu search. *International Journal of Engineering-Transactions C: Aspects*, *29*(6), 797–808. https://doi.org/10.5829/idosi.ije.2016.29.06c.09

Mohammadi, M., Tavakkoli-Moghaddam, R., & Rostami, H. (2011). A multi-objective imperialist competitive algorithm for a capacitated hub covering location problem. *International Journal of Industrial Engineering Computations*, *2*(3), 671–688. https://www.ije.ir/article_72131.html

Peker, M., & Kara, B. Y. (2015). The p-hub maximal covering problem and extensions for gradual decay functions. *Omega*, *54*, 158–172. https://doi.org/10.1016/j.omega.2015.01.009

Sedehzadeh, S., Tavakkoli, M. R., & Jolai, F. (2015). A new multi-mode and multi-product hub covering problem: A priority m/m/cqueueapproach. *International Journal of Industrial Mathematics*, *7*(2), 1–10. https://sanad.iau.ir/fa/Article/846119

Sener, N., & Feyzioğlu, O. (2021). Capacitated multiple allocation hub covering flow problem. *MANAS Journal of Engineering*, *9*(1), 72–84. https://doi.org/10.51354/mjen.809844

Todosijević, R., Stančić, O., Stanimirović, Z., & Mišković, S. (2024). General variable neighborhood search for the capacitated single allocation hub maximal covering problem. *TOP*, 1–42. https://doi.org/10.1007/s11750-024-00685-9

Wagner, B. (2008). Model formulations for hub covering problems. *Journal of the Operational Research Society*, *59*(7), 932–938. https://doi.org/10.1057/palgrave.jors.2602424

DRAFT VERSION

MINING, GEOLOGY AND ENERGY

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Review article

ANALYSIS OF COAL EXPLOITATION TRENDS

DRAGAN MILOŠEVIĆ[1], BRANKA JOVANOVIĆ[2]

[1] Rudarski institut d.o.o. Beograd, dragan.milosevic@ribeograd.ac.rs, 0000-0003-3714-1950
[2] Rudarski institut d.o.o. Beograd, povrsinska@ribeograd.ac.rs, 0000-0002-4095-6790

Abstract: *The mining industry plays a key role in the planetary economy, from exploration to the exploitation of mineral resources, ensuring the production of electricity necessary for modern life. In addition, it employs a large number of people. Since the beginning of industrialization in the 19th century, coal has had a significant impact on the world's energy supply. Despite widespread calls for a shift to clean energy sources, to this day, coal represents the second largest primary energy source in the world. Issues of using renewable energy sources represent essential not only existential but also political views on the way and quality of life, as well as the fate of future generations.*

Keywords: *Coal, electricity, power plant.*

1. MINING INDUSTRY

Many mining companies produce a diverse range of minerals to minimize exposure to volatile and unpredictable market prices. They are known as diversified mining companies. Although many of the world's largest mining companies are based in the United Kingdom, mainland Europe or Australia, in recent years many Chinese mining companies have become increasingly prominent and have increasing revenues. BHP (Australia), Rio Tinto (UK/Australia), China Shenhua Energy (China), Glencore (Switzerland), Vale (Brazil), Anglo American (UK), Teck Resources (Canada), Saudi Arabian Mining Company (Ma'aden) (Saudi Arabia) are among the largest mining companies, which normally operate in several countries.

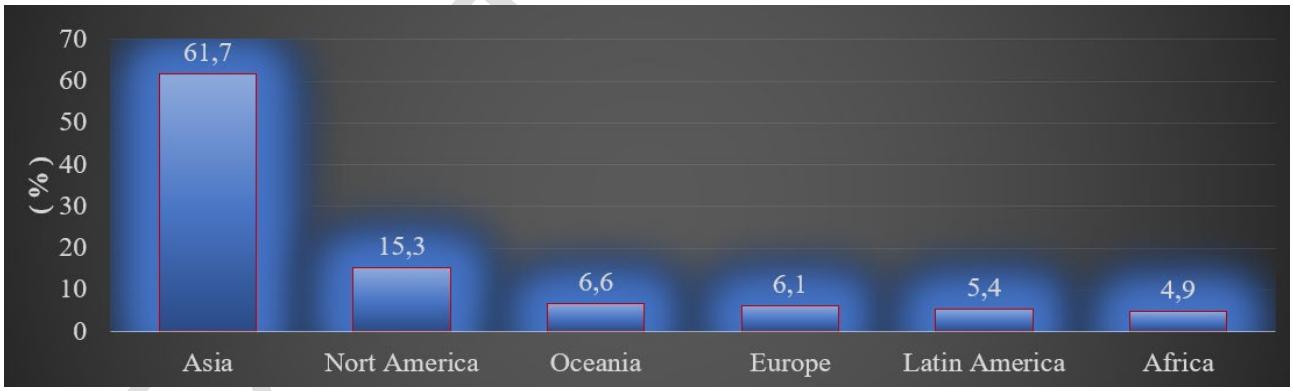


Figure 1. Global share of mine production volume in 2022 by region (M.Garside et al., 2024a)

2. COAL EMISSIONS AROUND THE WORLD

The burning of fossil fuels is the primary cause of rising global temperatures, and no fuel has contributed more to this than coal. Coal is the primary source of energy used as a means of producing electricity and heat. It was one of the first fuels used in the energy sector with a major contribution to the industrial revolution. Today, coal is still the world's largest source of electricity production. Despite being the dirtiest and most polluting fossil fuel, coal remains the largest source of electricity generation worldwide. In fact, coal-fired electricity was responsible for 21% of total global greenhouse gas emissions in 2021, making it the largest climate polluter. As the world aims to achieve climate neutrality by 2050, the transition from coal to clean energy sources will be a key area of focus for all governments in the future (Ian Tiseo, 2024).

DRAFT VERSION

3. GLOBAL COAL MINING INDUSTRY

Renewable energy production has been steadily increasing over the past decades, with solar and wind energy showing the greatest growth. However, the contribution of renewables to the world's electricity mix remains small compared to fossil fuels.

In the last three decades, coal has been the main source of electricity generation in the world. The total production of energy from coal in 2023 was almost 10,500 terawatt-hours. Overall, coal, natural gas and other fossil fuels accounted for about 60% of total electricity production. Fossil fuels remain the largest source of electricity generation worldwide. In 2023, coal accounted for roughly 35.5% of the global electricity mix, followed by natural gas at 23%.

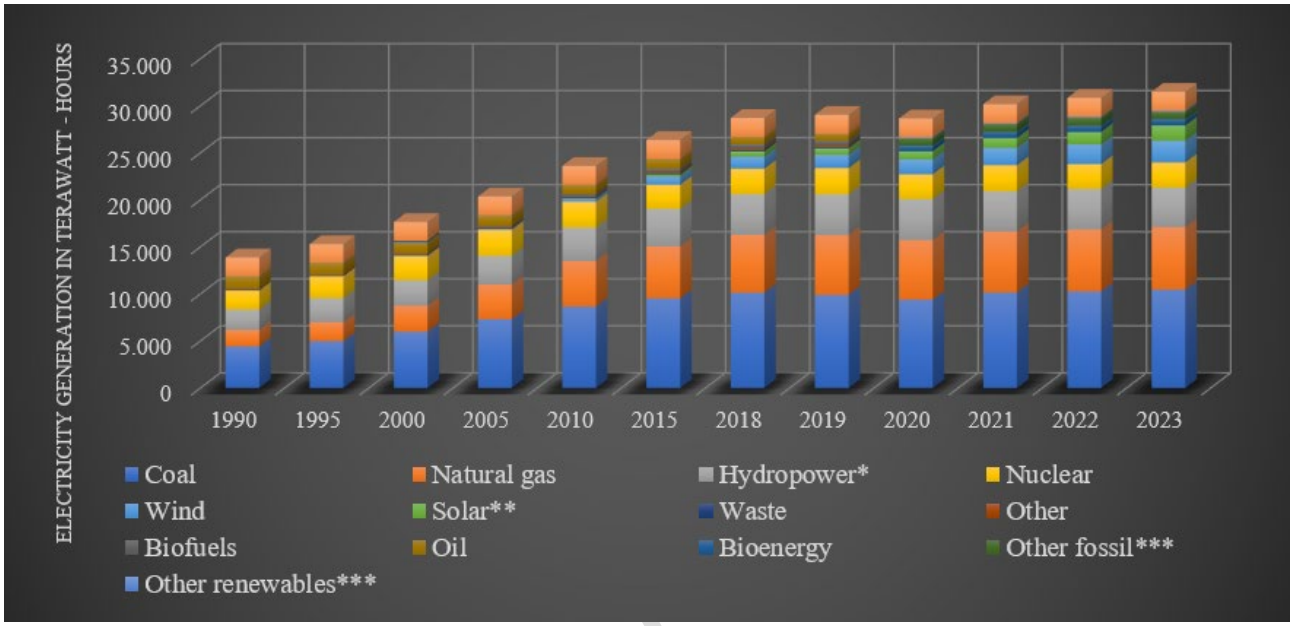


Figure 2. Electricity production. of energy by energy consumption in the world (1990 - 2023, in terawatt-hours) (Statista Research Departmen, 2024a)

To meet demand, the global mining industry has increased annual production since 2000 and has maintained it at a nearly constant level. However, it should also be noted that coal has by far the largest reserves of all non-renewable energy sources. In 2022, hard coal accounted for 45.8% of the world's proven non-renewable energy reserves. Meanwhile, conventional oil accounted for about 18% of reserves.

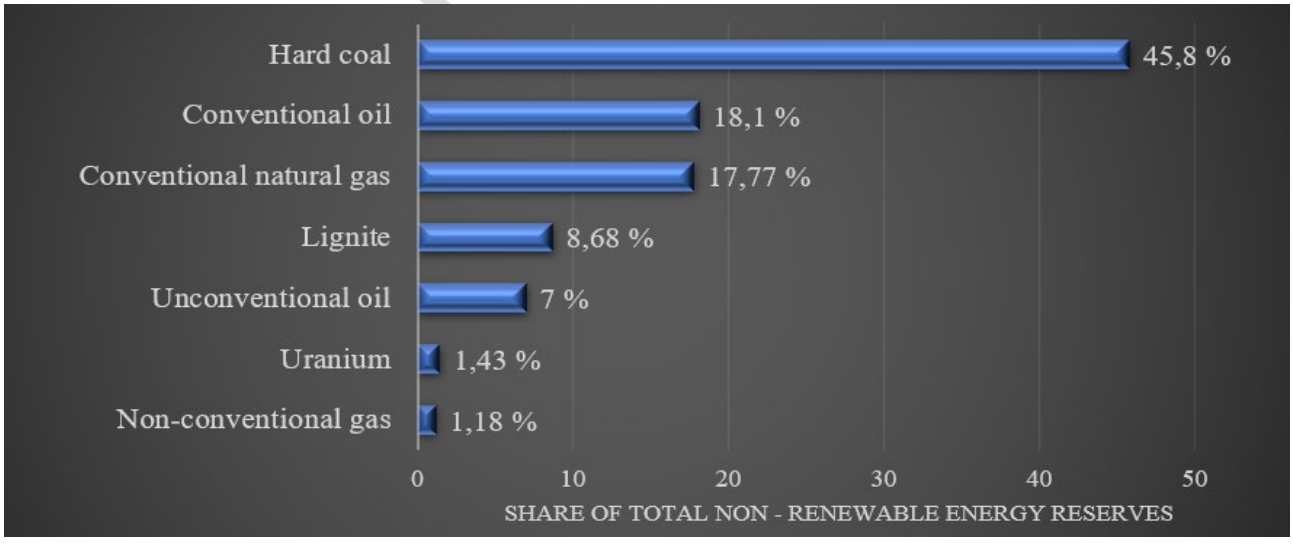


Figure 3. Overview of global non-renewable energy reserves in 2022 according to sources (Statista Research Department, 2024b)

At current electricity production rates, coal and lignite reserves are sufficient for more than a hundred years. Unlike oil and gas, coal is widely distributed throughout the world. The largest concentration of reserves is represented in the USA, Russia and China.

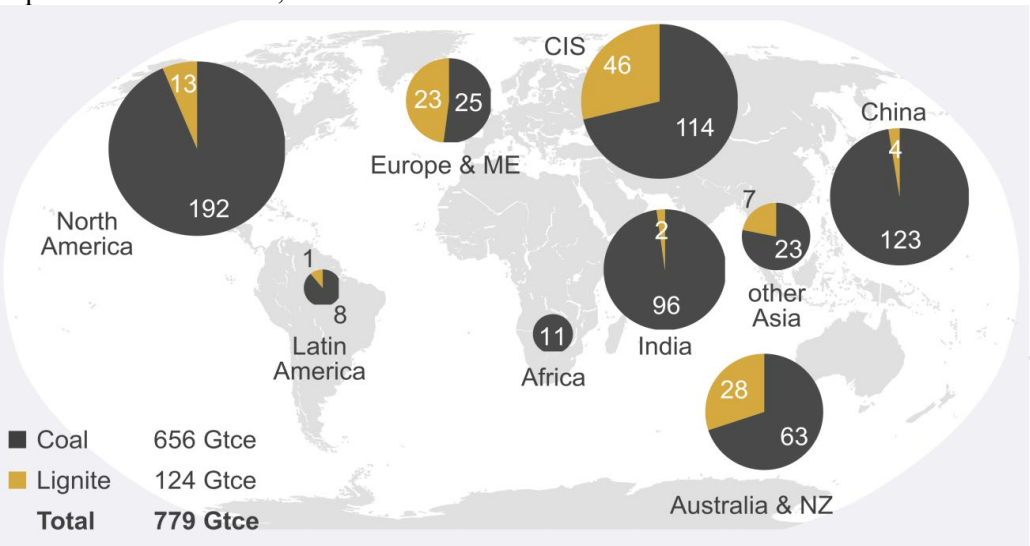


Figure 4. Global coal and lignite reserves, December 31, 2021 (Eurocoal)

Primary energy production in the Republic of Serbia is based mainly on the use of coal. It can also be stated that there has been a noticeable decline in the use of coal in recent years.

Table 1. Electricity production in the Republic of Serbia from 2017 to 2021, by energy sources (Statistical Office of the Republic of Serbia)

Energy	2017		2018		2019		2020		2021	
	TJ	%	TJ	%	TJ	%	TJ	%	TJ	%
Coal	302.127	68.23	276.720	65.37	285.772	66.22	295.064	65.11	263.034	61.58
Hydroelectric power	35.107	7.93	41.015	9.69	36.714	8.51	32.524	7.16	40.624	9.51
Solar photovoltaic energy	47	0.01	47	0.01	49	0.01	48	0.01	49	0.01
Wind energy	175	0.04	542	0.13	3.234	0.75	3.512	0.77	3.904	0.91
Crude oil and natural gas liquids	40.914	9.24	40.310	9.52	38.990	9.03	38.030	8.38	36.967	8.65
Firewood, wood chips and the rest	45.395	10.25	46.931	11.08	49.151	11.39	67.334	14.83	66.938	15.67
Natural gas	18.117	4.09	16.653	3.93	16.247	3.76	15.133	3.33	13.414	3.14
Geothermal energy	222	0.05	219	0.05	220	0.05	212	0.05	63	0.01
Biogas	723	0.16	939	0.22	1.190	0.28	1.624	0.36	2.221	0.52

Over the past half century, the world's electricity consumption has grown continuously, reaching approximately 27,000 terawatt-hours by 2023. Between 1980 and 2023, electricity consumption more than tripled, while the global population reached 8 billion people. The growth of industrialization and access to electricity further increased the demand for electricity. For example, China's GDP has grown by an incredible 15 times since 2000, making it the world's second largest economy, behind the United States. To fuel its industrial development, China needs more energy than any other country. As a result, it has become the largest consumer of electricity in the world and requires more than 8,000 terawatt-hours.

DRAFT VERSION

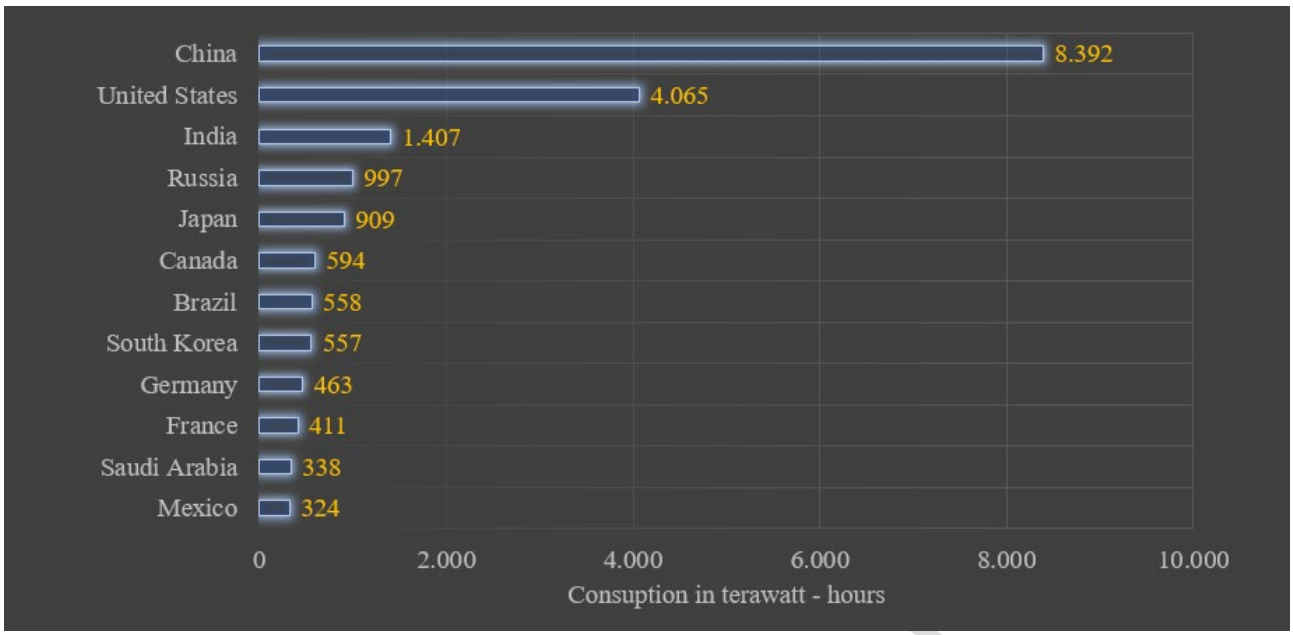


Figure 5. Electricity consumption in the world in 2023 by leading country (in terawatt-hours) (Statista Research Departmen, 2024c)

However, the per capita electricity consumption of the world's largest population is almost ten times less than that of Iceland, even though the energy used in this country came almost entirely from clean sources. Iceland is by far the largest consumer of electricity per capita in the world. On average, it consumes 53.9 megawatt-hours per person in 2023. This is the result of a combination of factors, such as cheap electricity generation, increased demand for heating and the presence of energy-intensive industries in the country.

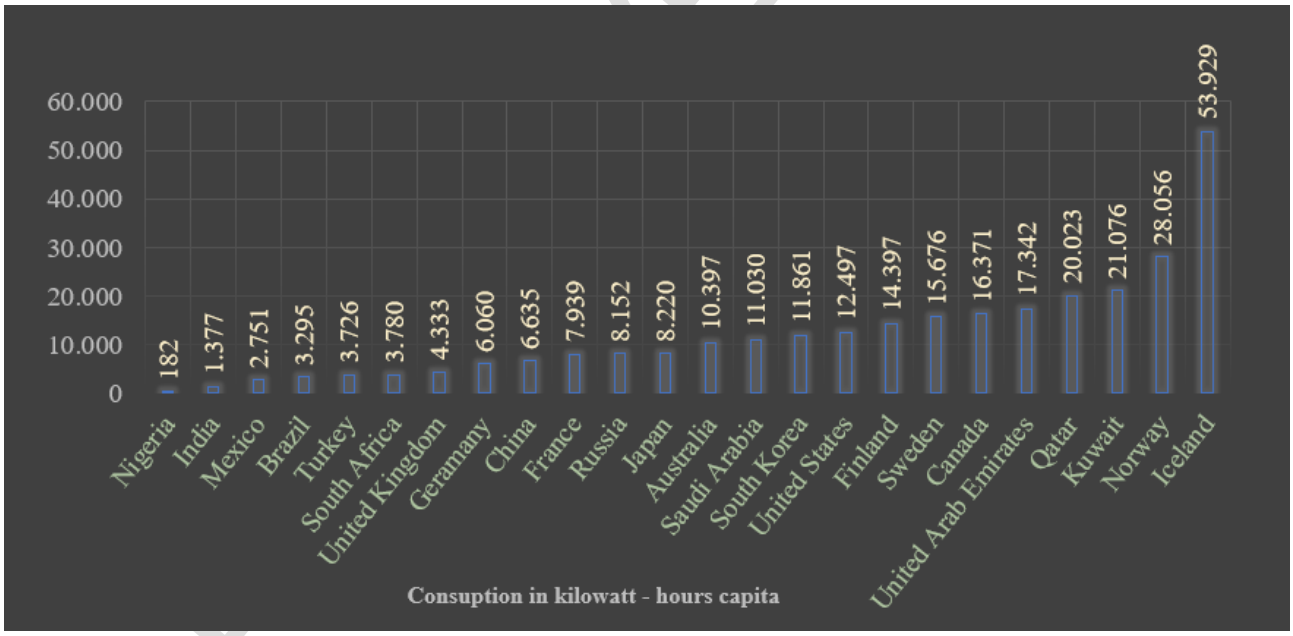


Figure 6. Electricity consumption per capita in 2023 (Statista Research Departmen, 2024d)

China has the largest number of coal-fired power plants in the world. As of July 2024, there were 1,161 operational coal-fired power plants on the Chinese mainland, four times the number of such plants in second-place India, for example. Over 50% of the total global electricity production from coal is realized in China.

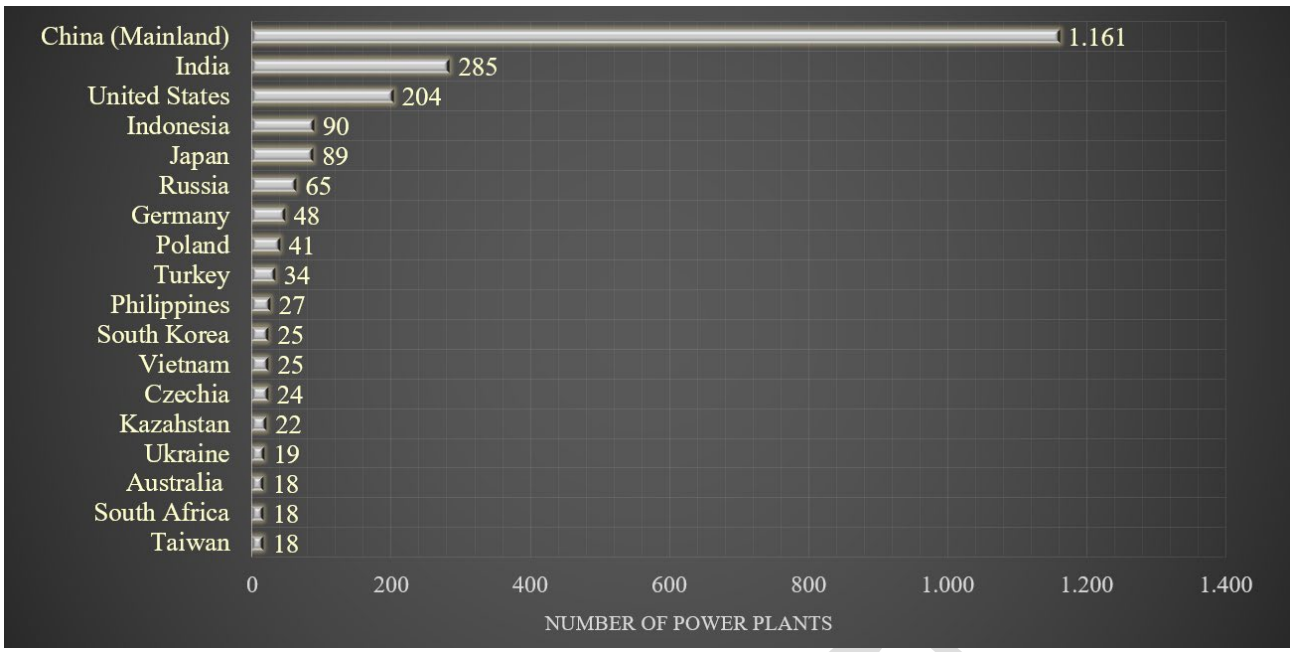


Figure 7. Countries with the largest number of coal-fired power plants in the world as of July 2024 (Statista Research Department, 2024e)

The capacities of the Republic of Serbia for the production of electricity managed by EPS have a total capacity of 7,855 MW. Eight thermal power plants with 25 coal-fired units, with a total capacity of 5,171 MW, produce about 70 percent of the electricity in the Republic of Serbia, while about 30 percent is obtained from 16 hydroelectric plants. The average coal production at mines in EPS ranges from 37 to 40 million tons of coal (EPS). In 2022, Indonesia exported about 471 million metric tons of coal, making it the world's largest coal exporting country. Australia is in second place, with coal exports of 359 million metric tons that year.

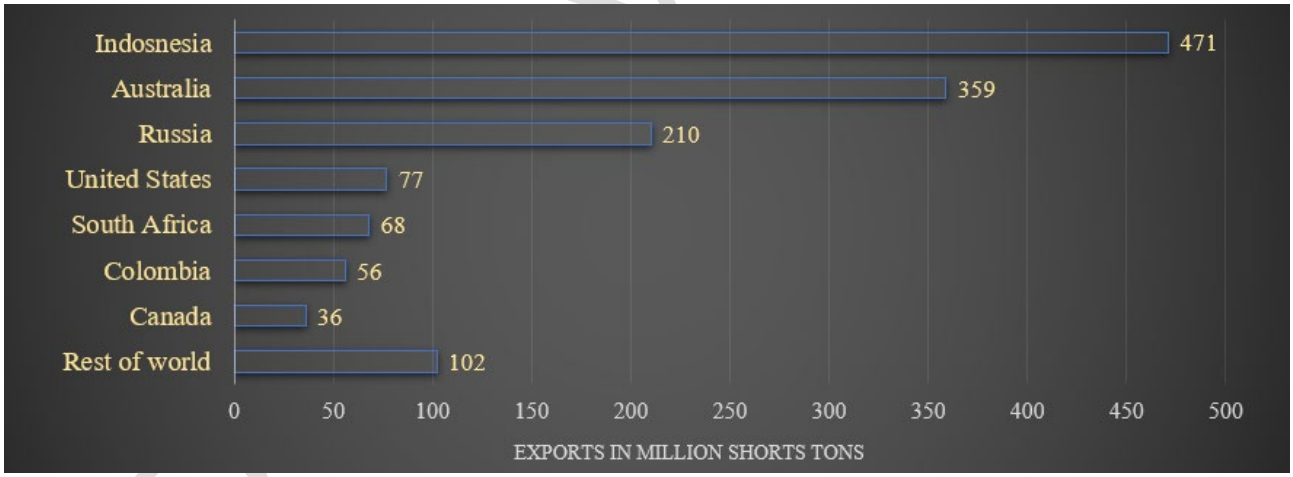


Figure 8. Leading coal exporting countries in 2022 (M. Garside, 2024b)

4. CONCLUSION

The structure of primary energy production is largely determined by the natural resources of a territory, but also by its strategic political decisions that affect the development of specific energy sources, such as renewable energy sources. Following the increased efforts of governments to fight against global warming, the share of renewable sources in the production of electricity has experienced a more pronounced growth in recent years compared to the use of fossil fuels. Projections show that renewable energy sources will surpass fossil fuels as the main source of energy by 2040. The key message of the work is the fact that the use of renewable energy sources is far below the expected result to which the countries of the energy community have committed themselves. The lack of results is not only a consequence of the lack of investor interest, but the presence of various influencing factors. Economic, political, social are only some of the most significant, which stand in the way of the construction of most projects of renewable energy sources. In this sense, it is

necessary to define achievable goals and guidelines for the improvement of renewable energy policy at the most intensive level.

LITERATURE

Garside, M. (2024a, December). *Global share of mine production volume 2022, by region*. Statista. https://www.statista.com/statistics/1474291/global-mine-production-volume-distribution-by-region/

Tiseo, I. (2024a, July). *Coal emissions worldwide – statistics & facts*. Statista. https://www.statista.com/topics/10428/coal-emissions-worldwide/#topicOverview

Statista Research Department. (2024b, October). *Global electricity generation 1990–2023, by source*. Statista. https://www.statista.com/statistics/273273/world-electricity-generation-by-energy-source/

Statista Research Department. (2024, May). *Global non-renewable energy reserves breakdown 2022, by source*. Statista. https://www.statista.com/statistics/263230/proportion-of-non-renewable-energy-reserves/

Euracoal. (n.d.). *Coal use worldwide*. https://euracoal.eu/coal/coal-use-worldwide/

Statistical Office of the Republic of Serbia. (2020–2024). *Statistical calendar of the Republic of Serbia*. Belgrade: SORS.

Statista Research Department. (2024c, October). *Global electricity consumption 2023, by country*. Statista. https://www.statista.com/statistics/267081/electricity-consumption-in-selected-countries-worldwide/

Statista Research Department. (2024d, August). *Per capita electricity consumption worldwide 2023, by selected country*. Statista. https://www.statista.com/statistics/383633/worldwide-consumption-of-electricity-by-country/

Statista Research Department. (2024e, September). *Global operational coal-fired power stations by country 2024*. Statista. https://www.statista.com/statistics/859266/number-of-coal-power-plants-by-country/

Elektroprivreda Srbije (EPS). (n.d.). *Official website*. https://www.eps.rs

Garside, M. (2024b, August). *Global leading coal exporting countries 2022*. Statista. https://www.statista.com/statistics/270952/global-hard-coal-exports-2009/

DRAFT VERSION

MULTICRITERIA ANALYSIS AND OPTIMIZATION

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

OPTIMIZING THE TRAJECTORY FOR IMPROVING QUALITY EDUCATION USING SLACK-BASED DATA ENVELOPMENT ANALYSIS

ANA KRSTIĆ SREJOVIĆ[1], PREDRAG MIMOVIĆ[2]

[1] University of Kragujevac, Faculty of Economics, Kragujevac, anakrstic@kg.ac.rs, 0000-0001-5649-3199
[2] University of Kragujevac, Faculty of Economics, Kragujevac, mimovicp@kg.ac.rs, 0000-0003-0323-8033

Abstract: *The paper analyzes the performance of sustainable development in thirty-five European countries in the year 2021. SDG 4 – Quality Education – was taken as a benchmark for evaluation. Performance of the countries was evaluated using the Slack-Based Measure Data Envelopment Analysis (DEA) model, which has a higher discriminating power than the standard DEA model. Then, stratification of the model was carried out in order to identify several levels of efficiency and reduce the analysis to a partial set of inefficient units. A multi-stage benchmarking method was applied to generate stepwise improvement paths and reveal peer references for inefficient units. The results showed that from the point of view of the effectiveness of national policies for the implementation of SDG 4, the reference countries in 2021 were Belgium, Ireland, Greece, Italy, Luxembourg, Finland, Sweden, Iceland, and Switzerland. Through the process of elimination of efficient states, the optimal benchmarking trajectory was identified using the example of Germany.*

Keywords: *sustainability development goals, quality of education, performance, efficiency, benchmarking, data envelopment analysis, multi-stage evaluation, policy improvement trajectory.*

1. INTRODUCTION

According to the United Nations agenda approved by the General Assembly in 2017, the quality of education - SDG 4 has its quantitative and qualitative dimensions. The quantitative dimension is reflected in the structure of the SDG 4 framework, which consists of two levels with 12 global indicators and 31 thematic indicators, in order to complement the depth and breadth of the global set of indicators (UIS, 2023). The qualitative dimension includes concepts such as inclusion, equality, lifelong learning, international cooperation, etc. Together, these indicators form a multidimensional structure that poses significant methodological challenges for evaluation and cross-country comparison. Hence, the monitoring and follow-up of progress in achieving the goals of SDG 4 and, as a result, the benchmarking of countries, are of key importance for progress in the domain of the national education sector. Achieving the goals set within SDG-4 is of great importance for the achievement of other sustainable development goals (Ferguson & Roofe, 2020; Agbedahin, 2019; Leicht et al., 2018).

In this context, the aim of this paper, on the one hand, is to evaluate the achieved SDG 4 performance through a relative comparison of the observed countries, so that the indicators of the SDG 4 goal, as defined by the European Commission, will be considered as output variables of the Data Envelopment Analysis (DEA) model whose mutual influences are also taken into account. On the other hand, although methodologically derived, the actual primary goal is to reach the simplest possible set of the worst states through the multi-stage elimination of efficient states, and in that way, through a reversible process, identify their benchmarking trajectories, as well as the optimal one. Such an approach enables policy makers to identify stepwise improvement scenarios, prioritize achievable benchmarks, and allocate resources more effectively. Chronologically, therefore, the procedure is top - down, that is, it starts from the complete set, and the decision-making process is bottom-up. In this paper, Slack-Based Measure Data Envelopment Analysis was used to calculate the results, which, compared to the standard DEA model (Charnes et al., 1978), has a higher discriminating power and therefore provides a better and more harmonized efficiency calculation. By combining SBM-DEA with a multi-stage stratification mechanism, the methodology also contributes to dynamic benchmarking, offering not just a snapshot but a trajectory for policy improvement.

DRAFT VERSION

2. RESEARCH METHODOLOGY

Assume that there are *n* decision units, denoted as DMUs. Let *DMUj* (*j*=1,2,...,*n*) consume m_1 variable input $x_{i_{1j}}\,(i = 1,2,..,m_1)$ and m_2 fixed input $x_{i_{2j}}\,(i = 1,2,..,m_2)$, to produce k_1 variable output $y_{r_{1j}}\,(r_1 = 1,2,..,k_1)$ and k_2 fixed output $y_{r_{2j}}\,(r_2 = 1,2,..,k_2)$. Then, according to (Tone, 2001), we can evaluate the performance of the observed DMU_0 through the following Slack Based Measure (SBM) input-oriented DEA model (1):

$$h_0 = \frac{1-\frac{1}{m}\sum_{i=1}^{m}\frac{s_i^-}{x_{i0}}}{1+\frac{1}{s}\sum_{r=1}^{s}\frac{s_r^+}{y_{r0}}} \tag{1}$$

$$x_{i0} = \sum_{j=1}^{n}\lambda_j \times x_{ij} + s_i^{(-)}$$

$$y_{r0} = \sum_{j=1}^{n}\lambda_j \times y_{rj} - s_r^{(+)}$$

$$s^- \geq 0, s^+ \geq 0$$

The optimal solution of model (1) is denoted as $(\theta^*, \lambda_j^*, s_i^{*(-)}, s_r^{*(+)})$. In model (2) the performance of DMU_0 can be improved by $x'_{i0} = x_{i0} - s_i^{*(-)}$ and $y'_{r0} = y_{r0} + s_r^{*(+)}$, wherein h^* the efficiency of the DMU_0, λ_j^* contribution of the reference efficient decision-making unit in achieving the goal of the analyzed decision-making unit, and $(x'_{i0}, \forall i)$ i $(y'_{r0}, \forall r)$ optimally improved input and output variables, respectively. Unlike radial models, the SBM model measures efficiency by taking into account disproportionate surpluses and deficits of inputs and outputs, which can lead to a more accurate ranking of DMUs, especially when there is an excess of inputs or a deficit of outputs.

3. STRUCTURING OF THE MODEL

The Data Envelopment Analysis has often been used, in different variants, in the measurement and analysis of sustainable development performance (Zhang et al., 2008; Tian et al., 2019; Li et al., 2019; Wang et al., 2018; Long, 2021; Nocera Alvares Junior et al., 2024). The appropriate SBM DEA output oriented model, with two inputs and four outputs are defined as follows:

1) A group consisting of thirty-five European countries for which there is available data is observed (Appendix, table A);

2) The available data were collected from the *Eurostat* and *World Bank* databases**,** for the year 2021. The selection of this specific year ensures temporal consistency and comparability across countries, given that 2021 was a post-COVID recovery period with significant implications for education systems.

3) Two input variables were chosen: government expenditure on education (% of GDP) and teacher-to-student ratio, reflecting the availability of financial and human resources in the educational system.

4) Four output indicators were selected, representing key components of SDG 4 in the EU context, as defined by the European Commission.

The output-oriented SBM DEA model is an adequate and methodologically justified choice in the analysis of the efficiency of educational systems in situations where we want to assess the potential for improving educational results, the data contain disproportionate surpluses and deficits, or we have complex indicators with different directions of desirability (such as SDG4-10). SDG4-10 (early leavers from education and training), measures the percentage of individuals (typically aged 18–24) who have left education and training with at most lower secondary education and are not in further education or training. A high rate suggests challenges in retaining young people in the education system, which can negatively affect employment, economic growth, and social inclusion. So, a lower value for early leavers from education and training is preferable. This is in contrast to standard efficiency measurement in DEA, where a higher output value typically signifies better performance. In order to account for this inverse relationship and maintain consistency with the DEA framework, a multiplicative inversion of the SDG4-10 output value was applied, according to the transformation: O = 1 / O. This ensures the model rewards improvements in this indicator appropriately.

The second output refers to the SDG 4.20 indicator – tertiary educational attainment by sex. This indicator measures the share of the population aged 25–34 who have successfully completed tertiary

DRAFT VERSION

education. It reflects the effectiveness of the education system in enabling young people to attain higher qualifications, which is essential for fostering innovation, competitiveness, and inclusive economic growth.

The third output refers to early childhood education participation by sex, indicating the percentage of children aged three to the start of compulsory primary education who are enrolled in structured early learning programmes (ISCED level 0). These programmes aim to support children's early development, and the EU has set a 2030 target of at least 96% participation, emphasizing the foundational role of early education in lifelong learning.

The fourth output refers to the indicator adult participation in learning in the past four weeks by sex, which tracks the percentage of individuals aged 25 to 64 who reported taking part in formal or non-formal education or training within the four weeks prior to the survey. Adult learning plays a vital role throughout life, supporting professional development, adaptation to new technologies, and reintegration into the labour market, while also contributing to personal well-being, active citizenship, and social inclusion.

The stratification of decision-making units is based on the procedure formulated by Seiford & Zhu, 2003. Let U^l be the set of observed decision-making units, and E^l, the subset of effective decision-making units, so that $E^l \in U^L$. Then the procedure has the following steps:

Step 1: Identification of the set of all decision-making units U^l;

Step 2: Evaluation of the efficiency of decision units in order to obtain the first level of efficient decision units E^l;

Step 3: Exclusion of efficient units from the set of DMUs at the first level; $U^{l+1} = U^l - E^l$. If $U^{l+1} = \emptyset$, the procedure stops;

Step 4: Evaluate the new subset of ineficient DMUs, U^{l+1}, in order to obtain a new subset of efficient DMUs, E^{l+1};

Step 5*:* Let $l = l + 1$. Go to step 2.

Stopping rule: $E^{l+1} = \emptyset$, the algorithm stops.

This iterative elimination of efficient units enables a deeper decomposition of inefficiencies and facilitates the construction of benchmarking trajectories for less efficient countries. By applying this procedure, it is possible to determine a multi-stage benchmarking trajectory for inefficient DMUs. In this way, a country can gradually improve its efficiency by benchmarking against a single efficient DMU or a convex combination of efficient DMUs, which overcomes the problem of unrealistic or unfeasible improvement paths (Lim et al., 2011). Such gradual transitions are particularly important in the policy domain, where resource reallocation and systemic reforms are typically incremental rather than immediate.

4. RESULTS AND DISCUSSION

The process begins with a top-down efficiency evaluation, including all observed decision-making units (DMUs), and gradually identifies optimal benchmarking trajectories starting from the least efficient units. In the first iteration, which consists of a set of 35 observed countries, with two inputs and four output variables, the efficient DMUs are Belgium, Ireland, Greece, Italy, Luxembourg, Finland, Sweden, Iceland, and Switzerland. In the second iteration, the set of decision units was reduced by excluding the efficient countries from the first iteration. Efficient DMUs now include Denmark, Spain, Cyprus, Lithuania, Hungary, Malta, Netherlands, Austria, Romania, Slovenia, Norway, Montenegro, and Serbia. In the third iteration, after excluding efficient DMUs from the second iteration, the reference countries are Estonia, Croatia, Latvia, Portugal, Slovakia, North Macedonia, and Türkiye. Finally, in the fourth iteration, after eliminating all previously identified efficient DMUs, the only remaining inefficient countries are the Czech Republic and Germany. This stage marks the point where the discriminatory power of the DEA model is significantly weakened due to the reduced number of DMUs and a high proportion of efficient units. Therefore, it is the threshold at which further stratification becomes analytically redundant. Germany's efficiency score in the first iteration was only 40.24% of the best practice, revealing considerable inefficiency in its 2021 performance on SDG 4 implementation. Luxembourg was identified as the key reference country contributing most significantly (λ = 1.419) to the target efficiency for Germany. The data suggest major output-side inefficiencies for Germany, particularly in its teacher-to-student ratio (13.2), which is notably less favorable compared to, for example, Malta (8.7). For example, for sub-goal SDG4-10, while Germany's original value in 2021 was 0.1, the target value to match Luxembourg's efficiency was 0.295, representing an increase of 195.23% (Table 1). For sub-goal SDG4-60, the necessary increase was 229.95%. It is evident that immediate alignment with Luxembourg's level of efficiency is highly unrealistic for Germany in the short term, which underscores the rationale for a step-by-step benchmarking trajectory. Hence, a more feasible solution lies in multi-level benchmarking, allowing for progressive adjustment of output values. In

DRAFT VERSION

the backward analysis, beginning with the fourth iteration, the reference countries for Germany are France (32.1%) and Poland (77.3%). After adjusting Germany's output variables according to these contributions, the target value of SDG4-60 is recalculated:

$$Germany\ (O_{4_{SDG\ 4-60}}{}^{*}) = \lambda_{France}{}^{*} \times O_{4_{SDG\ 4-60}} + \lambda_{Poland}{}^{*} \times O_{4_{SDG\ 4-60}} =$$

$$= 0.321 \times 11 + 0.774 \times 5.4 = 7.6$$

In a similar way, according to the contributions of the reference countries, in this iteration of France and Poland, the target values of other variables are calculated (Table 5). Now Germany, with the necessary correction of the value of the output variables SDG 4, joins the observed set of decision-making units at level III, so the model is solved anew. The reference country is Portugal, and Germany's efficiency is 58.93% of the reference. After correcting the values of the variables in terms of contribution of France and Poland, Germany becomes efficient in terms of achieving SDG4 progress, after which, with such optimized values of the model variables, it moves to the level corresponding to iteration IV, in the forward analysis. The model is re-solved where the efficiency of Germany is 68.35% of the best practice, and the reference countries in terms of contribution are Latvia (0.106) and Portugal (1.077). The procedure is repeated by correcting the values of the input and output variables for Germany, but now in terms of contributions at this level. These step-by-step corrections form a trajectory that is both methodologically robust and practically attainable, allowing a progressive approach to the optimization of educational performance. The corrected values of the variables for Germany, at all levels in the backward analysis, are shown in Table 6. The optimized values of the variables at this level are now applied to the set of countries in iteration III. The model is solved again, and the German efficiency is now 93.21% of the best practice, and the reference countries are Malta (0.016) and Spain (0.595). Once more, values are corrected based on the contributions of these reference units, and Germany is reinserted into the original full set of DMUs corresponding to the first iteration. By solving the model thus formed, Germany's efficiency is 77.85% of the best practice - a notable improvement compared to the initial 40.24% in the original forward analysis. The reference country is once again Luxembourg, as in the forward analysis, but now the contribution is 0.738 - significantly less than the previous 1.42, demonstrating a less intensive and more realistic path toward alignment with best practice. It is clear that harmonization in order to achieve the targeted efficiency is much more favorable and simpler for Germany when a multi-level benchmarking approach is applied. By correcting the values of the output variables according to the contributions of the reference countries, other inefficient countries also reach the threshold of best practice. Bearing the above in mind, starting from the chronological SBM DEA procedure, the optimal benchmarking trajectory would be to compare and reconcile through the following sequence of reference comparisons:

- France and Poland (Iteration IV)
- Latvia and Portugal (Iteration III)
- Malta and Spain (Iteration II)
- Luxembourg (Iteration I)

with appropriate adjustments to the values of the input and output variables in terms of their contribution. In the final phase, the target values of the output variables, which are necessary for Germany to reach the best practice of the reference countries in 2021, are:

- Government Expenditure to Education = 2.73%
- Teachers/Students Ratio = 6.86%
- SDG 4-10* = 0.154%
- SDG 4-20* = 45.02%
- SDG 4-31* = 65.61%
- SDG 4-60.5* = 13.21%

Table 2 displays the original and optimal values of input and output variables for Germany, optimal in terms of efficiency at all stages of the backward analysis. First, Germany is compared with France and Poland as benchmarks. Based on their contributions shown, it is necessary to adjust the model's variables. At

DRAFT VERSION

the next level, Latvia and Portugal become the reference countries, with Germany aligning its effectiveness to match their performance. This is followed by benchmarking with Malta and Spain, and finally, Luxembourg. The entire procedure is multi-stage, somewhat slower from a methodological and decision-making standpoint, but offers a significantly higher feasibility of achieving the desired outcomes. It avoids abrupt reforms or unrealistic targets by providing phased, evidence-based improvement plans tailored to national capacities. In the same way, the optimal benchmarking trajectory for the other inefficient country identified in the forward analysis, the Czech Republic, can be determined. By applying the same reverse DEA model structure and adjusting for its unique variable values, a similarly viable path for SDG4 improvement can be proposed.

Table 1: Original and target values of output variables of the model for Germany (forward analysis)

	Expenditure	T/S ratio	SDG 4 -10	SDG 4 -20	SDG 4 - 31	SDG 4 - 60
Original values	5.5	13.2	0.1	37.1	93.1	7.7
Target values	5.25	13.2	0.295	86.58	126.18	25.41
Diff.(%)	-4.52	-	195.23	133.37	35.53	229.95

Source: authors

Table 2: Original and target values of output variables of the model for Germany (backward analysis)

	Expenditure	T/S ratio	SDG 4 -10	SDG 4 -20	SDG 4 - 31	SDG 4 - 60
Original values	5.5	13.2	0.1	37.1	93.1	7.7
II level	5.37	13.22	0.115	47.53	102.07	7.6
III level	3.00	6.86	0.146	30.73	58.50	8.79
IV level	2.73	6.86	0.154	45.02	65.61	13.21
Diff. Target/Original in %	-50.37%	-48.03%	54%	21.35%	-29.53%	71.55%

Source: authors

5. CONCLUSION

The results of the research indicated a generally high or reference level of efficiency among most European Union member states, with a few notable exceptions. The chosen SBM DEA model is an adequate and methodologically justified choice in the analysis of the efficiency of educational systems, because:

- the goal of the research was the maximization of educational results;
- Policymakers want to know how much educational outcomes can be improved with existing resources - which is exactly what the output-oriented model measures;
- Heterogeneity among countries: The SBM model can accurately capture disproportional ties among indicators and enable a more realistic ranking of countries.

Within the SBM DEA model, which included a set of 35 observed countries, two input variables, and four output indicators, the following countries were identified as efficient Decision Making Units (DMUs) in 2021 in terms of progress toward SDG 4: Belgium, Ireland, Greece, Italy, Luxembourg, Finland, Sweden, Iceland, and Switzerland. The most unexpected result was the low efficiency score of Germany, and to a lesser extent, the United Kingdom, given their economic and institutional strength. However, when considering the actual variable values within the model-such as the student-teacher ratio and specific SDG 4 output metrics - this result becomes less surprising and reveals potential structural weaknesses in their educational systems relative to others. The applied multi-stage benchmarking approach, while conceptually straightforward, is based on the mathematically complex, non-parametric DEA method. By successively eliminating efficient units - first from the complete set and then from reduced subsets, this approach facilitates the identification of a minimal set of inefficient DMUs. This is crucial in terms of maintaining the discriminatory power of the model, especially considering the balance between the number of DMUs and the number of variables. Unlike traditional forward benchmarking that expands the DMU set, this study introduced a backward analysis, beginning with the full set and iteratively narrowing down toward the least

DRAFT VERSION

efficient units. This approach is more realistic, as it supports gradual, data - driven improvements through comparisons with the nearest efficient benchmarks. It also provides more attainable target values for SDG 4 indicators a critical feature for policy implementation in countries facing larger gaps in performance.

The findings carry significant implications for benchmarking public policies in the area of sustainable development and education. In particular, they contribute to the policy discourse by offering a data-based framework for evaluating national education systems, including those of non-EU countries. By identifying nations with exemplary education policies, the analysis provides potential models for designing more effective systems, with attention to the broader societal implications of education-such as social equity, economic growth, and long-term sustainability. Furthermore, establishing a framework for monitoring SDG 4 implementation encourages countries to expand the scope of their national education policies. The inclusion of SDG 4 benchmarks introduces a key instrument not only for increasing accountability but also for fostering innovation and the development of tailored national tools in education policy. To achieve a comprehensive understanding of SDG 4 implementation, especially as promoted by the European Commission, it is necessary to replace static monitoring approaches with dynamic, iterative evaluation systems.

Finally, while the model used in this study focuses on a limited number of variables, this represents both a limitation and an opportunity. Future research would benefit from the inclusion of a broader range of qualitative and contextual factors, which would enhance the credibility and relevance of the findings.

References

Agbedahin, A. V. (2019), Sustainable development, Education for Sustainable Development, and the 2030 Agenda for Sustainable Development: Emergence, efficacy, eminence, and future, *Sustainable Development,* Vol.27, No.4, pp. 669–680. https://doi.org/10.1002/sd.1931

Charnes, A., Cooper, W.W., Rhodes, E. (1978), *"Measuring the efficiency of decision making units"*, European Journal of operations research, Vol.2, No.6, pp. 429-444. https://doi.org/10.1016/0377-2217(78)90138-8

Ferguson, T., Roofe, C.G. (2020), SDG 4 in higher education: Challenges and opportunities, *International Journal of Sustainability in Higher Education*, Vol. 8, No. 22, pp. 959–975. https://doi.org/10.1108/IJSHE-12-2019-0353

Leicht, A., Combes, B., Byun, W. J., Agbedahin, A. V. (2018), From Agenda 21 to Target 4.7: The development of education for sustainable development*, Issues and trends in Education for Sustainable Development*, pp. 25-38. https://unesdoc.unesco.org/ark:/48223/pf0000261801

Li, D., Wang, M., Lee, C. (2019), The waste treatment and recycling efficiency of industrial waste processing based on two-stage data envelopment analysis DEA with undesirable inputs, *Journal of Cleaner Production*, Vol.242, 118279. https://doi.org/10.1016/j.jclepro.2019.118279

Lim, S., Bae, H. & Lee, L.H. (2011). A study on the selection of benchmarking paths in DEA. *Expert systems with applications*, *38*, 7665-7673

Long, Lj. (2021), Eco-efficiency and effectiveness evaluation toward sustainable urban development in China: a super-efficiency SBM–DEA with undesirable outputs, *Environment, Development and Sustainability*, Vol.23, pp. 14982–14997. https://doi.org/10.1007/s10668-021-01282-7

Nocera Alves Junior, P., Leger, P., & Costa Melo, I. (2024). Efficiency analysis of engineering classes: A DEA approach encompassing active learning and expositive classes towards quality education., *Environmental Science & Policy*, 160, 103856. https://doi.org/10.1016/j.envsci.2024.103856

Seiford, L. M. & Zhu, J. (2003). Context-dependent data envelopment analysis–measuring attractiveness and progress. *OMEGA International Journal of Management Science*, *31*(5), 397–408.

DRAFT VERSION

Tian, N., Tang, S., Che, A., Wu, P. (2019), Measuring regional transport sustainability using super-efficiency SBM-DEA with weighting preference, *Journal of Cleaner Production*, Vol.242, 118474. http://dx.doi.org/10.1016/j.jclepro.2019.118474

Tone, K. (2001). A slacks: based measure of efficiency in data envelopment analysis. *European Journal of Operational Research*, *130*(3), 498-509. http://dx.doi.org/10.1016/S0377-2217(99)00407-5

Wang, K., Wei, Y., Huang, Z. (2018), Environmental efficiency and abatement efficiency measurements of China's thermal power industry: a data envelopment analysis based materials balance approach, *European Journal of Operational Research*, Vol.269, No.1, pp. 35-50. http://dx.doi.org/10.1016/j.ejor.2017.04.053

Zhang, B., Bi, J., Fan, Z., Yuan, Z., Ge, J. (2008), Eco-efficiency analysis of industrial system in China: a data envelopment analysis approach, *Ecological Economics*, Vol. 68, No. (1-2), pp. 306-316. http://dx.doi.org/10.1016/j.ecolecon.2008.03.009

UIS and GEM Report, 2021. SDG 4 Data Digest 2021: National SDG 4 benchmarks: Fulfilling our neglected commitment. Montreal/Paris. 〈https://unesdoc.unesco.org/ark:/48223/pf0000380387〉

UIS, 2023. Official list of SDG 4 indicators. Montreal: UNESCO Institute for Statistics. https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2023/02/SDG4_indicator_list_2023_02_28.pdf.

UN (2012a), The future we want. Resolution adopted by the general assembly on 27 July 2012 (A/RES/66/288). United Nations, New York

United Nations Educational, Scientific and Cultural Organization. (2021b, August 16). Learning Portal. Retrieved from UNESCO's International Institute for Educational Planning: https://learningportal.iiep.unesco.org/en/issue-briefs/monitor-learning/quality-and-learning-indicators

Internet sources

European Commission, 2022. Sustainable development in the European Union Monitoring report on progress towards the SDGs in an EU context. Brussels. http ://ec.europa.eu/commission/presscorner/detail/en/IP_22_3212 (accessed 26 July 2022).

European Environment Agency, 2020. Sustainable Development Goals and the environment in Europe: a cross-country analysis and 39 country profiles. Report 21/ 2020.

https://ec.europa.eu/eurostat/data/database

https://commission.europa.eu/strategy-and-policy/sustainable-development-goals_en

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

APPLICATION OF THE ANALYTICAL HIERARCHY PROCESS FOR SELECTING A MOTOR VEHICLE FOR TRANING CATEGORY "C" DRIVER CANDIDATES IN THE MINISTRY OF DEFENCE AND THE SERBIAN ARMED FORCES

VESKO LUKOVAC[1], LUKA IVEZIĆ [2], PAVLE GLADOVIĆ[3], VLADIMIR POPOVIĆ[4], MILAN STANKOVIĆ[5]

[1] University of Defence in Belgrade, Military Academy, Belgrade, lukovacvesko@gmail.com
[2] University of Belgrade, Faculty of Technology and Metallurgy, Belgrade, ivezicluka94@gmail.com
[3] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, anaipavle@gmail.com
[4] Academy of Tehnical and Educational vocational studies, Niš, vladimir.popovic@akademijanis.edu.rs
[5] Academy of Tehnical and Educational vocational studies, Niš, milan.stankovic@akademijanis.edu.rs

Abstract: *The paper presents a case study on the decision-making process for selecting a motor vehicle for training category "C" driver candidates in the Ministry of Defence and the Serbian Armed Forces. Two vehicle alternatives (MAN TGL 12.180 and FAP 1118) were evaluated based on nine criteria defined by the study participants: vehicle stability, vehicle handling, visibility from the vehicle, cabin comfort, presence of dual controls, vehicle reliability, maintenance costs, fuel consumption, and purchase price. The Analytical Hierarchy Process (AHP), a method from the decision support methodology family, was applied, with emphasis on group decision-making and subsequent synthesis of individual priorities. The results of the group decision-making process favored the MAN TGL 12.180 vehicle.*

Keywords: *The Analytical Hierarchy Process, Motor vehicle, Driver traning.*

1. INTRODUCTION

The functioning of a logistics system largely depends on the efficiency of its transport processes. Motor vehicle drivers operating heavy-duty vehicles represent a key resource within the transport component of large-scale logistics systems. The training process for acquiring the appropriate driving license to operate such vehicles is of strategic importance. The success of driver candidate training greatly depends on the conditions under which it is conducted, which in turn requires a suitable vehicle for instruction.

The subject of this paper is the selection of a motor vehicle for training category "C" driver candidates within the Ministry of Defence (MoD) and the Serbian Armed Forces (SAF). For this purpose, the Analytical Hierarchy Process was employed, one of the most widely used multi-criteria decision-making methods, applicable both at individual and group decision-making levels. The decision-makers (DMs) were driver training instructors from the MoD and the SAF Driver Training Center. The synthesis of individual decisions was carried out subsequently, by aggregating individual priorities using the Weighted Arithmetic Mean Method (WAMM) and the Geometric Mean Method (GMM).

2. AHP DECISION-MAKING

The Analytic Hierarchy Process was developed by Thomas Saaty in the early 1970s. The method is termed “analytic” and “hierarchical” because the decision-maker decomposes the decision problem into multiple decision elements and establishes a hierarchy among them. The hierarchy of the decision problem typically consists of several levels: the goal, criteria, subcriteria, and alternatives.

After the hierarchy has been established, the decision-maker performs pairwise comparisons of the elements at a given level with respect to each element on the level above, in order to determine their relative importance. In the standard AHP approach, elements are compared using linguistic (semantic) judgments of relative importance with respect to the element at the higher level, based on Saaty’s scale from table 1 (Saaty, 1980).

DRAFT VERSION

Table 1: Saaty's scale

Definition	Numerical value (a_{ij})
Absolute dominance of element *i* over element *j*	9
Very strong dominance of element *i* over element *j*	7
Strong dominance of element *i* over element *j*	5
Weak dominance of element *i* over element *j*	3
Equal importance of elements *i* and *j*	1
Weak dominance of element *j* over element *i*	1/3
Strong dominance of element *j* over element *i*	1/5
Very strong dominance of element *j* over element *i*	1/7
Absolute dominance of element *j* over element *i*	1/9
(Intermediates)	(2,4,6,8)

When the decision-maker evaluates n decision elements at a given level of the hierarchy with respect to the element on the level above, using the scale from table 1, their semantic judgments (based on the definitions in the left column) are translated into numerical values (from the right column) and entered into a square matrix A. This matrix is positive and reciprocal (symmetric with respect to the main diagonal). In other words, the elements above the diagonal are reciprocals of those below it, and all elements on the main diagonal are equal to 1 ($a_{ij} = 1/a_{ij}$, for all i and j; $a_{ii} = 1$ for all i), relation 1 (Saaty, 1980).

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \tag{1}$$

If the standard Saaty's scale is used, then each element a_{ij} can assume one of 17 values from the discrete interval [1/9,9]. The process of determining the weights of the compared elements based on the numerical values from matrix A is called prioritization. Prioritization is the process of deriving the priority vector $w = (w_1, \dots, w_n)^T$ from matrix where each $w_i > 0$ and the following condition holds $\sum_{i=1}^{n} w_i = 1$. There are several matrix-based and optimization-based methods for prioritization, however, in this paper, the additive normalization method (AN) was used due to its simplicity and frequent application. To obtain the priority vector w it is sufficient to divide each element in a given column of matrix A by the sum of the elements in that column (normalization), then the normalized values in each row are summed, and each resulting row sum is divided by the rank of the matrix n. This procedure is described by relations 2 and 3 (Saaty, 1980):

$$a_{ij}' = \sum_{i=1}^{n} a_{ij}, ij = 1,2,\dots,n \tag{2}$$

$$w_i = \frac{\sum_{j=1}^{n} a_{ij}'}{n}, i = 1,2,\dots,n \tag{3}$$

Based on the evaluations, the selected prioritization method is used to determine the local weights of decision elements, through synthesis, specifically additive synthesis, the final weights of the alternatives on the lowest level of the hierarchy are calculated with respect to the element at the highest level (the goal), thereby completing the individual AHP decision-making process. Additive synthesis is represented by relation 4 (Saaty, 1980):

$$u_i = \sum_j w_j d_{ij} \tag{4}$$

DRAFT VERSION

where:

- u_i – the final (global) priority of the alternative i,
- w_j – the weight of the criterion j,
- d_{ij} – the local weight of the alternative i with respect to the criterion j [4].

In addition to prioritization methods, one of the key features of the AHP is that the consistency of decision-makers' judgments is verified at all levels of the hierarchy. To check for consistency, Saaty proposed the Consistency Ratio (CR), which is used in AN prioritization methods. Calculating the CR involves two steps. In the first step, the Consistency Index (CI) is calculated using the relation 5 (Saaty, 1980):

$$CI = \frac{\lambda_{max} - n}{n-1} \tag{5}$$

where:

- n – the rank of the matrix,
- $\lambda_{\max}$ – the maximum eigenvalue of the comparison matrix.

In the second step, the Consistency Ratio (CR) is calculated as the ratio of the Consistency Index (CI) to the Random Consistency Index (RI), using relation 6 (Saaty, 1980):

$$CR = \frac{CI}{RI} \tag{6}$$

The Random Consistency Index (RI) depends on the rank of the matrix, and its values are obtained by randomly generating 500 matrices, as shown in table 2 (Saaty, 1980).

Table 2: The values of the random consistency index (RI) depending on the rank of the matrix (Saaty, 1980)

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
RI	0.00	0.00	0.52	0.89	1.11	1.25	1.35	1.40	1.45	1.49	1.52	1.54	1.56	1.58	1.59

If the CR ≤ 0.10, the result indicates that the decision-maker was consistent, and there is no need to repeat the evaluation. If CR > 0.10, the decision-maker should repeat (or modify) their evaluations to improve their consistency (Saaty, 1980; Jandrić & Srđević, 2000).

To combine individual decisions into a group decision in AHP, there are several methods, and in this study, the group decision was made by aggregating individual priorities (*Aggregation of Individual Priorities – AIP*), which is characterized by two aggregation methods:

1) *Weight Arithmetic Mean Method–WAMM*. Given the alternative A_i and its weight (priority) $w_i^{(k)}$ for the k-th decision-maker. If all group members (g) are assigned corresponding weights α_k, the weighted arithmetic mean is calculated according to relation 7:

$$w_i^{(g)} = \sum_{k=1}^{m} w_i^{(k)} \alpha_k \tag{7}$$

where:

- $w_i^{(g)}$ – the final (composite) priority of the alternative A_i,
- m – the number of decision-makers.

Assuming that the individual weights α_k of group members are previously additively normalized, i.e.,

$$\sum_{k=1}^{m} \alpha_k = 1$$

2) *Geometric Mean Method – GMM.* In this method, the aggregation consists of applying relation 8:

DRAFT VERSION

$$w_i^{(g)} = \prod_{k=1}^{m} \left(w_i^{(k)} \right)^{\alpha_k} \tag{8}$$

The weights α_k of the group members are also previously additively normalized.

In both mentioned methods, a final additive normalization of the priorities of all alternatives is required.

3. FORMULATION OF THE HIERARCHICAL DECISON - MAKING PROBLEM

The hierarchy of the AHP decision-making problem is depicted in figure 1. The selection of a motor vehicle for the training of category "C" driver candidates in the Ministry of Defence and the Serbian Armed Forces is the goal of the AHP decision-making process and is positioned at the top of the hierarchy. At the next level are the decision-making criteria (C1, C2, ..., C9): vehicle stability, vehicle handling, visibility from the vehicle, cabin comfort, presence of dual controls, vehicle reliability, maintenance costs, fuel consumption, and purchase price. These criteria were predefined by the research participants (decision-makers) through survey questionnaires. The alternatives (A1 and A2) are at the lowest level of the AHP hierarchy, represented by the motor vehicles MAN TGL 12.180 and FAP 1118, which emerged as favorites based on the results of a survey conducted with the participants, among the vehicles available at the Driver Training Center in the MoD and the SAF (Ivezić, 2025).

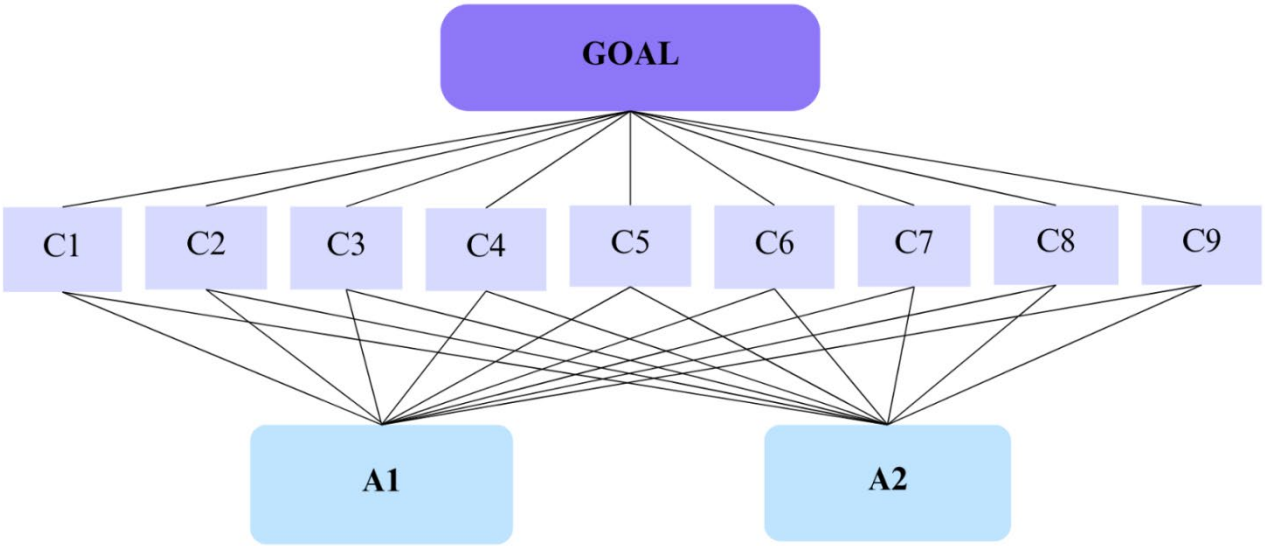


Figure 1: The AHP decision-making hierarchy (Ivezić, 2025)

The group of decision-makers consisted of twelve instructors from the Driver Training Center in the MoD and the SAF, who are directly involved in the training process for category "C" driver candidates. Regarding the work experience of the instructors-decision-makers, 75% of the instructors had more than 20 years of work experience, 17% had between 10 and 20 years of work experience, and 8% had between 5 and 10 years of work experience (Ivezić, 2025).

The AHP was implemented by first familiarizing the decision-makers with the AHP decision-making problem hierarchy and the method of comparing hierarchy elements using Saaty's scale. Subsequently, they were provided with a form containing matrices for pairwise comparisons of the elements of the decision-making problem hierarchy.

4. RESULTS OF AHP DECISION-MAKING

According to the previously described methodology for conducting AHP, the results from the decision-makers' comparison matrices were processed using the Expert Choice 2000 (EC 2000) software. Table 3 presents the results of the individual evaluations of the decision-makers and the Consistency Ratio (CR) they exhibited.

DRAFT VERSION

Table 3: Individual evaluations (Ivezić, 2025)

DM	A1	A2	CR
DM1	0.640	0.360	0.02
DM2	0.753	0.247	0.03
DM3	0.562	0.438	0.01
DM4	0.686	0.314	0.06
DM5	0.670	0.330	0.04
DM6	0.661	0.339	0.04
DM7	0.479	0.521	0.02
DM8	0.683	0.317	0.03
DM9	0.692	0.308	0.02
DM10	0.443	0.557	0.03
DM11	0.494	0.506	0.05
DM12	0.720	0.280	0.05

The group synthesis of individual weight vectors was performed using the AIP WAM and AIP GMM methods. Both methods were tested for two types of weights assigned to the decision-makers, i.e., weights obtained by normalizing the reciprocal values of the decision-makers' Consistency Ratios (1/CR) and weights obtained by normalizing the ratings of their work quality (Rating), as shown in Table 4.

Table 4: Variants of DM weights (Ivezić, 2025)

DM	1/CR	Rating
DM1	0.109	0.081
DM2	0.073	0.078
DM3	0.219	0.093
DM4	0.036	0.082
DM5	0.055	0.089
DM6	0.055	0.076
DM7	0.109	0.092
DM8	0.073	0.083
DM9	0.109	0.077
DM10	0.073	0.085
DM11	0.044	0.087
DM12	0.044	0.079

When the variants of decision-maker weights from table 4 are used for group AIP WAM and AIP GMM synthesis according to relations 7 and 8, the final priorities of the alternatives are obtained, as shown in figure 2.

DRAFT VERSION

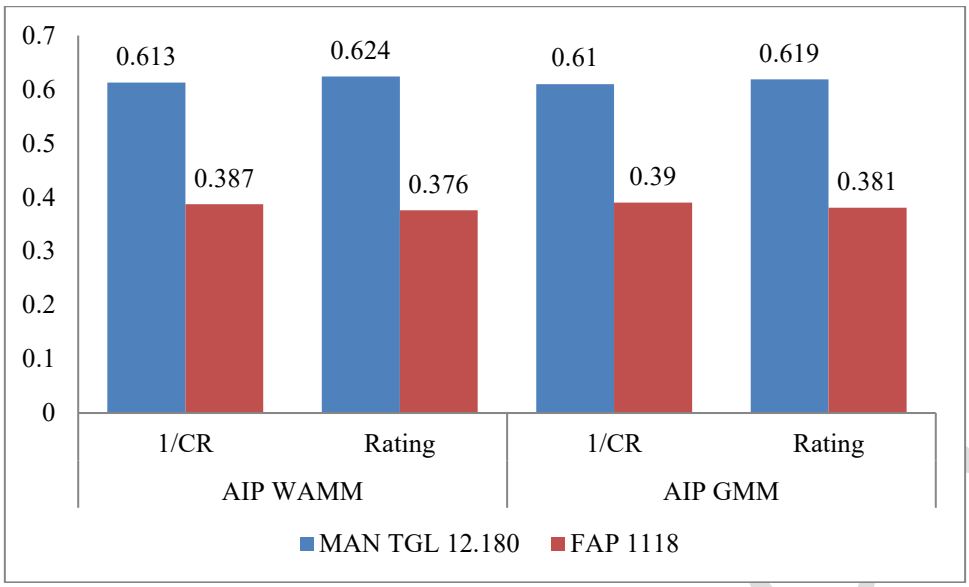


Figure 2: Group AHP synthesis

Since no change in the ranking of alternatives occurred during the dynamic sensitivity analysis (using the Dynamic option of the EC 2000 software) with a 5% variation in the importance of all criteria across all combinations, the final results of the conducted AHP can be considered stable (Hot, 2014).

5. CONCLUSION

The efficiency of the defense system largely depends on the quality of its personnel. Drivers of category "C" motor vehicles in the Ministry of Defense and Serbian Armed Forces represent one of the fundamental pillars of the logistical support system. Operating vehicles with higher payload capacities in specific, often complex conditions requires a high level of knowledge, skills, and psychophysical preparedness. In this context, their training process must not only be formally well-designed but also substantively tailored to real needs and modern standards of safety, effectiveness, and efficiency.

Successful training of driver candidates, in addition to pedagogical methods and quality teaching staff, largely depends on the conditions under which it is conducted, which also requires suitable motor vehicles for their training. This study demonstrates how the well-known multi-criteria analysis method - AHP, can be utilized to select an appropriate motor vehicle for training category "C" driver candidates in the Ministry of Defense and Serbian Armed Forces. The results of the conducted group AHP indicate that the MAN TGL 12.180 is the vehicle that should be used for the practical training of category "C" driver candidates in the MoD and the SAF. Its structural suitability for the driver training process, modern safety systems, high level of reliability, and comfort for both candidates and instructors make it preferable compared to other vehicles available at the Driver Training Center in the MoD and the SAF.

LITERATURE

Hot, I. (2014). Management of conceptual designs creation in field of infrastructure by use of multi-criteria analysis, doctoral dissertation, Faculty of Tehnical Sciences, University of Novi Sad.

Ivezić, L. (2025). Selection of a motor vehicle for traning category "C" driver candidates in the Ministry of Defence and the Serbian Armed Forces, graduation thesis, Military Academy, University of Defence.

Jandrić, Z., Srđević, B. (2000). Analytical Hierarchy Process as a decision-making in water management, Water management 186-188: 327–335.

Saaty, T.L. (1980). The Analytical Hierarchy Process, McGraw–Hill, New York.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

ANYTIME ALGORITHM TO GENERATE NON-DOMINATED VECTORS TO BI-OBJECTIVE INTEGER PROGRAMMING PROBLEMS

NADA MLADENOVIĆ[1], BOGDANA STANOJEVIĆ[2]

[1] University of Belgrade, Faculty of Organizational Sciences, Belgrade, nada.mladenovic@fon.bg.ac.rs, 0009-0002-2825-8348
[2] University of Belgrade, Faculty of Organizational Sciences, Belgrade, bogdana.stanojevic@fon.bg.ac.rs, 0000-0003-4524-5354

Abstract: *In this study we address the bi-objective mixed-integer programming problems. Our solution approach is an adaptation of a general procedure proposed in the recent literature to solve multiple objective optimization problems. The integrality of the decision variables generates a discrete Pareto front that is not easy to be fully disclosed using a framework developed for continual problems. We combine the manner of defining and ordering the sub-sets of the Pareto front (from the continual approach) with a new way of choosing hypothetical bounds to be targeted (specific to the integer optimization problems). We carried out experiments on bi-objective knapsack, and integer problems. Preliminary results show that the proposed approach performs well in comparison to the results reported in the literature.*

Keywords: *Multiobjective optimization, combinatorial problem, anytime algorithm*

1. Introduction

In this paper we propose an anytime solution algorithm to bi-objective mixed-integer programming (2objMIP) problems that is able to generate well distributed vectors on the Pareto front, and report the numerical results of our experiments carried out on bi-objective knapsack problems (2objKP), and an integer programming (2objIP) problem. Within experiments we used Octave programming language and its GLPK solver for all single objective linear optimizations.

Generally, multiple objecvtive (MOIP) problems are widely studied in the recent literature. See for instance (Toloo et al., 2022), (Domínguez-Rios et al., 2021) and (Cai et al., 2019). The state of art and main challenges in solving multiple objective combinatorial problems can be found in (Ehrgott et al., 2016). The so called anytime algorithms are those algorithms that produce a valid (0-level approximation) solution even when they are interrupted before they finish. The quality of the solution derived by such algorithms must be directly improved with the time the algorithms run. In our context: (i) the exact solution (i.e. the Pareto front of a 2objMIP problem) is a set of points, and its 0-level approximation means any of its subsets; (ii) the quality of the approximate solution increases with the running time if and only if, whenever the execution of the algorithm stops, a set of well distributed points along the exact set is yielded. See (Domínguez-Rios et al., 2021) for more details.

(Ehrgott et al., 2016) provided a wide survey of the exact methods to solving multiple objective combinatorial optimization (MOCO) problems. (Deb and Miettinen, 2009) provided a review of nadir point estimation procedures based on evolutionary approaches. They discussed the dimensionality reduction issue. We recall an example from (Ozpeynirci and Koksalan, 2010) where an exact algorithm for finding extreme supported nondominated points to multiple objective mixed integer programs was introduced.

Our algorithm is essentially based on the TDM model introduced in (Stanojevic and Glover, 2020) to solve multiple objective continual problems. We propose a variant of this algorithm, obtained by exploiting the particularity of discrete Pareto front of MOIP problems. The original TDM model was formulated with the help of a hypothetical bound defined as a convex combination of certain already generated non-dominated vectors.

Our algorithm, C-TDM, generates an approximation to the Pareto front of the bi-objective continual problem, namely the problem obtained by ignoring the integral constraints, and use the vectors from that approximation as hypothetical bounds in the integer form of the classic TDM model. C-TDM can be easily adapted to be used in a more general context.

Our paper is structured as follows: Section 2 provides the basic notation and terminology; Section 3 briefly describes the specificity of our solution approach, while Section 4 reports some of our numerical results. The final conclusion and some directions for further researches are included in Section 5.

DRAFT VERSION

2. Preliminaries

The general model of a multiple objective optimization (MOO) problem is given in (1).

$$\max_{x \in X} \left(f_i(x)\right)_{i=\overline{1,p}}, \tag{1}$$

where $p \geq 2$ is the number of objectives, $X \subseteq R^n$ is the feasible set of the decision variables $x \in R^n$, and $f_i : X \to R$, $i = \overline{1,p}$ are the objective functions that all have to be maximized. Particular models are obtained from (1) by imposing various special conditions on variables, feasible sets, and/or objective functions.

Let $F : X \to R^p$ be defined by $F(x) = (f_1(x), f_2(x), \ldots, f_p(x))$. The image of the feasible set in R^p defined by $F(X)$ represents the criterion space. The main concept used in defining solutions to MOO problems is the concept of dominance.

A feasible solution $x' \in X$ is said to dominate another feasible solution x'' – written $F(x') \geq F(x'')$ – if $f_i(x') \geq f_i(x'')$ for all $i \in M$. Further on, x' strictly dominates x'' if and only if x' dominates x'' and in addition $f_i(x') > f_i(x'')$ for at least one index $i = \overline{1,p}$. A feasible solution is called Pareto optimal or Pareto efficient if there is no feasible solution that strictly dominates it.

A vector/point from the criterion space is called non-dominated if it is the image of a Pareto optimal solution. The set of all non-dominated vectors of an MOO problem represents the Pareto front. For more details we refer the reader to (Pardalos et al., 2017).

PESA algorithm was proposed in (Stanojevic and Glover, 2020) to provide a pattern efficient approximation to the Pareto front of a general MOO problem. It used the TDM optimization model that we recalled here in its simplified form

$$\max \left\{ \sum_{i=\overline{1,p}} c_i f_i(x) + p_0 \lambda | x \in X, f_i(x) \geq \lambda y_i^0, i = \overline{1,p}, \lambda \geq 0 \right\}. \tag{2}$$

Compared to its general form, Model (2) uses only one out of two parameters, namely $p_0 \geq 0$, that can be tuned in accordance to the specificity of each approached instance. The other parameter does not affect the optimization results, thus it was removed based on empirically derived conclusions.

In what follows we address multiple objective integer optimization problems, hence we restrict the decision variables x in (1) to take only integer values; and adapt PESA to solve it. The main idea of PESA was to generate one by one well-distributed non-dominated (ND) vectors. Each ND vector, except the extreme ones, was obtained with the help of an hypothetical bound and aimed to cover a gap on the Pareto front. The novelty of our approach is related to the choices of the hypothetical bounds y^0. Details are provided in Section 3.

To evaluate the performances of our algorithms we used the coverage gap (CG) and inverted generational distance (IGD) as diversity indicators within our experiments. Strengths and weaknesses of a wide class of performance indicators were systematically analyzed in (Audet et al., 2021).

IGD metricis widely used and considered as a trusty performance indicator. It calculates the average distance from uniformly distributed points along the Pareto front (P^*) to discrete Pareto front approximation, A, i.e.

$$IGD(A, P^*) = \frac{\sum_{v \in P^*} d(v, A)}{|P^*|}, \tag{3}$$

where $d(v,A)$ is minimal distance between v and points in A in Euclidean sense. See (Deb and Jain, 2014) for more details.

The coverage gap as performance indicator was introduced in (Ceyhan et al., 2019). Given the discrete Pareto front P and the subset A of P,

$$\alpha_A(z) = \min_{y \in A} \left(\max_{i=\overline{1,p}} (z_i - y_i) \right) \tag{4}$$

represents a measure of how well the vector z is covered by A. Then, the coverage gap of A is determined by the vector z^* with the worst coverage, i.e. the vector z^* with the greatest value $\alpha_A(z)$. In other words, the coverage gap of A is the value

$$\alpha_A = \alpha_A(z^*), z^* = \arg\max_{z \in P} (\alpha_A(z)). \tag{5}$$

DRAFT VERSION

Algorithm 1 Simplified PESA pseudocode (adapted from Stanojevic and Glover, 2020)

Input: An instance of (1) and the number of iterations *nrit*.

1: Find marginal solutions and their corresponding extreme ND vectors. Normalize the objective functions.
2: Define the first gap to be covered using the extreme ND vectors. Compute its hyper-volume.
3: Define the first optimization problem using the first gap, its hyper-volume, the extreme ND vectors and the coefficient vectors used as weights in the aggregation function that yielded the marginal solutions. Include this problem in the list of problems
4: **for** $q = \overline{1, nrit}$ **do**
5: Identify in the list the problem P^0 that has the greatest hyper-volume. Compute its corresponding hypothetical bound y^* and aggregation coefficients c^*.
6: Solve (2) with $y^0 = y^*$, $c^0 = c^*$, and derive y^{new} as its solution. Update with y^{new} the list of ND vectors.
7: Remove P^0 from the list of problems and insert p new problems with respect to P^0 and y^{new}.
8: **end for**

Output: The set of generated ND vectors.

Table 1: Summary of the running times obtained for 2KP instances (Gandibleux and Freville, 2000)

Instance	r	n	$\lvert E_{\text{PF}}\rvert$	Gandibleux et al., 2000		Our alg.
				Case 1	Case 2	C-TDM
2KP50-11	0.11	50	43	55	78 / 4	1.5620
2KP50-50	0.5	50	51	801	349	3.1002
2KP50-92	0.92	50	2	7	11 / 1	2.7507
2KP100-50	0.5	100	155	-	-	59.4820

3. Solution algorithm

Aiming to solve 2objMIP problems, we exploit the specific shape of the Pareto front brought by the integer components of its ND vectors, and – compared to PESA – conveniently change the manner of choosing the hypothetical bounds to be used in the integer form of the classic TDM (Stanojevic and Glover, 2020) model.

In what follows we present our C-TDM solution approach. C-TDM algorithm uses the integer relaxation of the original problem to define the hypothetical bound y^0 needed in the TDM model (2). Then, we construct gaps and order them with respect to their size, as described in PESA (Stanojevic and Glover, 2020).

The quality of being an anytime algorithm is assured by the inclusion of PESA framework within our solution approach. PESA is briefly described by Algorithm 1.

Each time, the hypothetical bound used by C-TDM to derive a ND vector to the original 2objMIP problem is a ND vector of the Pareto front of the corresponding continual problem obtained by ignoring the integrality constraints on the original decision variables. An approximation of the Pareto front of the continual problem is generated using the classic PESA method. In other words, y^0 is an already generated ND vector to the integer relaxation of the original 2objMIP problem.

4. Numerical results

4.1. Bi-objectives KP instances

The first group of experiments were performed on Set 1.A 2KP instances downloaded from (International Society on Multiple Criteria Decision Making, Created : October 1998 — Last update: November 2004), and having the following characteristics: two objectives ($p = 2$), one constraint only ($k = 1$), and several tightness ratio (r). See (Gandibleux and Freville, 2000) where the instances were firstly introduced.

The obtained results are summarized in Table 1:, which shows that our C-TDM algorithm is competitive with the algorithms reported in the literature.

The strength of the C-TDM algorithm can be concluded from the detailed results reported for Instance *2KP50-50*. The values of the diversity metrics, IGD and CG, are reported in Table 2:. Values $IGD = 0$ and $CG = 0$ were obtained after 395 iterations proving that all points on the true Pareto front are obtained.

Analyzing the values reported in Table 2: we draw the following conclusions: (i) the generated ND vectors are well dispersed alon the Pareto front; (ii) the quality of the generated approximations to the true Pareto front increases with the number of iterations; and (iii) 300 iterations were needed to generate 50 ND vectors and additional 95 to generate the 51th one.

Figure 1: shows the well-distributed ND vectors generated after a small number of iterations, and the full

DRAFT VERSION

Table 2: Instance 2KP50-50

C-TDM				
#it	#NDs	IGD	CG	Time (s)
8	8	18.4153	0.0138	0.1541
32	27	3.1162	0.0028	0.4099
100	39	0.9906	0.0028	0.9181
300	50	0.0608	0.0005	2.4536
395	51	0	0	3.1002

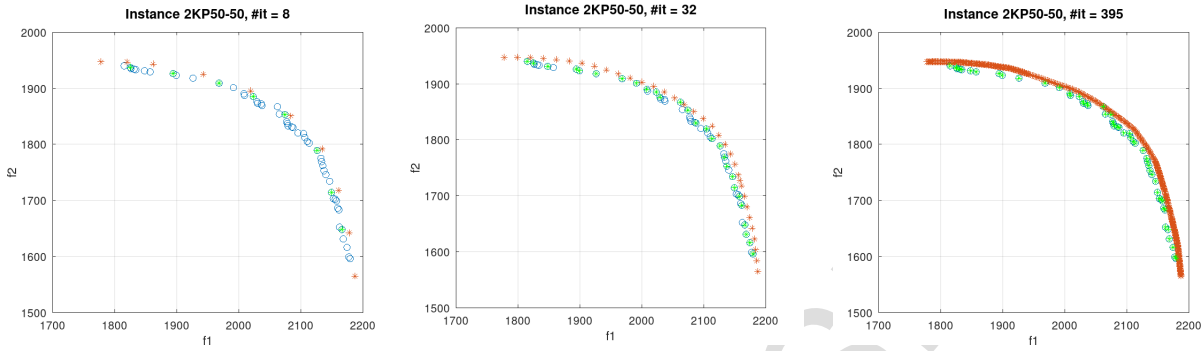


Figure 1: Generated ND vectors on Pareto front of Inst. 2KP50-50 after 8, 32, and 395 iterations respectively

coverage of the Pareto front achieved after 395 iterations. In the first sub-figure can be easily seen that the extrem ND vectors were not disclosed from the beginning.

4.2. Bi-objective MOIP instance

We solved the bi-objective MOIP instance defined by the matrices

$$c = \begin{bmatrix} 6 & 0 & -10 & -23 & 33 & 25 & 27 & -90 & -9 & 27 \\ 92 & 91 & 17 & 75 & 53 & 97 & 31 & -50 & 85 & 93 \end{bmatrix}, \tag{6}$$

$$A = \begin{bmatrix} 33 & 75 & -22 & 80 & 96 & 76 & 69 & 88 & 60 & 81 \\ 20 & 75 & 5 & 55 & -60 & 39 & 0 & 40 & 86 & 0 \\ 6 & 37 & 86 & 45 & 64 & 85 & 44 & 0 & -89 & 0 \\ 23 & -47 & 26 & 0 & -80 & -81 & -18 & 71 & 0 & 57 \\ 2 & -33 & 35 & 0 & 37 & 17 & -13 & 17 & 100 & 57 \end{bmatrix}, b = \begin{bmatrix} 314 \\ 241 \\ 224 \\ 37 \\ 219 \end{bmatrix}. \tag{7}$$

Figure 2: shows the derived ND vectors that fully cover the Pareto front.

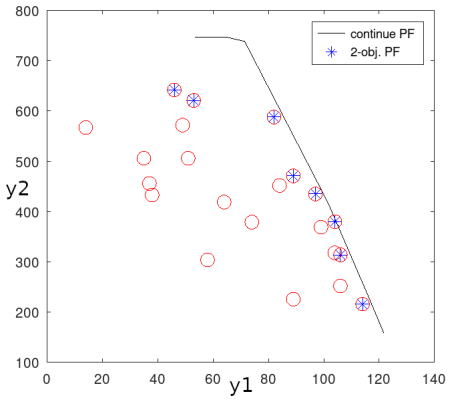


Figure 2: Generated ND vectors on the Pareto front of the bi-objective MIP problem defined using (6)-(7)

DRAFT VERSION

5. Conclusion and further researches

In this study we addressed the 2objMIP problems. Our solution approach is an adaptation of a general procedure proposed in the recent literature to solve continual multiple objective optimization problems. We combined the manner of defining and ordering the sub-sets of the Pareto front (from the continual approach) with a new way of choosing hypothetical bounds to aim. These ways were introduced in accordance to the specific shape of the Pareto fronts of combinatorial optimization problems. We carried out experiments and reported results on bi-objective knapsack and MIP problems. The results obtained for the knapsack problems clearly illustrated that the proposed approach performed well in comparison to the results reported in the literature.

Our plans for future research include wider experiments on larger instances by employing various heuristics for solving single objective mixed-integer programming problems in the framework of our current approach. We will also carry out some more experiments on the same class of problems, aiming to find a way to efficiently tune the problem-specific parameters.

ACKNOWLEDGMENT

This work was supported by the Serbian Ministry of Education through the Faculty of Organizational Sciences, University of Belgrade.

LITERATURE

Audet, C., Bigeon, J., Cartier, D., Le Digabel, S., & Salomon, L. (2021). Performance indicators in multiobjective optimization. *European Journal of Operational Research*, *292*(2), 397–422. https://doi.org/10.1016/j.ejor.2020.11.016

Cai, X., Hu, M., Gong, D., Guo, Y.-n., Zhang, Y., Fan, Z., & Huang, Y. (2019). A decomposition-based coevolutionary multiobjective local search for combinatorial multiobjective optimization. *Swarm and Evolutionary Computation*, *49*, 178–193. https://doi.org/10.1016/j.swevo.2019.05.007

Ceyhan, G., Köksalan, M., & Lokman, B. (2019). Finding a representative nondominated set for multi-objective mixed integer programs. *European Journal of Operational Research*, *272*(1), 61–77. https://doi.org/10.1016/j.ejor.2018.06.012

Deb, K., & Jain, H. (2014). An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: Solving problems with box constraints. *IEEE Transactions on Evolutionary Computation*, *18*(4), 577–601. https://doi.org/10.1109/TEVC.2013.2281535

Deb, K., & Miettinen, K. (2009). A review of nadir point estimation procedures using evolutionary approaches: A tale of dimensionality reduction. In or AUTHOR (Ed.). Proceedings of the Multiple Criterion Decision Making (MCDM-2008) Conference.

Domínguez-Rios, M. A., Chicano, F., & Alba, E. (2021). Effective anytime algorithm for multiobjective combinatorial optimization problems. *Information Sciences*, *565*, 210–228. https://doi.org/10.1016/j.ins.2021.02.074

Ehrgott, M., Gandibleux, X., & Przybylski, A. (2016). Exact methods for multi-objective combinatorial optimisation. In S. Greco, M. Ehrgott, & J. R. Figueira (Eds.), *Multiple criteria decision analysis: State of the art surveys* (pp. 817–850). Springer New York. https://doi.org/10.1007/978-1-4939-3094-4_19

Gandibleux, X., & Freville, A. (2000). Tabu search based procedure for solving the 0-1 multiobjective knapsack problem : The two objectives case. *ournal of Heuristics*, *6*(3), 361–383.

International Society on Multiple Criteria Decision Making. (Created : October 1998 — Last update: November 2004). MCDM Numerical Instances Library [Accessed: January 22nd, 2025].

Ozpeynirci, O., & Koksalan, M. (2010). An exact algorithm for finding extreme supported nondominated points of multiobjective mixed integer programs. *Management Science*, *56*(12), 2302–2315.

Pardalos, P. M., Zilinskas, A., & Zilinskas, J. (2017). *Non-convex multi-objective optimization*. Springer.

Stanojevic, B., & Glover, F. (2020). A new approach to generate pattern-efficient sets of non-dominated vectors for multi-objective optimization. *Information Sciences*, *530*, 22–42.

Toloo, M., Talatahari, S., Gandomi, A. H., & Rahimi, I. (2022). 1 - multiobjective combinatorial optimization problems: Social, keywords, and journal maps. In M. Toloo, S. Talatahari, & I. Rahimi (Eds.), *Multi-objective combinatorial optimization problems and solution methods* (pp. 1–9). Academic Press. https://doi.org/10.1016/B978-0-12-823799-1.00010-3

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

ADAPTIVE MULTI-CRITERIA OPTIMIZATION OF EV CHARGING DECISIONS USING XGBOOST AND TOPSIS

JELENA VELIMIROVIĆ[1], LAZAR VELIMIROVIĆ [2], PETAR VRANIĆ[3]

[1] Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade, jelena.velimirovic@mi.sanu.ac.rs
[2] Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade, lazar.velimirovic@mi.sanu.ac.rs
[3] Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade, petarvvv@mi.sanu.ac.rs

Abstract: *The paper presents user-specific multi-criteria analysis integrating predictive machine learning to develop a decision-making framework for electric vehicle charging optimization. The framework utilizes an XGBoost model trained on both synthetic and real-world data to predict optimal charger selections across various station-time combinations, with the goal of minimizing charging duration. Three user profiles — time-sensitive, price-sensitive, and speed-sensitive — are incorporated into a TOPSIS-based optimization, which utilizes predictions and includes charging price and speed. The results show that varying user types have significantly different preference for choices: each time-sensitive user chooses fast chargers day / time, each price-sensitive user chooses standard but less expensive chargers, and speed-sensitive users select chargers that provide the highest power charging rate independent of wait time and price. The limitations of static optimization are demonstrated, as well as value of a custom and adaptable approach. The model achieves high prediction accuracy and provides intelligent, user-centric recommendations regarding electric vehicle charging options.*

Keywords: *XGBoost, TOPSIS, Multi-criteria decision making, EV charging.*

1. INTRODUCTION

The accelerating transition toward electric vehicles (EVs) has created needs for intelligent, personalized, and adaptive EV charging systems. With the growth in charging infrastructure, considerations of user heterogeneity pose an increasing challenge to classical recommendation-based models that have considered static parameters such as location or price. Traditional methods often overlook dynamic user preferences, resulting in low user satisfaction (Habbal & Alrifaie, 2024; Lee et al., 2025). Hence, there is a growing need for decision-making that can predict user's needs and adapt according to user specification.

Artificial intelligence (AI) evolution and the multi-criteria decision making (MCDM) paradigm are seen as promising for improving efficiency in EV charging. For instance, Habbal and Alrifaie (2024) noted that TOPSIS-AHP greatly improves recommendation reliability when compared to heuristics considering price or distance only. Zuo et al. (2023) pointed out the importance of multidimensional profile of users and offered an optimization approach that includes collaborative filtering that aim to balance cost, time and energy. In parallel, there is a growing body of research that applies machine learning techniques to forecast dynamic patterns of demand. For example, Shern et al. (2024) compared efficiency of several prediction models and found out that XGBoost is the one that most persistently outperform other in decreasing forecasting errors. Mahato et al. (2023) introduced a hybrid Jaya–Multi-Verse Optimizer to minimize charging costs and maximize scheduling efficiency.

Increasing demand for flexible decision-making is also evident from recent research on EV route planning. Relying on real-world data, Velimirović et al. (2023a) introduced hybrid methods based on AHP and dynamic programming that allow each EV user to schedule charging stations and routes based on predefined values such as travel time, charging duration, and location attractiveness. Building upon this, Velimirović et al. (2023b) further extended their approach by proposing a hybrid algorithm combining dynamic programming with heuristic components to solve the Electric Vehicle Routing Problem (EVRP). Their model not only balances user needs with travel and charging constraints but also incorporates route optimization under real-world constraints, demonstrating the importance of personalized and adaptive planning in EV mobility

DRAFT VERSION

scenarios. Similar study for planning routing and infrastructure includes the main personal limitations on space, behavior, and networks applying hybrid *p*-median – AHP approach (Janjić et al., 2021).

Two investigations carried out by Ou et al. (2024) and Feng et al. (2025) additionally support the advantages of hybrid models. Studies presented agent systems based on a large language model for understanding user language and consequently for proposing an optimal decision. These researches explain the recent incline toward user-friendly optimization systems, with real-time data-driven information being essential for decision support systems in an EV charging infrastructure.

This study proposes a hybrid framework that integrates predictive machine learning with MCDM to support adaptive, user-specific EV charging decisions. Unlike prior approaches that treat prediction and optimization separately, our model combines both, using three user profiles to generate dynamic, personalized charger-time recommendations.

2. METHODOLOGY

Extreme Gradient Boosting - XGBoost - is a significant component of our methodology and allows us to predict how EV users will use an EV charging station, particularly the predicted duration of each charging session. XGBoost works based on gradient-boosted decision trees (Mustapha & Saeed, 2016; Xia et al., 2017), which is a type of ensemble learning method that captures complex, and relatively unknown non-linear relationships and interactions among some variables. In our case, variables refer to the type of station, time of day, prior congestion level, and trends in user behavior. A substantial element in our optimization methodology is the estimated charging time at each option, which feeds into our multi-criteria weighted ranking. Ultimately, XGBoost will provide an improved input data for the TOPSIS algorithm, which will improve realism and personalization of the optimization process in consideration of actual feedback from the user and station interactions that shape frequent user behavior.

The set of functions f_t in the regression tree model can be learned by minimizing the objective function (Chen et al., 2019):

$$Obj^{(t)} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 = \sum_{j=1}^{T}\left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \quad (1)$$

where $I_j = \{ i | q(x_i) = j \}$ indicates the instance set of leaf j, while the gain formula is described in detail in (Chen et al., 2019).

The proposed model merges TOPSIS decision-making with predictive data modeling using the XGBoost algorithm. To produce predictions based on how long users will stay at each station throughout the day, we utilize historical data from a four-station charging facility, which has two fast and two standard stations.

2.1. Formation of Charging Alternatives Based on Station-Time Slot Combinations

In order to construct an authentic and thorough decision space for EV users, we capture both the spatial (charging station) and temporal (time of charging) aspects by defining a discrete choice set of 16 options. This is accomplished through the combination of four stations — two fast-charge (S1_Fast and S2_Fast) and two standard-charge (S3_Stand and S4_Stand) — together with four time intervals (08:00–11:00, 11:00–14:00, 14:00–17:00, and 17:00–20:00). Any given combination of station and time frame represents a plausible charging alternative, forming the basis of the user's decision problem: "When and where should I charge my vehicle?" Each of these options is evaluated across three quantitative criteria. First, expected user time at the station — including both waiting and charging duration — is predicted using a trained XGBoost regression model based on historical user data, which reflects temporal patterns such as day-of-week, seasonal variation, and typical user behavior. Second, charging cost per kWh varies by charger type and time of day, reflecting dynamic pricing mechanisms that are especially sensitive to peak demand periods (e.g., 11:00–17:00). Third, power output differs by station type, with standard chargers providing 22 kW and fast chargers offering 50 kW, directly impacting the effective charging speed and user satisfaction.

To formalize the decision process, we apply the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). Here, the three criteria are classified as either cost criteria (duration and cost) or benefit criteria (speed). The resulting decision matrix is normalized and weighted according to user-specific preferences. We define three representative user profiles: (1) time-sensitive users, who prioritize minimizing wait and charge duration (weights: [0.6, 0.2, 0.2]); (2) price-sensitive users, who focus on minimizing cost (weights: [0.2, 0.6, 0.2]); and (3) speed-sensitive users, who prioritize maximum charging power regardless of

DRAFT VERSION

time or cost (weights: [0.2, 0.2, 0.6]). Each user case is processed separately through the TOPSIS method to generate a ranked ordering of the 16 available charging options.

A context-sensitive decision process involves selecting both a charging station and time slot, reflecting dynamic pricing and congestion. This setup enables the use of MCDM techniques like TOPSIS and supports adaptive optimization based on user preferences.

2.2. Construction of the Decision Matrix for TOPSIS Evaluation

After defining the set of 16 alternatives through combinations of charging stations and time intervals, we develop a decision matrix suitable for multi-criteria decision making. The three evaluation criteria established previously will quantify each alternative: 1. Cost criteria is anticipated duration (minutes); 2. Cost criteria is charging price (€/kWh); 3. Benefit criteria of charging power output (kW).

These three criteria are used to populate a decision matrix $D \in \mathbb{R}^{16\times 3}$, where each row corresponds to a unique charging alternative (station × time slot), and each column represents one evaluation criterion. Formally, the matrix D is defined as:

$$D = \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ \vdots & \vdots & \vdots \\ d_{161} & d_{162} & d_{163} \end{bmatrix}, \tag{2}$$

where:

- d_{i1} : predicted duration for alternative i,
- d_{i2} : charging price for alternative i,
- d_{i3} : power output for alternative i.

Since the criteria are measured in different units (minutes, €/kWh, kW), min-max normalization is applied to scale all values to the [0,1] range. For cost-type criteria (duration and price), values are inverted after normalization so that higher values consistently indicate better options, aligning all criteria under a "higher is better" interpretation. After normalization, a weight vector $w = \left[w_1 , w_2 , w_3 \right]$ is applied to reflect the user's preferences for each criterion: w_1 = weight assigned to predicted duration, w_2 = weight assigned to charging price, and w_3 = weight assigned to charging speed.

The set of weight vectors will simulate varying user profiles, as described in the adaptive TOPSIS assessment in the next section. The constructed and normalized matrix will be the input to the TOPSIS algorithm; ranking the alternatives based on their similarity to the ideal dataset and their distance from the worst-case data set allows for multi-criteria recommendation systems that are flexible and sensitive to user strategies and priorities.

2.3. Application of the TOPSIS Method for Multi-Criteria Ranking

To find the best charging option based on user preferences, we use the TOPSIS method, a commonly applied multi-criteria decision-making (MCDM) method (Yoon & Hwang, 1981; Shanian & Savadogo, 2006; Siregar, 2019). TOPSIS ranks alternatives based on their relative distance to the ideal solution, valuing both benefit-type and cost-type criteria.

To portray realistic and contextual decision-making behavior we consider a typical EV user who will employ three decision strategies associated with dynamic user preferences. Rather than consider static priorities, our user will evaluate charging station alternatives from three unique vantages: charging station options that minimize total charging time, minimize the costs to charge the EV, and maximize the chargers speed to charge the EV. Each user profile is defined by a specific weight vector across the three criteria, generating a unique preference-based ranking. Time-sensitive users seek to minimize total charging time, so they input a high weight of 0.6 for duration, and low weights of 0.2 for price and charger speed. Price-sensitive users seek to minimize total cost, so they put their greatest weight (0.6) on charging price and only a weight of 0.2 on both duration and speed. Speed-sensitive users are looking for the highest power output regardless of cost or wait time and mean to transfer as much energy as possible to the battery as fast as possible, so they weight charger speed at 0.6 and duration and price both at 0.2.

DRAFT VERSION

This way of creating rankings allows analysis of the optimal station-time-slot options vary depending on the user's current preferred profile. The adaptive rankings that we generate from our TOPSIS method, based on each weight assumption, provide the user with a ranking, and an understanding of how those rankings vary based on changing decision contexts caused by trading off the different alternative combinations.

Each user profile generates a distinct TOPSIS ranking of the 16 station-time alternatives, reflecting personalized trade-offs between duration, cost, and charging speed. This enables adaptive recommendations tailored to individual user priorities. The framework supports flexible, real-time decision-making in dynamic environments. Detailed numerical results, including scores and ranked options for each profile, are presented in the following section.

3. Numeric Results and Discussion

The dataset in this study was real-world data from a real-world application, including 3.395 records of EV charging sessions from 85 users of 105 charging stations at 25 sites (Asensio et al., 2020). The real-world dataset initially included 24 attributes, including time-related session attributes, energy attributes, user IDs, and station attributes. A thorough feature selection process was completed in order to narrow the criterion variables for model relevance and interpretability which resulted in selected times, charging duration, user ID, user and station IDs, and location indicators. Additional features were engineered to provide better review of charging performance such as total energy delivered, cost estimates based on dynamic pricing, and congestion measures by utilizing historical data. The synthetic dataset replicates user–station interactions from the original data and simulates time-based behavior over a full quarter. The variety of the full dataset with the modeling capability of XGBoost developed robust prediction of user duration at charging alternatives, this predicted duration served as base input for the TOPSIS-based optimization framework.

To assess model performance, the standard regression performance measures were employed: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2), based on a strong 5-fold cross validation (generalizability and reduced risk of overfitting).

The XGBoost model represented the greatest performance at MAE = 2.15, RMSE = 2.83, and $R^2 = 0.85$, reflecting an improvement of up to 16% in RMSE, and greater than 20% in MAE over Decision Tree and Random Forest models, underscoring its strength and efficacy in accurately modeling EV charging durations.

Figure 1 indicates a strong linear relationship in the predicted versus observed congestion values along the diagonal, also showing that the bias is approximately equal over the range of values.

Along with regression-based validation, we also evaluated the model's top-1 user-level decision accuracy in selected time slots during the day. We found that the model accurately predicted the preferred charging option for 76.1% of users at 08-11h, 78.0% at 11-14h, 74.2 at 14-17h, and 75.8% at 17-20h. These values present evidence of the operational validity of the model, particularly where predictive as it behaved correctly at applicable times for charging owing to the mid-day decision opportunity. This consistency supports its use as a real-time decision support tool.

Adaptive TOPSIS results show that optimal charging decisions vary significantly by user profile (Table 1). When the model prefers shorter charging durations (time-sensitive users), fast chargers in time slots of lower congestion such as midday (11-14h) are soon viable options. However, we still find some standard stations in the top one or two choices, if they have particularly low predicted waiting times. This indicates that users with the time constraint may sometimes place convenience ahead of the speed of charging.

Alternative choices change dramatically for users who are price sensitive. The highest ranked charging options are almost all standard stations with the lowest cost rates. While standard stations typically lead to longer charge times, users willing to put up with the time trade-off in order to save costs will tend to select such options.

DRAFT VERSION

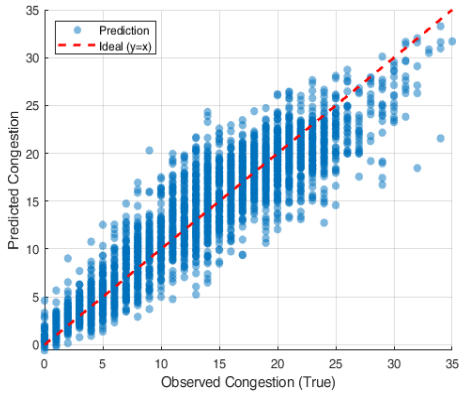


Figure 1: XGBoost: True vs. Predicted Value of Congestion

Table 1: TOPSIS Rankings of Charging Alternatives by User Profile

Alternative	Duration	Price	Speed	TOPSIS TimeSensitive	Rank	TOPSIS PriceSensitive	Rank	TOPSIS SpeedSensitive	Rank
S2_Fast_2	25.804	0.4	50	0.759747	1	0.320377	13	0.759747	3
S1_Fast_2	29.472	0.4	50	0.754812	2	0.31753	14	0.759212	4
S4_Stand_2	28.702	0.3	22	0.726891	3	0.498401	9	0.265536	9
S4_Stand_1	138.99	0.2	22	0.306195	11	0.700567	1	0.246901	11
S3_Stand_1	146.63	0.2	22	0.28106	13	0.694707	2	0.244404	12
S3_Stand_4	151.73	0.2	22	0.266949	15	0.690766	3	0.243005	13
S1_Fast_1	139.95	0.3	50	0.334073	9	0.515048	5	0.761372	1
S2_Fast_4	141.01	0.3	50	0.330328	10	0.514415	6	0.76012	2
S2_Fast_2	25.804	0.4	50	0.759747	1	0.320377	13	0.759747	3

The TOPSIS scores illustrate this shift in alternatives by penalizing higher price stations irrespective of charging speed. Thus alternatives that would be nearly eliminated in cost- or time-based analyses emerge as highest recommendations in price-based optimization scenarios.

On the other hand, speed-sensitive users exhibit a strong and consistent preference over charging with high-power fast chargers over options linked with longer predicted times, even at moderate cost. For speed-sensitive users, the technical capability of a 50 kW station is better than the potential wait times and costs. While alternatives in terms of early morning or late afternoon on fast stations are still top scores, their success reflects the relationship between technical capability and user intentions that change time related to cost.

One key insight from the three profiles is the variation in rankings. The same alternative can be in the lead in one context, and not lead a few months later. For example, a fast station in a low-congestion time slot, may rank first for a time-sensitive user, but when price dominates the decision then it can fall considerably down the ranking. This illustrates the need to update the decision-making framework to represent personal priorities rather than to adopt an optimal solution for all.

In addition to the analysis of preferences across the subgroups of profiles, the outcomes provide another useful perspective to develop our understanding of temporal patterns. The time slots selected by some profiles, in particular, time slots such as 11-14h are characteristics preferred by several profiles, which may indicate an architecture style (station load and price structure) that may be helpful when planning infrastructure. For instance, there were time slots that were selected at very low rates (e.g., 17-20h) that were frequently shown as the lowest preferred alternatives, providing support for sending congestion-aware decisions.

The model effectively captures interactions between user type, station type, and time of day—crucial for developing intelligent EV guidance systems. These systems can adapt based on user behavior, enabling highly personalized charging plans. Unlike static methods, this approach supports real-time, data-driven recommendations aligned with individual preferences. Overall, the findings highlight the need for user-centric, preference-aware optimization in EV infrastructure. Combining machine learning with adaptive MCDM methods like TOPSIS enables personalized navigation and energy management, reflecting user-specific trade-offs beyond technical parameters alone.

4. CONCLUSION

DRAFT VERSION

This paper presents a hybrid framework that integrates XGBoost predictions with TOPSIS-based multi-criteria optimization to generate personalized EV charging recommendations. By modeling three user profiles, the approach adjusts to individual preferences and system conditions, and performs better than static decision methods. Results show high predictive accuracy and effective preference-based differentiation which supports the use of real-time, user-focused guidance in smart EV infrastructure.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia through the Mathematical Institute SANU.

LITERATURE

Asensio, O. I., Apablaza, C. Z., Lawson, M., & Walsh, S. E. (2020). Replication data for: A field experiment on workplace norms and electric vehicle charging etiquette. Tech. Rep., doi:10.7910/DVN/NFPQLW

Chen, M., Liu, Q., Chen, S., Liu, Y., Zhang, C. H., & Liu, R. (2019). XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system. *IEEE Access*, *7*, 13149-13158.

Feng, J., Cui, C., Zhang, C., & Fan, Z. (2025). Large Language Model Based Agent Framework for Electric Vehicle Charging Behavior Simulation. In T.T., Lie, C., Shen, G.,Li & Y., Liu. (Eds.), *2025 10th Asia Conference on Power and Electrical Engineering (ACPEE)*, (pp. 2683–2687). https://doi.org/10.1109/ACPEE64358.2025.11040586

Habbal, A., & Alrifaie, M. F. (2024). A User-Preference-Based Charging Station Recommendation for Electric Vehicles. *IEEE Transactions on Intelligent Transportation Systems, 25*(9), 11617–11634. https://doi.org/10.1109/TITS.2024.3379469

Janjić, A., Velimirović, L., Velimirović, J., & Vranić, P. (2021). Estimating the optimal number and locations of electric vehicle charging stations: the application of multi-criteria p-median methodology. *Transportation Planning and Technology, 44*(8), 827–842. https://doi.org/10.1080/03081060.2021.1992177

Lee, D., Kim, D. S., Chung, B. J., & Chung, Y. M. (2025). Recommendation of Electric Vehicle Charging Stations in Driving Situations Based on a Preference Objective Function. *World Electric Vehicle Journal, 16*(4), 192. https://doi.org/10.3390/wevj16040192

Mahato, D., Aharwal, V. K., & Sinha, A. (2023). Multi-objective optimisation model and hybrid optimization algorithm for Electric Vehicle Charge Scheduling. *Journal of Experimental & Theoretical Artificial Intelligence, 36*(8), 1645–1667. https://doi.org/10.1080/0952813X.2023.2165719

Mustapha, I. B., & Saeed, F. (2016). Bioactive molecule prediction using extreme gradient boosting. *Molecules*, *21*(8), 983. https://doi.org/10.3390/molecules21080983

Ou, P., Wang, Y., Lin, W., & Wu, J. (2024). An LLM-Based Modeling and Decision Optimization for User-Centric Electric Vehicle Charging. 2024 IEEE 8th Conference on Energy Internet and Energy System Integration (EI2), 4078–4083. https://doi.org/10.1109/EI264398.2024.10991378

Shanian, A., & Savadogo, O. (2006). TOPSIS multiple-criteria decision support analysis for material selection of metallic bipolar plates for polymer electrolyte fuel cell. *Journal of Power Sources*, *159*(2), 1095-1104. https://doi.org/10.1016/j.jpowsour.2005.12.092

Shern, S. J., Sarker, M. T., Haram, M. H. S. M., Ramasamy, G., Thiagarajah, S. P., & Al Farid, F. (2024). Artificial Intelligence Optimization for User Prediction and Efficient Energy Distribution in Electric Vehicle Smart Charging Systems. *Energies, 17*(22), 5772. https://doi.org/10.3390/en17225772

Siregar, I. (2019, December). Supplier selection by using analytical hierarchy process (ahp) and techniques for order preference methods with similarities to ideal solutions (topsis). In *Journal of Physics: Conference Series* (Vol. 1339, No. 1, p. 012023). IOP Publishing. doi:10.1088/1742-6596/1339/1/012023

Velimirović, L. Z., Janjić, A., Vranić, P., Velimirović, J. D., & Petkovski, I. (2023a). Determining the Optimal Route of Electric Vehicle Using a Hybrid Algorithm Based on Fuzzy Dynamic Programming. *IEEE Transactions on Fuzzy Systems, 31*(2), 609–618. https://doi.org/10.1109/TFUZZ.2022.3205045

Velimirović, L. Z., Janjić, A., Vranić, P., Petkovski, I., & Velimirović, J. D. (2023b, November). Solving Electric Vehicle Routing Problem Using Hybrid Algorithm Based on Dynamic Programming. In *2023*

DRAFT VERSION

31st Telecommunications Forum (TELFOR) (pp. 1-4). IEEE. https://doi.org/10.1109/TELFOR59449.2023.10372746

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert systems with applications*, *78*, 225-241. https://doi.org/10.1016/j.eswa.2017.02.017

Yoon, K., & Hwang, C. L. (1981). TOPSIS (technique for order preference by similarity to ideal solution)–a multiple attribute decision making, w: Multiple attribute decision making–methods and applications, a state-of-the-at survey. *Berlin: Springer Verlag*, *128*, 140. http://dx.doi.org/10.1007/978-3-642-48318-9

Zuo, X., Bi, J., Wang, Y., & Du, Y. (2023). A Charging Guidance Optimization Model for Electric Vehicle Travel by Considering Multi-Dimensional Preferences of Users. *World Electric Vehicle Journal, 14*(7), 171. https://doi.org/10.3390/wevj14070171

DRAFT VERSION

OR APPLICATIONS IN CONSTRUCTION

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

Original article

FUZZY ADAPTIVE CPM-PERT METHOD

NATAŠA POPOVIĆ-MILETIĆ[1], GORAN ĆIROVIĆ[2], DRAGAN PAMUČAR [3]

[1] University of Banja Luka, Faculty of Architecture, Civil Engineering and Geodesy, Banja Luka, natasa.popovic-miletic@aggf.unibl.org 0009-0000-4478-8432
[2] University od Novi Sad, Faculty of Technical Sciences, Novi Sad, cirovic@orion.rs 0000-0001-7419-8695
[3] University of Belgrade, Faculty of Organizational Sciences, Belgrade, dragan.pamucar@fon.bg.ac.rs 0000-0001-8522-1942

Abstract: *This paper presents a novel methodology for assessing the criticality of project paths and activities within a fuzzy logic framework. The proposed approach integrates the relative degree of criticality and the agreement index into a unified indicator, offering a more precise and comprehensive measure of criticality. Research findings indicate that the fuzzy adaptive CPM-PERT method can be effectively applied in project planning. Its application is particularly suitable for the construction industry, where activity durations and cost estimates are often subject to a high degree of uncertainty and subjectivity.*

Keywords: *Project planning, Fuzzy, CPM, PERT, Critical Degree.*

1. INTRODUCTION

Project management in construction is a complex task, filled with uncertainty and risks. The planned project completion time is often not met, leading to delays in the scheduled construction period. Quality planning includes a realistic assessment of the optimal balance between time, available resources, and construction costs. Scientists and researchers have devoted most of their attention to improving the most popular project planning methods: the Critical Path Method (CPM) and the Program Evaluation and Review Technique (PERT), which consider activity durations as deterministic or probabilistic values, respectively.

Activity durations are imprecise and vague because they are mostly estimated by experts based on their experience. For this reason, the use of fuzzy set theory is proposed for modeling uncertainty instead of probabilistic methods in network planning theory, resulting in new, modified project planning methods, the most well-known of which are the Fuzzy Critical Path Method (FCPM) and the Fuzzy PERT Method (FPERT). This paper presents a new modified method called the Fuzzy Adaptive CPM-PERT, designed for improved project planning under uncertainty.

2. LITERATURE REVIEW

The theory of fuzzy sets (Fuzzy Set Theory) is an alternative approach for handling uncertainty in activity durations. Fuzzy logic measures imprecision or vagueness in estimation and may be preferable to probability theory in situations where historical data on activity durations is unavailable or unreliable (Liberatore, 2016).

Planning in cases where the duration of activities or project costs are given as fuzzy numbers is particularly suitable for application in construction, considering the high level of subjectivity that arises during the estimation of construction activities (Bakri et al. 2013).

In this paper, activity durations are represented using triangular fuzzy numbers Ft(a, b, c), as shown in Figure 1.

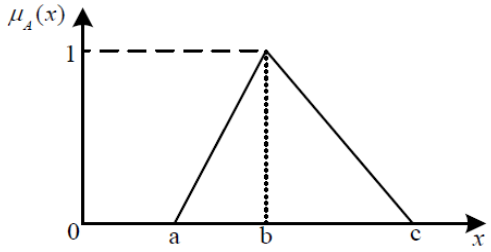


Figure 1: Triangular fuzzy number A = (a, b, c)

DRAFT VERSION

The triangular fuzzy number A = (a, b, c) has a membership function μA(x): $\mathbb{R} \rightarrow [0, 1]$ defined as follows:

$$\mu_A(x) = \begin{cases} 0, & x<a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ \frac{c-x}{c-b}, & b \leq x \leq c \\ 0, & x>c \end{cases} \quad (1)$$

The values a, b, and c represent real numbers that satisfy the condition a > 0 and a ≤ b ≤ c.

Below are the basic arithmetic operations of addition, subtraction, and multiplication on triangular fuzzy numbers A = (a_1, b_1, c_1) and B = (a_2, b_2, c_2) (Dubois & Prade, 1978).

$$A + B = (a_1, b_1, c_1) + (a_2, b_2, c_2) = (a_1 + a_2,\ b_1 + b_2,\ c_1 + c_2) \quad (2)$$

$$A - B = (a_1, b_1, c_1) - (a_2, b_2, c_2) = (a_1 - a_2,\ b_1 - b_2,\ c_1 - c_2) \quad (3)$$

$$A \times B = (a_1, b_1, c_1)x(a_2, b_2, c_2) = (a_1 \times a_2,\ b_1 \times b_2,\ c_1 \times c_2) \quad (3)$$

$$k \times A = k \times (a_1, b_1, c_1) = (k \times a_1, k \times b_1, k \times c_1) \quad (4)$$

Defuzzification is the process of representing a fuzzy set with a specific number, i.e., a fuzzy (uncertain) number is expressed as a deterministic (clear or "crisp") number, that is, a real number. In project management, input data can be represented by fuzzy numbers; however, it is often required that the output data (e.g., project duration) be a clear "crisp" number.

There are several methods that enable defuzzification, and the most commonly used are: the maximum method, the center of area method, the mean of maximum method, the bisector method, the smallest of maximum method, etc. The choice of which method to apply depends on the nature of the problem the fuzzy model is intended to solve. The most frequently used defuzzification method is the *Center of Area (COA)* method, which determines the center of the area of a fuzzy set or fuzzy number (Shaheen et al., 2007). In the literature, another name for the COA method can be found, which is the *Center of Gravity (COG)* method.

Defuzzification of a fuzzy number using the COA (Center of Area) method is performed using the following formula (Moselhi, 2015):

$$y^* = \frac{\int x \cdot \mu_i(x_i)}{\int \mu_i(x_i)} \quad (5)$$

where is:

y*	clear "crisp" number, obtained by defuzzification of a fuzzy number
μ	membership function of the fuzzy number
x	output variable

$$E(A) = \frac{a+2b+c}{4} \quad (6)$$

$$Var\ (A) = \sigma^2(A) = \frac{a^2+b^2+c^2-ab-ac-bc}{18} \quad (7)$$

The use of fuzzy numbers in project planning facilitates the project planner in defining activity durations. Activity durations do not need to be expressed as simple deterministic ("crisp") values, as is typically the case in traditional project planning methods. For example, the project planner does not have to express their subjective estimate that an activity will last approximately 15 days as a crisp number (15); instead, it can be represented using a triangular fuzzy number (10, 15, 20). The planner can determine a confidence interval and estimate the duration as "approximately between 13 and 18 days," which can also be modeled using the triangular fuzzy number (10, 15, 20).

3. RESEARCH METHODOLOGY

In the first phase of the experiment, fuzzy durations of activities were determined. First, it is necessary to define the list of activities with their mutual relationships in the network diagram and to define the fuzzy durations of the activities. This study will use a small project consisting of 10 activities, whose network plan is shown in Figure 2, while the durations of the activities and their interdependencies are

DRAFT VERSION

presented in Table 1. For simplicity of calculations, a "finish-to-start" (FS) relationship between activities is assumed.

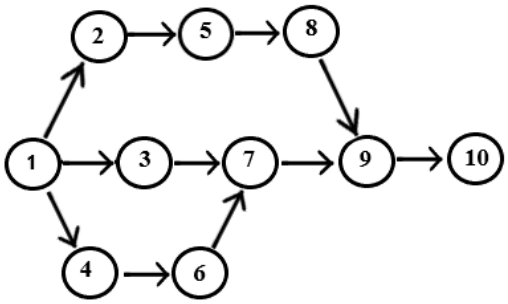


Figure 2: Project Network Diagram

Table 1: PERT activity durations

			Activity durations according to PERT						
Activity	PA	NA	t_o	t_m	t_p	t_e	σ	σ^2	Critical activity
1		2,3,4	3	4	7	4.33	0.67	0.44	yes
2	1	5	4	5	7	5.17	0.5	0.25	
3	1	7	11	14	18	14.17	1.17	1.361	yes
4	1	6	2	5	7	4.83	0.83	0.694	
5	2	8	5	7	10	7.17	0.83	0.694	
6	4	7	4	6	8	6	0.67	0.444	
7	3,6	9	8	12	14	11.67	1	1	yes
8	5	9	2	4	6	4	0.67	0.444	
9	7,8	10	11	14	18	14.17	1.17	1.361	yes
10	9		8	10	12	10	0.67	0.444	yes
For the critical path Σ						54.33	4.67	4.611	

Meaning of the symbols in Table 1: PA – Predecessor activity; NA – Successor activity; t_o – Optimistic duration; t_m – Most likely duration; t_p – Pessimistic duration; t_e – Expected duration; σ – Standard deviation; σ^2 – Variance

The duration of each activity is defined by three time estimates according to the PERT method: optimistic, most likely, and pessimistic durations, as shown in Table 1.

Next, each duration is replaced with a triangular fuzzy number, as shown in Table 2. For each activity, at least three fuzzy duration estimates are provided. A weighting coefficient is assigned to each estimate so that the sum of all weighting coefficients equals 1.

Finally, the reliable fuzzy duration for each activity is determined using the fuzzy Delphi method, as shown in Table 2. The calculation of the reliable fuzzy duration according to the Delphi method is carried out by computing the average value of the proposed estimates for each activity using formula 8 (Habibi et al., 2018).

$$Ft_{p,i}(a_{p,i.}, b_{p,i}, c_{p,i}) = \left(\frac{\sum_{j=1}^{m} w_j * a_{j,i,}}{\sum_{j=1}^{m} w_j}, \frac{\sum_{j=1}^{m} w_j * b_{j,i,}}{\sum_{j=1}^{m} w_j}, \frac{\sum_{j=1}^{m} w_j * c_{j,i,}}{\sum_{j=1}^{m} w_j}\right) \quad (8)$$

Where is:

$Ft_{p,i}$ $(a_{p,i}, b_{p,i}, c_{p,i})$	Reliable fuzzy duration of activity "i"
w_j	The weight coefficients assigned to each estimate of the fuzzy duration of activity "i"
$j= 1, ...,m$	The number of fuzzy duration estimates
Ft_j (a_j, b_j, c_j)	Fuzzy duration of activity *i* for each estimate $j = 1, ..., m$

DRAFT VERSION

Table 2: Reliable fuzzy activity duration

				Fuzzy duration			Reliable fuzzy duration Ft_i		
Activity	PA	NA	w_j	a	b	c	a	b	c
1		2,3,4	w_1= 0.2	2	4	6	3.1	5.1	6.8
			w_2= 0.5	3	5	7			
			w_3= 0.3	4	6	7			
2	1	5	w_1= 0.2	2	4	6	2.8	4.5	6.3
			w_2= 0.5	3	5	6			
			w_3= 0.3	3	4	7			
3	1	7	w_1= 0.2	8	12	16	10.2	13.9	17.3
			w_2= 0.5	10	14	18			
			w_3= 0.3	12	15	17			
4	1	6	w_1= 0.2	2	3	4	3.1	4.1	5.1
			w_2= 0.5	3	4	5			
			w_3= 0.3	4	5	6			
5	2	8	w_1= 0.2	5	7	9	5.5	8	9.5
			w_2= 0.5	6	9	10			
			w_3= 0.3	5	7	9			
6	4	7	w_1= 0.2	2	4	8	2.8	4.5	6.9
			w_2= 0.5	3	5	7			
			w_3= 0.3	3	4	6			
7	3,6	3,6	w_1= 0.2	7	10	13	8.3	11.1	12.8
			w_2= 0.5	9	11	12			
			w_3= 0.3	8	12	14			
8	5	5	w_1= 0.2	2	4	5	2.5	4.5	5.5
			w_2= 0.5	3	5	6			
			w_3= 0.3	2	4	5			
9	7,8	7,8	w_1= 0.2	9	13	15	10.3	13.8	16.9
			w_2= 0.5	11	14	17			
			w_3= 0.3	10	14	18			
10	9		w_1= 0.2	5	7	8	6.9	8.9	11.5
			w_2= 0.5	7	9	12			
			w_3= 0.3	8	10	13			

In the second phase of the experiment, it is necessary to determine the Critical Degree (CD) of each path in the network diagram and of each activity, i.e., to identify the critical path.

To determine the Critical Degree (CD) for each path in the network diagram, it is first necessary to assign reliable fuzzy durations to all activities using triangular fuzzy numbers, based on the results from the first phase of the experiment, as shown in Table 2. After that, a forward pass calculation is performed to determine the early start and early finish times for each activity, expressed as triangular fuzzy numbers. In this way, the fuzzy project duration is also determined (see Table 3).

DRAFT VERSION

Table 3: Calculation of Fuzzy Early Start and Fuzzy Early Finish of Activities

Activities	PA	NA	Reliable fuzzy duration Ft_i	Fuzzy Early Start FES_i	Fuzzy Early Finish FEF_i
1		2,3,4	(3.1; 5.1; 6.8)	(0;0;0)	(3.1; 5.1; 6.8)
2	1	5	(2.8; 4.5; 6.3)	(3.1; 5.1; 6.8)	(5.9; 9.6; 13.1)
3	1	7	(10.2; 13.9; 17.3)	(3.1; 5.1; 6.8)	(13.3; 19.0; 24.1)
4	1	6	(3.1; 4.1; 5.1)	(3.1; 5.1; 6.8)	(6.2; 9.2; 11.9)
5	2	8	(5.5; 8.0; 9.5)	(5.9; 9.6; 13.1)	(11.4; 17.6; 22.6)
6	4	7	(2.8; 4.5; 6.9)	(6.2; 9.2; 11.9)	(9.0; 13.7; 18.8)
7	3,6	9	(8.3; 11.1; 12.8)	(13.3; 19.0; 24.1)	(21.6; 30.1; 36.9)
8	5	9	(2.5; 4.5; 5.5)	(11.4; 17.6; 22.6)	(13.9; 22.1; 28.1)
9	7,8	10	(10.3; 13.8; 16.9)	(21.6; 30.1; 36.9)	(31.9; 43.9; 53.8)
10	9		(6.9; 8.9; 11.5)	(31.9; 43.9; 53.8)	(38.8; 52.8; 65.3)

$$FES_i = max[FEF(PA)] \tag{9}$$

$$FEF_i = max[FEF(PA) + Ft_i] \tag{10}$$

Where is:
FEF(PA) - Fuzzy early finish of the predecessor activity

Now it is necessary to determine the fuzzy duration for each path in the network, i.e., for each path in the network diagram, the path length is determined using a triangular fuzzy number, as shown in Table 4. Then, for each path in the network diagram, the "crisp" (deterministic) duration, i.e., the "crisp" length of each path DPL_k, is determined (Ammar & Abd-El Khalek, 2019). This is done by defuzzification using the center of area (centroid) method COA, i.e., by converting the fuzzy (uncertain) number into a deterministic (clear) number using the COA method, according to formula (6). In the same way, the "crisp" (deterministic) project duration DPD is determined, and the values are shown in Table 4.

Then, according to Ammar and Abd-El Khalek (2019), the following indicators are calculated for each path in the network diagram: Crisp path ratio – DPR_k , Intersection ratio – IR_{kp} , Critical Degree Index – CD_k

$$DPR_k = {DPL_k}/{DPD} \tag{11}$$

Where is:
DPL_k Crisp length of the k-th path in the network diagram
DPD Crisp project duration

$$IR_{kp} = \begin{cases} \frac{c_k - a_p}{c_p - a_p}, & if \quad c_k > a_p \\ 0 & if \quad c_k \le a_p \end{cases} \tag{12}$$

$$CD_k = DPR_k \times IR_{kp} \tag{13}$$

The parameters used in the formula for calculating the intersection ratio IRkp are shown in Figure 3:

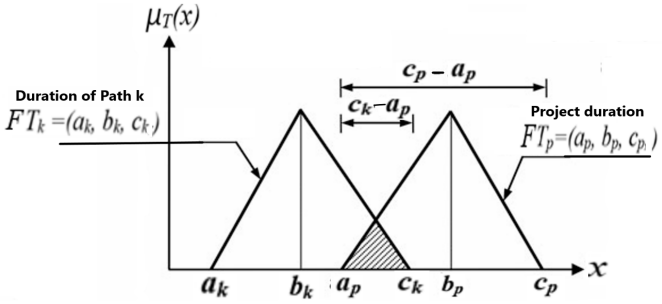


Figure 3: The parameters used in the formula for calculating the intersection ratio IR_{kp} (adapted and modified from Ammar and Abd-El Khalek, 2019)

DRAFT VERSION

4. RESEARCH RESULTS

Table 4 shows, for each path in the network diagram, the following parameters: DPLk, DPD, DPRk, IRkp, and the criticality index CDk.

Table 4: Fuzzy parameters of the project

Path in the network	FT_k	DPL_k	DP_k	IR_{kp}	CD_k
1-2-5-8-9-10	(31.3; 44.8; 56.5)	44.35	0.8316	0.6423	0.5341
1-3-7-9-10	(38.8; 52.8; 65.3)	52.42	0.9829	0.9808	0.9640
1-4-6-7-9-10	(34.5; 47.5; 60.0)	47.37	0.8882	0.7769	0.6900
Project duration	(39.8; 53.8; 65.8)	53.33			

Based on the data from Table 4, it can be concluded that the path 1-3-7-9-10 is most likely critical, as it has the highest criticality index value of CDk = 0.9640. The other paths have lower criticality index values CDk and are most likely non-critical paths.

5. CONCLUSION

The implementation of any project, from conception to completion, takes place under conditions of varying degrees of uncertainty. Fuzzy logic addresses imprecision or ambiguity in estimation and may be preferable to probability theory in situations where historical data on activity durations are unavailable or unreliable. For this reason, the use of fuzzy set theory is proposed for modeling uncertainty, as an alternative to traditional probabilistic methods. This paper presents a novel methodology for assessing the degree of criticality of project paths and activities within a fuzzy logic framework. In the proposed methodology, the relative degree of criticality and the agreement index are integrated into a unified indicator to provide a more precise and comprehensive measure of criticality. The research results have shown that the fuzzy adaptive CPM-PERT method can be successfully applied in project planning. It is particularly suitable for use in the construction industry, due to the high degree of subjectivity in estimating activity durations and costs..

LITERATURE

Ammar, M.A., & Abd-Elkhalek, S.I. (2019). Criticality measurement in fuzzy project scheduling, *International Journal of Construction Management, Vol 22, Issue 2*, pp. 252-261 https://doi.org/10.1080/15623599.2019.1619226

Bakry, I., Moselhi, O., & Zayed, T., (2013). Fuzzy scheduling of repetitive construction projects, *Proceedings of the 30th ISARC*, Montreal, Canada, pp. 395–403, https://doi.org/10.22260/ISARC2013/0043

Dubois, D., & Prade, H., (1978). Operations on fuzzy numbers, *International Journal of Systems Sciences, Vol. 9, no. 6,* p. 613-626. https://doi.org/10.1080/00207727808941724

Habibi, F., Taghipour Birgani, O. , Koppelaar, H., & Radenović, S., (2018). Using fuzzy logic to improve the project time and cost estimation based on Project Evaluation and Review Technique (PERT), *Journal of Project Management Vol3(4)*, pp. 183-196. DOI:10.5267/j.jpm.2018.4.002

Liberatore, M. (2016). A counterexample to the forward recursion in fuzzy critical path analysis under discrete fuzzy sets, *International Journal of Fuzzy Logic Systems (IJFLS), Vol. 6(2)*, pp. 53–62. DOI:10.5121/ijfls.2016.6204

Moselhi, A.S.O, (2015). Contingency modelling for construction projects using fuzzyset theory, *Engineering, Construction and Architectural Management, Vol. 22 Iss 2,* pp. 214 – 241 https://doi.org/10.1108/ECAM-03-2014-0039

Shaheen, A., Robinson Fayek, A., & Abourizk, S.M. (2007). Fuzzy Numbers in Cost Range Estimating, *Journal of Construction Engineering and Management, Volume 133 Issue 4* https://doi.org/10.1061/(ASCE)0733-9364(2007)133:4(32

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

THE USE OF DESCRIPTIVE STATISTICS IN ANALYZING THE PERCEPTION OF OCCUPATIONAL RISKS: THE CASE OF SILICA DUST EXPOSURE AMONG CONSTRUCTION WORKERS

ŽELJANA KUŽET[1], VLADIMIR MUČENSKI[2], JOVANA TOPALIĆ[3]

[1] Department of Environmental Protection and Occupational Safety Engineering, Faculty of Technical Sciences, University of Novi Sad, Republic of Serbia, Scholarship holder of the Ministry of Science, Technological Development and Innovation, zeljana.kuzet.12@outlook.com, 0009-0004-0958-517X

[2] Department of Civil Engineering and Geodesy, Faculty of Technical Sciences, University of Novi Sad, Republic of Serbia, mucenskiv@uns.ac.rs , 0000-0001-9830-4747

[3] Department of Civil Engineering and Geodesy, Faculty of Technical Sciences, Novi Sad, Republic of Serbia, jovanatopalic90@uns.ac.rs 0000-0001-7854-3257

Abstract: *Silica dust is present on the construction sites during digging, demolition, and cutting of wood and tiles. With no smell and no color, silica dust can cause various diseases, where silicosis represents the most common disease, with no cure. Preventive measures and medical exams are essential to reduce the risk of silica. The aim of this research is to assess how construction workers perceive the level of awareness, risk perception, and use of PPE related to occupational exposure to silica dust through the use of survey and descriptive statistical analysis. This study surveyed 55 male construction workers in Novi Sad to assess their awareness of risks related to silica dust exposure. The outcome of the study revealed that approximately 50% of employees described their understanding as minimal while the majority did not know about silicosis health risks. The majority of participants confirmed that they used safety masks but a significant portion reported using unsuitable protection and experiencing discomfort from this safety measure. According to the study results more than 60% of workers did not receive training from their employers regarding silica dust dangers. The results indicate that enhanced education alongside better personal protective equipment implementation would boost workplace safety standards.*

Keywords: *construction workers, silicate dust, survey, descriptive statistics*

1. INTRODUCTION

Over 3 million construction workers are exposed to silica dust in Europe, and in the United States of America, that number is 1.7 million construction workers (Sauvé et al., 2013). The construction industry represents one of the primary sources of air pollution (Cheriyan & Choi, 2020), where the dust that is created on construction sites is usually of natural origin, but still dangerous for the respiratory system. One of the most common types of dust present on construction sites is silicate dust (Mučenski, 2018). Exposure to dust on construction sites occurs during excavation, drilling, rock processing, tunneling, demolition of buildings, cutting and processing of concrete, brick, ceramic, and granite tiles, etc. (HSE UK, 2020). Silicate dust is a complex mixture of particles that vary in shape, size, and composition. The crystalline silicon is the most common form associated with the development of silicosis and other related diseases (Alabi & Bakare, 2011). The construction site is naturally loaded with dust particles, but the problem is when the free crystalline silica appears in the polluted air of a construction site. Once inhaled, it can pose various diseases such as silicosis, chronic obstructive pulmonary disease, and lung cancer (Keramydas et al., 2020).

Silicosis is a lung disease caused by inhalation of dust containing particles of silicon dioxide, which are defined as "respirable", that is, the given form of silicon is called "respirable crystalline silicon" (RCS). Exposed workers may suffer not only from silicosis but also from certain cardiovascular diseases, sarcoidosis, pulmonary tuberculosis, lung infections, chronic obstructive disease, certain types of tumors, lung cancer, autoimmune diseases, and kidney disorders (Kreuzer et al., 2013). Various studies (Pérez-Alonso et al., 2014; Hoy et al., 2018) indicate that the occurrence of silicosis is directly related to the availability and proper use of personal protective equipment (PPE). Since the exposure to silica dust is continuously increasing, so is the

DRAFT VERSION

risk of mortality and increased mortality, making silica dust one of the major health problems worldwide (Requena-Mullor et al., 2021).

Although there is no adequate cure, preventive measures and a medical examination should be implemented (Requena-Mullor et al., 2021; OSHA, 2012). Silicosis, an incurable disease, requires quantitative and qualitative control of respirable crystalline silica and the development of appropriate control measures (Tavakol et al., 2016). HSE UK (2020) is proposing that when construction workers are exposed to chemical substances, in this case – silicate dust, there should be done risk assessment, risk control, and a review of the control measures. As the main preventive measures against silica dust, Mučenski (2018) proposes water curtains, extraction on the tool, organization of work tasks, ventilation, proper selection of PPE, and conducting health surveillance of exposed workers.

The aim of this research is to assess how construction workers perceive the level of awareness, risk perception, and use of PPE related to occupational exposure to silica dust through the use of survey and descriptive statistical analysis.

2. METHODOLOGY

With the aim of better understanding worker awareness of risks associated with silicate dust, quantitative research was conducted among construction workers in the form of a survey. The research was realized with an online survey that was created on the platform Google Forms, with full anonymity of respondents. The survey was composed of closed and semi-open questions, with a total number of 15 questions grouped into four sections:

1. Sociodemographic data – structured at the beginning of the questionnaire, with the aim of providing the segmentation of respondents according to gender, age, work experience in the construction industry and workplace on the construction site;
2. Awareness of silica dust – this section included questions about familiarity with the term silica dust, the effects that silica dust has on health, and to assess the risk of exposure to silica dust in the workplace;
3. The use of personal protective equipment (PPE) – this section included questions about provided PPE, use of PPE, types of PPE masks against dust, reasons for not wearing PPE
4. Health and safety (HSE) training, awareness and consequences of exposure to silica dust – information about the harmful effects of silica dust as a part of HSE training, information channels about silica dust, preventive measures against silica dust and adequacy of applied preventive measures against silica dust on construction site.

Most of the questions were in form with only one choice of answer, while some of the questions enabled multiple selections (checkbox type) or option "Other" for additional comments. Due to the impossibility of filling out the survey in online mode, the survey was printed and respondents were filling the survey manually on construction sites. After collecting the completed surveys, the results were uploaded using Google Forms. The target group consisted of construction workers and the data collection was done voluntarily. All of the construction workers were informed that the survey was anonymous and that collected data would be used only for the purpose of scientific research. The length of the survey was a couple of minutes.

Collected data is statistically processed and analyzed using descriptive statistics in order to analyze sociodemographic data, and identify patterns in risk perception, level of information, and frequency of use of PPE.

When analyzing the sociodemographic data, statistical analysis was conducted on two continuous variables in this study (age and work experience) to determine their arithmetic mean, median, and mode values. The standard statistical formulas were used for grouped data to calculate these measures because the data was presented in interval groups which required the use of class midpoints. The equations for arithmetic mean, median, and mode values are shown as follows. The arithmetic mean is calculated using the following equation (1):

$$\bar{x} = \frac{\sum(fi*xi)}{\sum fi}, \tag{1}$$

where f_i represents the frequency of the i-class, x_i represents the midpoint of the i-class and $\sum f_i$ represents the total number of observations.

The median class was calculated based on the cumulative frequency distribution using the equation (2):

DRAFT VERSION

$$Median = L + \left(\frac{\frac{N}{2} - F}{fm}\right) * h, \quad (2)$$

where L represents the lower boundary of the median class, N represents the total number of observations, F represents cumulative frequency before the median class, f_m represents the frequency of the median class, and h width of the class interval.

The mode values were approximated as the midpoint of the most frequent class, using the equation (3):

$$Mode = midpoint\ of\ modal\ class \quad (3)$$

Patterns in risk perception, level of information, and frequency of use of PPE were displayed graphically by using bar graphs and pie charts. The applied method enabled clearer interpretation and understanding of collected results.

3. RESULTS AND DISCUSSION

A survey about construction workers' awareness of the risks associated with silica dust was conducted on 55 male construction workers in Novi Sad. A survey about workers' awareness of the risks associated with silica dust was conducted on 55 male construction workers, where:

- 29.1% of workers were between 35-44 years old,
- 21.8% were between 45-54 years old,
- 21.8% were between 55-64 years old,
- 12.7% were between 18-24 years old,
- 9.1% were between 25-34 years old,
- 3.6% were 65 years old or older and
- only 1% were underage.

When talking about the working experience in the construction industry, 21.8% of respondents have been working between 16-20 years, 20% of respondents have been working between 6-10 years, 20% of respondents have been working between 1-5 years, 16.4% of respondents have been working between 11-15 years, 10.9% of respondents have been working between 21-30 years, while only 5.5% of respondents have been working for more than 31 years, and 5.5% of respondents have been working less than a year.

Based on collected data, the central tendencies – the mean, median, and mode were estimated for all analyzed continuous variables (age and work experience) in Table 1. The data shows that construction workers have an average age of 43.4 years and a median of 39 years thus indicating an older population on construction sites. Most participants fall within the 35-44 age range which places them in middle adulthood. The data indicates that participants on average have 13 years of work experience but the median value of 8 years demonstrates that fewer highly experienced respondents play a significant role in the average. The two main professional tenures of 16-20 years in the data represent the existence of two separate groups within the sample population for analysis. The sample's heterogeneity in these findings should be considered when conducting additional data analysis.

Table 1: Descriptive Statistics – Central Tendencies of Age and Work Experience Among Construction Workers (n=55)

Variable	Arithmetic mean	Median	Mode	Additional notes
Age (years)	43.4	39	35-44	Based on grouped responses (n=55)
Work experience (years)	13.0	8	16-20	

The target group was construction workers exposed to silicate dust, so the largest representation is the construction workers who perform preparatory work as the main job on the construction site (47.3%), in addition, there are also other jobs such as painting and decorating works (23.6%), masonry works (9.1%), carpentry works (7.3%), earthworks (1.8%), concrete works (3.6%), reinforcement works (1.8%), ceramic works (1.8%), operating machines and equipment (1.8%), manual worker (1.8%).

Analyzing the section about construction workers' awareness of silica dust, the survey indicated some interesting facts. The construction workers had an assessment to evaluate their own knowledge about the definition of silica dust and its effects on human health from 1 to 5, where 1 was defined as completely unfamiliar, and 5 as very well-informed. The survey showed that most of the construction workers (49.1%) rated their knowledge with 2, which indicates a low level of information about this type of professional risk. On the other side, a small percentage of construction workers evaluated their knowledge with 4 (16.4%) and 5 (1.8%) showing that the percentage of construction workers who are aware of the risks and health consequences that silica dust can induce are negligible (Figure 1). When talking about the consequences that silica dust can induce on human health during long-term exposure to silicate dust, several different answers

DRAFT VERSION

were offered, some of which were incorrect, while some truly represented the consequences of exposure to silica dust, to see how familiar construction workers were. This question enabled multiple answers. Results showed that most of the construction workers are not informed (56.4%), while only 5.5% of construction workers are informed of silicosis as a health consequence (Figure 1). On the other hand, construction workers showed that they are familiar with the basic consequences after long-term exposure to silica dust because 10.9% of respondents stated bronchitis and 5.5% of respondents stated lung cancer as a source of consequence when being exposed to silica dust for a long time. By indicating skin allergies, high blood pressure, and headaches as a source of consequences during long-term exposure to silica dust, it was proven that familiarity with health consequences is low, and it should be improved by adapted HSE medicine lectures with the support of occupational medicine, all with the aim of increasing the awareness and knowledge of workers.

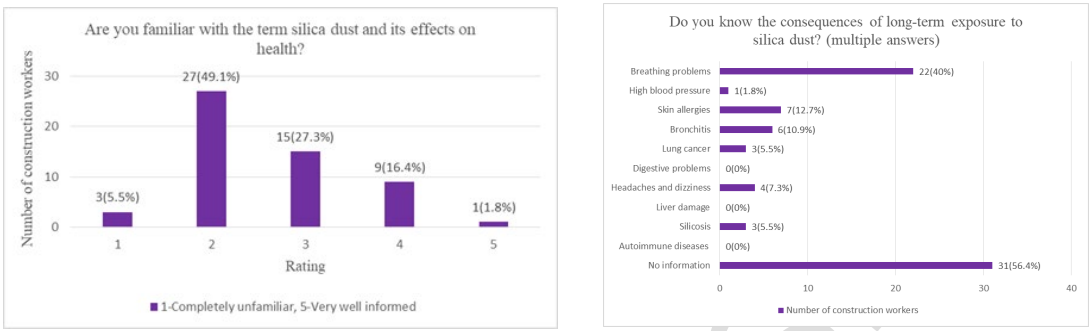


Figure 1: Familiarity with the term silica dust and its effects on health among construction workers (left) and the knowledge about consequences of long-term exposure to silica dust among construction workers (right)

At the final of this section, construction workers were asked to assess the risk of exposure to silica dust at their workplace (1-Very low risk, 5-Very high risk). Results showed that 38.2% of respondents assessed the risk as moderate (3), 23.6% of respondents estimated the risk as 2, and 23.6% with 4. A low number of respondents believe that the risk is very low (9.1%), while only 5.5% believe that the risk is very high (5). These results show that most of the construction workers perceive risk as present, but not extremely high, which can be due to lack of information during HSE training, lack of dust measurements, and lack of visually observable effects of exposure. Considering that the level of information and perception of silica dust as a risk represents an important aspect when assessing the awareness of construction workers, it is also important to assess their relation with the use of PPE. Effective health and safety system also depends on whether the protective equipment is provided by the employer. 38.2% of respondents stated that the PPE against silica dust is always provided and 61.8% of respondents stated that PPE against silica dust is provided only sometimes. In this section, the two questions drew attention to the sincerity of construction workers. Firstly, on the question: "Do you use a protective mask during work in a dusty environment?" most of the respondents (74.5%) said that they always use a protective mask when working in a dusty environment, and 25.5% of respondents stated that they use the protective mask sometimes when working in a dusty environment. On the other side, the question "If you don't use a protective mask, what is the main reason for not wearing?" somehow revealed the real truth about the use of protective masks. As the main reason for not wearing a protective mask during work in a dusty environment, 36.4% of respondents revealed that the mask bothers them during the work (Figure 2)

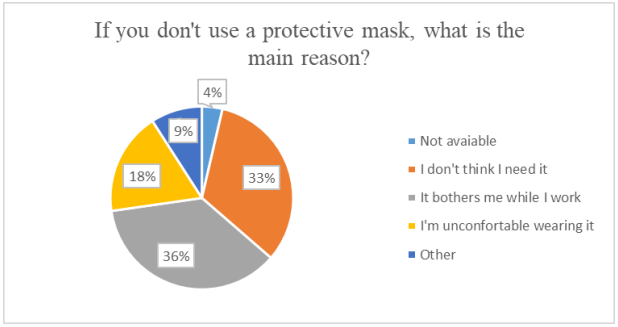


Figure 2: The main reason for not using a protection mask among construction workers

Also, when talking about the type of personal protective mask that construction workers use when working in dusty environments, most of the workers stated that they use protective mask with basic protection from dust (FFP1), and a small percentage of respondents said that they are using the protective masks that provides better protection from dust and pollution. An interesting fact brought the attention when analyzing the collected

DRAFT VERSION

responses of the survey this question - 32.7% of respondents said that they are using ordinary surgical masks when working in dusty environments and 25.5% of respondents said that they don't know exactly what type of personal protective mask are using. In conclusion to this results, it is obvious that there is a gap between the use of PPE and the real awareness of its importance and correct application. Although most workers state that they are using PPE on a regular basis, additional questions reveal that inadequate types of masks are often used or that construction workers are not sure what equipment they are using. A significant percentage of construction workers complain about masks (bothering during work operations), which points to the need for ergonomic and functional adaption of PPE. Therefore, workers are not irresponsible, but adequate information, equipment, and support are often not available for them. On the other side, employers and the HSE team should recognize the problems and act proactively (HSE training, availability of equipment).

The last section of the survey discovered the level of information among the construction workers across the formal HSE training, as well as the sources of information, the perception of existing preventive measures, and their opinions on the necessary preventive measures. When talking about the formal HSE training the survey showed that 63.6% of respondents stated that they were not informed about the harmful effects of silicate dust and protection measures against silica dust, and 36.4% of respondents stated that they were informed about the harmful effects of silica dust through HSE training. The survey showed that two main sources of information about hazards connected to silica dust are employers (through obligatory HSE training) and colleagues (Figure 3). The perception of existing preventive measures is shown in Figure 3. It is obvious that construction workers are aware that one of the existing preventive measures – The use of personal protective masks (FFP1, FFP2, FFP3) is necessary when working in a dusty environment (63.6%). Concerning fact is that 29.1% of respondents consider that surgical masks should be used as a preventive measure against silica dust.

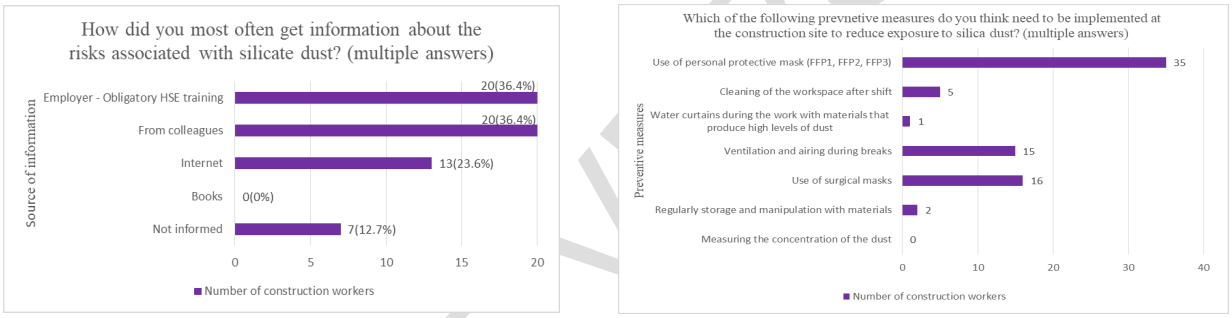


Figure 3: Information sources about the risks associated with silicate dust among construction workers (left) and preventive measures for silica dust proposed by construction workers (right)

At the end of the survey, construction workers were asked to evaluate the adequacy of preventive measures against silica dust on the construction site where they work with a scale from 1 to 5, where 1 means that preventive measures against silica dust are inadequate and 5 means that preventive measures against silica dust are adequate. Most of the respondents choose average values (3) as a neutral or medium rating, 27.3% with rate 4, 21.8% with rate 2, and the very low and very high rates are rare – rate 1 was given by only 1.8% of respondents, and rate 5 only 5.5% of respondents. This indicates that construction workers mostly think of preventive measures as something that is not completely inadequate, but also not completely satisfactory either. Most of the construction workers express moderate satisfaction, which means that there is a basic level of protection, but additional improvements are needed when applying preventive measures against silica dust.

4. CONCLUSION

The survey conducted among 55 male construction workers in Novi Sad demonstrates an insufficient understanding of silica dust health risks from prolonged exposure. Despite the fact that the surveyed population mostly contains middle-aged workers who possess extensive construction experience, they demonstrate limited comprehension of silica dust and their health impacts. The study reveals that only a small percentage of construction workers identified silicosis, bronchitis or lung cancer. Also, study showed that construction workers demonstrate a strong tendency to use PPE but their practice doesn't align with the actual level of provided protection. Workers either wear surgical masks or don't understand the proper kind of mask. Worker education about silica dust remains low due to inadequate program of HSE training, and most workers depend on co-workers to learn about the topic. Preventive measures are perceived as moderate, but the use of inadequate PPE and a general lack of risk perception highlight the necessity for better safety practices and improvement.

DRAFT VERSION

This research demonstrates an immediate requirement for: (1) specialized awareness initiatives and educational initiatives that address the particular dangers from silica dust, (2) both the accessibility and practicality of PPE, (3) strengthened the participation of employers and HSE teams in their ability to notice and resolve worker issues, (4) Monitoring and measuring dust level measurements and occupational health monitoring of workers who are exposed to silica dust, (5) Improving the worker education, attitudes and practices regarding silica dust exposure.

ACKNOWLEDGMENT

This research has been supported by 'Karin Komerc' LLC, Novi Sad, Republic of Serbia, which provided access to the construction site of 'VELEBIT Gradnja' LLC, Novi Sad, Republic of Serbia, whose management and workers kindly agreed to participate in the study, and 'GAT' LLC, Novi Sad, Republic of Serbia whose management and workers also kindly agreed to participate in the study. This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451- 03-137/2025-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad 2025" (No. 01-50/295). (V.M.) This research has been supported by the Ministry of Science, Technological Development and Innovation through Contract No. 451-03-136/2025-03/200156. (J.T)

LITERATURE

Alabi OA, Bakare AA. (2011). Genotoxicity and mutagenicity of electronic waste leachates using animal bioassays. *Environ Toxicol Chem*.; 93(5):1073–1088 ps.

Cheriyan, D., & Choi, J. H. (2020). A review of research on particulate matter pollution in the construction industry. *Journal of Cleaner Production*, *254*, 120077.

Hoy, R.F.; Baird, T.; Hammerschlag, G.; Hart, D.; Johnson, A.R.; King, P.; Putt, M.; Yates, D.H. (2018) Artificial stone-associated silicosis: A rapidly emerging occupational lung disease. Occup. Environ. Med., 75, 3–5.

HSE UK (2020). Construction dust, Construction Information Sheet No 36 (Revision 3), https://www.hse.gov.uk/pubns/cis36.pdf

Keramydas, D., Bakakos, P., Alchanatis, M., Papalexis, P., Konstantakopoulos, I., Tavernaraki, K., ... & Lazaris, A. C. (2020). Investigation of the health effects on workers exposed to respirable crystalline silica during outdoor and underground construction projects. Experimental and Therapeutic Medicine, 20(2), 882- 889.

Kreuzer, M.; Sogl, M.; Brüske, I.; Möhner, M.; Nowak, D.; Schnelzer, M.; Walsh, L. (2013). Silica dust, radon and death from non-malignant respiratory diseases in German uranium miners. Occup. Environ. Med., 70, 869–875.

Mučenski V. (2018). Teorijske osnove bezbednosti i zdravlja na radu u građevinarstvu (Theoretical foundations of safety and health at work in construction), Faculty of Technical Sciences, University of Novi Sad, pp. 39 – 41. (in Serbian), ISBN: 978-86- 6022-049-5

OSHA, (2012). Health Hazards in Construction Workbook, https://www.osha.gov/sites/default/files/2019-03/health_hazards_construction_pt2.pptx

Pérez-Alonso, A.; Córdoba-Doña, J.A.; Millares-Lorenzo, J.L.; Figueroa-Murillo, E.; GarcíaVadillo, C.; Romero-Morillos, J. (2014). Outbreak of silicosis in Spanish quartz conglomerate workers. Int. J. Occup. Environ. Health, 20, 26–32.

Requena-Mullor, M., Alarcón-Rodríguez, R., Parrón-Carreño, T., Martínez-López, J. J., Lozano-Paniagua, D., & Hernández, A. F. (2021). Association between crystalline silica dust exposure and silicosis development in artificial stone workers. International Journal of Environmental Research and Public Health, 18(11), 5625.

Sauvé JF, Beaudry C, Bégin D, Dion C, Gérin M and Lavoué J (2013). Silica exposure during construction activities: statistical modeling of task-based measurements from the literature. Ann Occup Hyg 57: 432-443,

Tavakol E, Azari MR, Salehpour S and Khodakarim S (2016). Determination of Construction Workers' Exposure to Respirable Crystalline Silica and Respirable Dust. J Saf Promot Inj Prev 3: 263-270

DRAFT VERSION

PERFORMANCE ANALYTICS

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

ADVANTAGES OF THE SUPER EFFICIENCY SBM DEA MODEL

JELENA NOVAKOVIĆ[1]

[1] University of Belgrade, Faculty of Organizational Sciences, Belgrade, jekanovakovic@gmail.com , 0009-0003-7863-5399

Abstract: *This paper aims to present the key advantages of the super efficiency Slacks-Based Measure (SBM) Data Envelopment Analysis (DEA) model. The Super Efficiency SBM DEA model extends the conventional SBM framework by enabling efficiency scores to exceed unity, thus allowing for the full ranking of all decision-making units (DMUs), including those deemed efficient under traditional DEA models. By incorporating input and output slacks explicitly into the efficiency measurement, the Super Efficiency SBM model addresses the issue of multiple DMUs achieving an efficiency score of one, which limits discriminative power in benchmarking analyses. This model operates in a non-radial, non-angular manner, effectively capturing inefficiencies arising from both proportional reductions and slack adjustments in inputs and outputs. Moreover, the Super Efficiency SBM model accommodates undesirable outputs and maintains strong properties such as unit invariance and translation invariance, making it robust in various practical applications. The enhanced discriminatory capability and ability to rank efficient DMUs uniquely provide critical insights for performance evaluation, resource allocation, and managerial decision-making in complex operational environments.*

Keywords: *Super Efficiency, Slacks-Based Measure (SBM), Data Envelopment Analysis (DEA), Decision-Making Units (DMUs), Input and output slacks, Performance evaluation, Managerial decision-making.*

1. INTRODUCTION

Data Envelopment Analysis (DEA), originally developed by Charnes and co-authors, is a nonparametric mathematical technique for measuring and evaluating the relative efficiency of a set of decision-making units (DMUs) with multiple inputs and outputs (Charnes et al, 1978). Although DEA provides a valuable approach for assessing performance and measuring efficiency, it also has its limitations. The disadvantage of the previously mentioned DEA models is that they cannot distinguish between efficient DMUs, all of which have an efficiency score of unity. Therefore, it is necessary to improve the DEA model in order to expand its application. Andersen and Petersen (Andersen & Petersen , 1993) proposed a super-efficient DEA model which can rank efficient decision units. Also Banker and co-authors (Banker et al., 1989) in their paper proposed a model that overcomes the limitations. The input-oriented (output-oriented) super-efficiency DEA model eliminates the DMU under evaluation from the reference set. Then, that efficient DMUs may have efficiency grades larger (smaller) than or equal to one. The authentic super-efficiency DEA model is introduced under the condition of constant returns to scale (CRS) and is achievable if all inputs and outputs of DMUs are positive. But, the infeasibility question might appear in VRS super-efficiency DEA models (Seiford and Zhu, 1999).

There are several more weaknesses of traditional DEA models, among other things is that they consider that all the inputs and outputs are non-negative values. In many examples in practice, especially for outputs, values can be negative, for example expected return for assessing the efficiency of an investment fund, profit for measuring the efficiency of a project etc (Lin & Chen, 2022).

More than forty years, the DEA method has been extensively used in measuring the efficiency in different fields. Some authors have used models that are closely related to the basic DEA models to assess the efficiency of bank management (Li, 2020; Lin, Lee, & Chiu, 2009; Matthew, 2013; Mercan, Reisman, Yolalan, & Emeld, 2003; Schaffnit, Rosen, & Paradi, 1997; Shao, You, Xu, & Shao, 2020). Recommended super-efficiency models can be found in numerous significant papers (Chen, 2005; Chen & Liang, 2011; Cook et al., 2009; Lee et al., 2011; Lee & Zhu, 2012; Lovell & Rouse, 2003; Seiford & Zhu, 1999).

The integration of super efficiency SBM-DEA or DEA and neural networks can contribute improved tools to overcome the previously mentioned problems (Zhong et al., 2021; Alizadeh & Omrani, 2019;

DRAFT VERSION

Mostafa, 2009). Some autors also have researching the fusion of DEA and neural networks (Santin et al., 2004; Chuang et al., 2011; Bose & Patel, 2015; Kwon & Lee, 2015; Kwon et al., 2016; Tsolas et al., 2020). A significant contribution of the SBM (Slack-Based Measure) super-efficiency model to the analysis of efficiency and productivity can be found in the numeros papers (Lee, 2021; Xiao-ning et al., 2013; Tian et al., 2020, Bolos et al., 2023, Du et al., 2010).

The radial DEA model based on the CCR cannot fully consider the effect of slackness on efficiency. To solve this problem, Tone (2001) proposed the SBM model (Tone, 2001 ; 2004) and super-efficiency SBM model (Tone, 2002). Based on slackness measurement, SBM is a non-radial method suitable for measuring efficiency when input and output may vary in a non-proportional manner (Hong, Fang, Wei, & Hong, 2013). The super-efficiency SBM model is a method that combines super efficiency and the SBM model. The basic idea of the super-efficiency evaluation method is to remove the effective evaluation unit from the set and re-evaluate; thus, the original non-effective value evaluation remains unchanged, and the original effective value evaluation can be greater than 1, then it can be compared.

2. SUPER-EFFICIENCY DEA MODELS

Estimating super efficiency assumes a modification of the DEA model in such a way that efficient units can be assigned an index greater than 1 and thus allow discrimination between them. They allow ranking among efficient DMUs by measuring how far a DMU can deviate from the efficient frontier and still remain efficient. Common super-efficiency DEA models are:

- **Andersen–Petersen (AP) Model (1993)** - This is the most popular super-efficiency model. It removes the DMU under evaluation from the reference set and solves a standard DEA model, and allows scores greater than 1 for efficient DMUs. This model can be used in both input-oriented and output-oriented contexts.
- **Super-SBM (Slack-Based Measure) -** Developed by Tone (Tone, 2002). Non-radial and non-oriented: considers input/output slacks directly. Offers more discrimination than radial models like AP. Particularly useful when inputs/outputs have slacks that radial models ignore.
- **Directional Distance Function (DDF) Super-Efficiency** - Allows simultaneous contraction of inputs and expansion of outputs in a specific direction. More flexible and adaptable to different decision-making settings.

Super-efficiency models, while useful in performance analysis, can have some issues. Specifically, under variable returns to scale (VRS), infeasibility can occur, especially when trying to exclude the evaluated DMU from the reference set, and this infeasibility can be a major obstacle to their application. VRS constraints include a convexity condition (sum of lambdas = 1). When the efficient DMU is removed from the reference set, it may no longer be possible to construct a feasible convex combination of the remaining DMUs to dominate the input-output vector of the evaluated DMU. This leads to infeasibility, especially in small or sparse datasets and highly efficient DMUs (that form the frontier themselves).

Solutions to address infeasibility in super-efficiency under VRS:

- **Use of Modified Super-Efficiency Models.** Slack-Based Super-Efficiency Models incorporate input and output slacks explicitly and may relax the $\sum\lambda = 1$ constraint. For Example: Slack-Based Measure (SBM) super-efficiency models proposed by Tone (Tone, 2002) avoid infeasibility more robustly. Non-Radial or Directional Distance Function (DDF) Models evaluate, instead of radial expansion/contraction, efficiency in a given direction, which can often avoid infeasibility. DDF super-efficiency models typically modify the reference set structure.,
- **Relaxing the Convexity Constraint.** Transition from VRS to CRS, because in CRS models, $\sum\lambda$ is not constrained to 1, which allows the reference point to lie on the ray from the origin. This change avoids infeasibility but changes the scale assumption. This is useful for robustness checks or when VRS leads to frequent infeasibility. On the other side, it's possible to impose Bounded Returns to Scale (BRS), in way that adds constraints like: $\alpha \leq \sum\lambda \leq \beta$. This introduces flexibility around the convexity constraint while maintaining partial VRS assumptions.,
- **Reinsertion or Modification of DMU_0 in Reference Set.** Instead of completely removing DMU_0, it possible to keep with a small weight ($\lambda_0 \neq 0$). This relaxes the impact of exclusion and may preserve feasibility. Also, it possible slightly perturb the inputs/outputs of DMU_0 so it's technically excluded but doesn't cause infeasibility. This trick avoids the issue of perfect self-dominance without violating model structure.,
- **Augmenting the Dataset.** By add synthetic or artificial DMUs that maintain the shape of the frontier and allow convex combinations. This is especially useful in sparse datasets.,

DRAFT VERSION

- **Alternative Efficiency Ranking Methods.** When infeasibility persists, consider alternative ranking methods: Cross-efficiency evaluation, Bootstrap DEA, Malmquist index-based comparisons (for panel data).

Super efficiency SBM models are foundational to several advanced DEA techniques, including: Network DEA, Malmquist productivity indexes, Dynamic and Window DEA models,Stochastic and fuzzy DEA models.

3. SUPER EFFICIENCY SBM DEA MODEL

Tone incorporated the idea of the super-efficiency proposed by Andersen and Petersen into SBM, and based on the SBM model, suggested the super-efficiency measure, which is referred to as Super-SBM (Tone, 2002). The basic idea is the similar as in the Andersen-Petersen DEA model. The unit being evaluated is excluded from the production set that forms the efficiency frontier. After this unit excluded, a DMU unit with inputs $x_i^*(x_i^* \geq x_{ik},\ i=1,...,m)$ and outputs $y_r^*(y_r^* \leq y_{rk},\ r=1,...,s)$ is required, which will be SBM efficient, that is, it will have an efficiency index of 1 (model M 1)., as follows:

MODEL (M1)

$$(min)\varphi = \frac{\frac{1}{m}\sum_{i=1}^{m}\frac{x^*_i}{x_{ik}}}{\frac{1}{s}\sum_{r=1}^{s}\frac{y^*_r}{y_{rk}}} \quad (1)$$

s.t.

$$\sum_{\substack{j=1 \\ j\neq k}}^{n} x_{ij}\lambda_j + s_i^- = x_{ik} \qquad i=1,...,m \quad (2)$$

$$\sum_{\substack{j=1 \\ j\neq k}}^{n} y_{rj}\lambda_j + s_r^+ = y_{rk} \qquad r=1,...,s \quad (3)$$

$$x_i^* \geq x_{ik} \qquad r=1,...,s \quad (4)$$

$$x_i^* \geq x_{ik} \qquad r=1,...,s \quad (5)$$

where (1) denotes the serial number of the objective function.

The presented model provides an efficiency index greater than or equal to 1, simultaneously taking into account the input and output values and their distance from the reference point at the efficiency limit. The counter in the objective function (1) shows the rate of possible average increase of all inputs and tells how much it is possible to increase each of the inputs (x_i^*-x_{ik}, $i=1,...,m$) individually, while the super-efficient unit DMU*k* still remains efficient. On the other side, the denominator in the objective function shows the rate of possible average reduction of all outputs and tells about how much it is possible to reduce each of the outputs (y_r^*-y_{rk}, $r=1,...,s$) individually, while the super-efficient unit DMU*k* still remains efficient. The problem with this way of evaluating super-efficiency is that each inefficient DMU will be assigned an efficiency index equal to 1, which makes it impossible to discriminate them.

4. PURPOSE OF THE SUPER EFFICIENCY SBM DEA MODEL

Slacks based measure (SBM) is a non-radial model to solve the problem in the additive model developed by Charnes, Cooper and Rhodes in 1978. This model can discriminate between efficient and inefficient DMU. Still it does not provide an efficiency measure, so a decision maker can't clarify the DMU performance. Accordingly super SBM model was proposed to established and rank efficient DMU. The efficiency scores are between 0 and 1, thus permitting the comparison of different DMU efficiencies.

SBM model uses “slacks” to present the excess amount of inputs and the lack of outputs then it directly deals with them by maximizing these slacks. SBM analysis provides an efficiency score which is units-invariant and a monotone function of input slacks and output slacks. Tone (2001), in aim to rank efficiency, was introduced super SBM. It is appropriate for evaluating efficiencies when inputs and outputs may change non-proportionally.

Super SBM model can be expanded by the super efficiency to the DEA models. Super efficiency rate refers to the distance between the inputs and outputs of both units. The distance is shown in variable ρ. Assuming that there exist a set of n DMU producing the same set of outputs which consume the same set of inputs.

DRAFT VERSION

The advantages of SBM for measuring super-efficiency are numerous. The SBM model directly incorporates slacks (input excesses and output shortfalls) into the efficiency measure, while Conventional DEA models (e.g., CCR, BCC) might classify a DMU as efficient even when it has input/output slacks. This leads to a more accurate and discriminating evaluation, especially important in super-efficiency analysis. Unlike radial models, which scale inputs or outputs proportionally, the SBM model allows for non-proportional adjustments. This flexibility gives a more realistic and detailed measure of efficiency, useful when proportional changes are impractical.

The super-efficiency SBM model can rank efficient DMUs by allowing their efficiency score to exceed 1. This helps in differentiating among efficient units, which is crucial for benchmarking or decision-making. Another advantage is Invariance to Units of Measurement. Because the SBM model is based on slacks, it is unit invariant—meaning changes in units (e.g., kg to tons) do not affect the efficiency score. By considering slack and allowing for super-efficiency scores, the SBM model provides better discrimination power among high-performing DMUs compared to radial models. Also, the SBM framework has been extended to handle zero or negative data, making it applicable to a broader range of real-world problems (with appropriate modifications). Theoretical Rigor and Interpretability are also advantages. The model has solid theoretical foundations, and the efficiency scores are easily interpretable due to the explicit use of slacks.

5. CONCLUSION

Since the appearance of the first DEA models, data envelopment analysis has flourished in its theory and application. Although DEA is widely used to quantitatively evaluate business performance, it is difficult to meet actual improvement needs. This is mainly due to the limitations of the model. The Super-SBM extends the Slack-Based Measure (SBM) to allow scores greater than 1, enabling a meaningful ranking of efficient units. It enhances discrimination among top-performing units, which is critical in competitive benchmarking, performance rewards, or resource allocation. Especially valuable when there are many efficient DMUs. Unlike traditional radial models (e.g., CCR, BCC), the SBM model includes input and output slacks directly into the efficiency score. This leads to a more accurate and nuanced measure of efficiency, particularly when slacks exist even for technically efficient units. The Super-SBM is non-radial (not based solely on proportional input reduction or output expansion) and it is non-oriented, meaning it treats input excesses and output shortfalls symmetrically, which is useful when both inputs and outputs matter equally. The Super-SBM model is more robust and often avoids infeasibility, offering a stable and consistent solution for ranking.

In reality, constraints often cannot be improved from the input side. How to determine the optimal output is a problem that cannot be solved by traditional DEA. For example, the integration of a superefficiency SBM model and machine learning algorithms allows the efficient frontier to be adapted to actual production needs (Zhong et al., 2021). This integration can allow decision makers to further improve their level of management. For future researches, the integration of a superefficiency SBM model and machine learning algorithms still needs more improvement, and there is also a lot of space for application in the efficiency evaluation of large-scale samples, which can be used as one of the directions.

LITERATURE

Alizadeh, A., & Omrani, H. (2019). An integrated multi response Taguchi-neural network-robust data envelopment analysis model for CO2 laser cutting. Measurement, 131, 69-78.

Andersen, P., & Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. Management science, 39(10), 1261-1264.

Banker, R. D., Das, S., & Datar, S. M. (1989). Analysis of cost variances for management control in hospitals. Research in governmental and nonprofit accounting, 5(1989), 269-291.

Bolós, V. J., Benítez, R., & Coll-Serrano, V. (2023). Continuous models combining slacks-based measures of efficiency and super-efficiency. Central European Journal of Operations Research, 31(2), 363-391.

Bose, A., & Patel, G. N. (2015). "NeuralDEA"–A framework using Neural Network to re-evaluate DEA benchmarks. OPSearch, 52, 18-41.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. European journal of operational research, 2(6), 429-444.

Chen, Y. (2005). Measuring super-efficiency in DEA in the presence of infeasibility. European journal of operational research, 161(2), 545-551.

DRAFT VERSION

Chen, Y., & Liang, L. (2011). Super-efficiency DEA in the presence of infeasibility: One model approach. European Journal of Operational Research, 213(1), 359-360.

Chuang, C. L., Chang, P. C., & Lin, R. H. (2011). An efficiency data envelopment analysis model reinforced by classification and regression tree for hospital performance evaluation. Journal of medical systems, 35, 1075-1083.

Cook, W. D., Liang, L., Zha, Y., & Zhu, J. (2009). A modified super-efficiency DEA model for infeasibility. Journal of the Operational Research Society, 60(2), 276-281.

Du, J., Liang, L., & Zhu, J. (2010). A slacks-based measure of super-efficiency in data envelopment analysis: A comment. European Journal of Operational Research, 204(3), 694-697.

Kwon, H. B., & Lee, J. (2015). Two-stage production modeling of large US banks: A DEA-neural network approach. Expert Systems with Applications, 42(19), 6758-6766.

Kwon, H. B., Lee, J., & Roh, J. J. (2016). Best performance modeling using complementary DEA-ANN approach: Application to Japanese electronics manufacturing firms. Benchmarking: An International Journal, 23(3), 704-721.

Lee, H. S., Chu, C. W., & Zhu, J. (2011). Super-efficiency DEA in the presence of infeasibility. European Journal of Operational Research, 212(1), 141-147.

Lee, H. S., & Zhu, J. (2012). Super-efficiency infeasibility and zero data in DEA. European Journal of Operational Research, 216(2), 429-433.

Lee, H. S. (2021). An integrated model for SBM and Super-SBM DEA models. Journal of the Operational Research Society, 72(5), 1174-1182.

Li, Y. (2020). Analyzing efficiencies of city commercial banks in China: An application of the bootstrapped DEA approach. Pacific-Basin Finance Journal, 62, 101372.

Lin, T. T., Lee, C. C., & Chiu, T. F. (2009). Application of DEA in analyzing a bank's operating performance. Expert systems with applications, 36(5), 8883-8891.

Lin, P. H., & Chen, W. H. (2022). Factors that influence consumers' sustainable apparel purchase intention: The moderating effect of generational cohorts. Sustainability, 14(14), 8950.

Lovell, C. K., & Rouse, A. P. B. (2003). Equivalent standard DEA models to provide super-efficiency scores. Journal of the Operational Research Society, 54(1), 101-108.

Matthews, K. (2013). Risk management and managerial efficiency in Chinese banks: A network DEA framework. Omega, 41(2), 207-215.

Mercan, M., Reisman, A., Yolalan, R., & Emel, A. B. (2003). The effect of scale and mode of ownership on the financial performance of the Turkish banking sector: results of a DEA-based analysis. Socio-Economic Planning Sciences, 37(3), 185-202.

Mostafa, M. M. (2009). Modeling the efficiency of top Arab banks: A DEA–neural network approach. Expert systems with applications, 36(1), 309-320.

Santin, D., Delgado, F. J., & Valino, A. (2004). The measurement of technical efficiency: a neural network approach. Applied Economics, 36(6), 627-635.

Seiford, L. M., & Zhu, J. (1999). Infeasibility of super-efficiency data envelopment analysis models. INFOR: Information Systems and Operational Research, 37(2), 174-187.

Schaffnit, C., Rosen, D., & Paradi, J. C. (1997). Best practice analysis of bank branches: an application of DEA in a large Canadian bank. European Journal of operational research, 98(2), 269-289.

Shao, L., You, J., Xu, T., & Shao, Y. (2020). Non-parametric model for evaluating the performance of Chinese Commercial Banks' product innovation. Sustainability, 12(4), 1523.

Tian, N., Tang, S., Che, A., & Wu, P. (2020). Measuring regional transport sustainability using super-efficiency SBM-DEA with weighting preference. Journal of Cleaner Production, 242, 118474.

Tone, K. (2001). A slacks-based measure of efficiency in data envelopment analysis. European journal of operational research, 130(3), 498-509.

Tone, K. (2002). A slacks-based measure of efficiency in data envelopment analysis. European journal of operational research, 143(1), 32-41.

Tsolas, I. E., Charles, V., & Gherman, T. (2020). Supporting better practice benchmarking: A DEA-ANN approach to bank branch performance assessment. Expert Systems with Applications, 160, 113599.

DRAFT VERSION

Xiao-ning, X., Feng, H., & Li-yun, Z. (2013, July). A slacks-based measure of super-efficiency in DEA with interval data. In 2013 International Conference on Management Science and Engineering 20th Annual Conference Proceedings (pp. 314-321). IEEE.

Zhong, K., Wang, Y., Pei, J., Tang, S., & Han, Z. (2021). Super efficiency SBM-DEA and neural network for performance evaluation. Information Processing & Management, 58(6), 102728.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

PERFORMANCE MANAGEMENT IN SERBIA'S PUBLIC SECTOR: INSIGHTS FROM QUANTITATIVE REGRESSION MODELING

ALEKSANDRA PAVLOVIĆ[1], ANDREA IVANIŠEVIĆ[2], KATARINA STOJANOVIĆ[3]

[1] The Academy of Applied Studies Polytechnic, Belgrade, apavlovic@politehnika.edu.rs,
[2] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, andreai@uns.ac.rs,
[3] Faculty of Contemporary Arst, Belgrade; University Business Academy, Faculty of Economics and Engineering Management, Novi Sad, katarina.stojanovic@fsu.edu.rs,

Abstract: *This paper explores the key factors influencing performance management within the public sector of Serbia through the application of quantitative regression modeling. By integrating publicly available data from international and national sources, such as the World Bank, OECD SIGMA, and Serbia's open data portal, the study constructs a composite performance index and examines its relationship with variables including government effectiveness, control of corruption, public expenditure, and digital governance indicators. The results indicate that institutional quality and digital infrastructure development significantly affect performance outcomes, while economic indicators show more nuanced correlations. The analysis provides empirical evidence supporting the need for stronger alignment between strategic management practices and administrative capacities. Additionally, the findings highlight the importance of continuous monitoring, transparent resource allocation, and investment in human capital as levers for enhancing public sector performance. This study contributes to the literature by offering a data-driven perspective on governance reform and by proposing a replicable model for evaluating performance in transitional public administrations. Implications for policymakers are discussed, particularly in the context of EU integration processes and the modernization of public services.*

Keywords: *Performance management, Government effectiveness, Public expenditures, Regression analysis.*

1. INTRODUCTION

Performance management (PM) in Serbia's public sector is undergoing gradual transformation, driven by institutional reforms, digitalisation initiatives, and alignment with European governance standards. The OECD's 2024 SIGMA Monitoring Report highlights critical advancements across areas such as strategic planning, human resource management, service delivery, and digital governance in Serbia (OECD, 2025). Simultaneously, Serbia's digital strategy, initiated in 2017, has expanded e-Government systems and introduced data-driven service delivery models, fostering transparency and enhancing productivity (Arsovska Tomovska, 2021). Despite this progress, enduring challenges, such as weak human resources (HR) competencies, fragmented performance appraisal systems, and limited integration of evaluation mechanisms, continue to hinder optimal performance management (Vukonjanski, 2019). Prior studies in transitional environments suggest that reforms frequently falter due to instability and institutional ambiguity (Šević, 2005).

In light of these dynamics, this study employs quantitative regression modeling to empirically assess how factors such as government effectiveness, corruption control, public spending, and digital governance correlate with overall performance levels in Serbia's public sector. The aim is to provide evidence-based insights for policymakers seeking to strengthen administrative performance and institutional resilience.

2. LITERATURE REVIEW

Understanding the evolution and current state of PM in the public sector requires a review of both international theoretical foundations and country-specific empirical research. This section provides an overview of key academic and professional contributions that examine how public institutions measure, evaluate, and improve their performance. Particular attention is given to studies focusing on transitional countries, including Serbia, where performance management systems are still developing within the broader context of administrative reform and EU integration. The reviewed literature highlights the complexity of

DRAFT VERSION

implementing effective performance frameworks and identifies institutional, technological, and organizational factors that influence outcomes across different governance systems. By synthesizing global insights and national findings, this review aims to establish a conceptual foundation for the empirical analysis presented in the following sections.

Vukonjanski (Vukonjanski, 2019) examines the human resources management (HRM) model in Serbia's public administration. The study finds that the current HRM framework combines contemporary European practices with outdated pre-transition structures. Dual legal models governing civil servants contribute to institutional inconsistencies and hinder reform efforts in HR governance. Group of authors from Serbia (Milojević, Domanović & Knežević, 2024) investigate PM within Serbia's healthcare sector. Based on a survey of 280 healthcare professionals, the authors identify major challenges – absence of a robust PM system, insufficient support from information systems, and rigid organizational structures. They further observe that increased obstacles in the evaluation process negatively affect perceived system flexibility and adaptability. Another group of authors (Vujko, Vuković, Demirović Bajrami, Zečević Stanojević & Zečević, 2022) analyze the relationship between employee engagement and PM in the Fruška Gora National Park, Serbia. Surveying 103 staff members, the study shows that employee engagement is strongly linked to internal communication and belonging. Positive appraisal from superiors enhances engagement, and effective PM practices further reinforce it.

Rana et al. (Rana, Steccolini & Bracci, 2022) examined the evolution of performance auditing in the public sector over the past 50 years. The authors identify two primary roles of performance audits: promoting accountability and enhancing performance. They emphasize the inherent tensions between audit independence and its effectiveness and highlight the need to explore digital and risk-related dimensions of performance auditing. Van Helden &Reichard (Van Helden & Reichard, 2016) the authors found, in their comparative study of PM practices in public versus private sectors, that public organizations tend to adopt a broader range of indicators focused on societal objectives, while private firms concentrate more on financial metrics. Furthermore, strategic alignment processes are reported to be more structured and effective in the private sector. Garengo & Sardi (Garengo & Sardi, 2020), in their bibliometric study, analyzed trends in public sector performance measurement and management. The authors note increasing academic interest in frameworks like the Balanced Scorecard but also critique the lack of alignment with the unique needs of public institutions. They call for future research on themes such as digitalization, stakeholder participation, and the measurement of public value. The study of Mauro et al. (Mauro, Cinquini & Grossi, 2017), by reviewing the literature on performance-based budgeting (PBB) between 1990 and 2014, evaluates its effectiveness in fostering transparency and accountability. The authors find that while PBB frameworks can improve efficiency, their implementation varies significantly across countries, largely depending on institutional and governance contexts. Empirical research of Veledar et al. (Veledar, Bašić & Kapić, 2014) investigates performance measurement practices in the public sector of Bosnia and Herzegovina, a transitional economy. The findings suggest that excessive measurement without reporting negatively impacts operational outcomes, while the lack of reliable performance indicators undermines efficiency and fiscal accountability.

The reviewed literature reveals a growing academic and practical interest in PM within the public sector, particularly in the context of enhancing accountability, transparency, and efficiency. Across various international studies, researchers emphasize the complexity of implementing performance systems in public institutions, where societal goals and institutional constraints often differ from those in the private sector. While models such as performance auditing and performance-based budgeting have shown promise in improving public sector outcomes, their effectiveness largely depends on contextual factors, including governance capacity, institutional culture, and the availability of reliable data. Moreover, scholars increasingly call attention to the limitations of *one-size-fits-all* models, stressing the importance of tailoring performance frameworks to local administrative and social realities. Emerging topics, such as digital transformation, citizen participation, and the measurement of public value, highlight the need for more dynamic and inclusive approaches to PM. This evolving landscape suggests that future research should focus not only on the technical design of measurement tools but also on their integration into decision-making processes and their broader socio-political implications. In summary, the literature supports a nuanced and context-sensitive approach to PM in the public sector, where strategic alignment, stakeholder engagement, and institutional maturity play critical roles in shaping success.

DRAFT VERSION

3. INSTITUTIONAL AND DIGITAL REFORM CONTEXT IN SERBIA

Serbia's public administration has undergone significant reforms, particularly in the last decade, aligning with OECD and EU recommendations. According to the OECD report 2022, key challenges include improving the professionalization of the public administration, developing strategic planning, and strengthening transparency and accountability (OECD, 2022). Digitalization is recognized as a pillar of reform, with e-government programs contributing to more efficient management and accessibility of public services (OECD, 2022).

Corruption remains a significant issue in Serbia, affecting the quality of public services and citizens' trust in institutions. Transparency International 2023 highlights that corrupt practices are particularly prevalent in public procurement, healthcare, and education sectors, negatively impacting public sector performance (Transparency International, 2023). Studies show that corruption control directly influences management efficiency and the implementation of public policies (Rose-Ackerman & Palifka, 2016). Reforms in human resource management are crucial for improving public sector performance. According to research by Janković (2019), performance management, professional development, and meritocracy are only partially implemented in Serbia, which limits the efficiency of public institutions (Janković, 2019). Introducing systems for monitoring and evaluating employee performance can enhance transparency and accountability.

Digital transformation represents a significant driver for enhancing public sector performance. According to Popović et al. (2021), the implementation of e-government in Serbia facilitates better inter-institutional connectivity and more transparent performance monitoring (Popović, Živković & Radenković, 2021). Additionally, the development of public data ecosystems and analytical capabilities supports higher quality decision-making and more efficient governance (Nikiforova, Lnenicka, Milić, Luterek & Rodríguez Bolívar, 2024).

This analysis confirms that reforms in human resource management, digitalization, and corruption control play key roles in enhancing public sector performance in Serbia. At the same time, challenges such as persistent corruption and uneven reform implementation remain barriers to greater progress. Digital tools and the development of public data offer new opportunities for future improvements in performance management systems.

4. DATA AND METHODOLOGY

To evaluate the determinants of public sector performance in Serbia, a multiple linear regression analysis is conducted, using data for the period from 2003 to 2023. Data was taken from the official website of the World Bank Governance Indicators and World Bank macroeconomic trends. The dependent variable in the model is the Public Sector Performance Index, while the independent variables include: Government Effectiveness, Control of Corruption, and GDP per Capita (USD). These variables were chosen based on the reviewed scientific literature, which suggests their relevance for explaining performance outcomes in the public sector. The SPSS statistical software package was used for statistical data processing.

4.1. Model specification

The linear regression model used in the analysis is expressed as follows:

$$Performance_Index = \beta_0 + \beta_1 (Gov_Effectiveness) + \beta_2 (Control_Corruption) + \beta_3 (GDP_per_Capita) + \varepsilon$$

Where:. β_0 is the intercept, β_1–β_3 are the coefficients of the independent variables, while ε is the error term.
A description of the variables is given in Table 1 below.

Table 1: Variable description

Variable	Description
Performance_Index	Composite index of public sector performance
Gov_Effectiveness	Quality of public services and policy formulation
Control_Corruption	Extent to which public power is exercised for private gain
GDP_per_Capita	Gross domestic product per capita in USD

DRAFT VERSION

4.2. Results of the Regression Analysis

The model explains approximately 92% of the variance in public sector performance (R-square: 0.919). After adjusting for the number of predictors, the model remains highly explanatory (Adjusted R-square: 0.904). The model as a whole is statistically significant (F-statistic: 63.90, $p < 0.0001$). This is concluded based on Table 2, which follows.

Table 2: Regression coefficients

Variable	Coefficient	Std. Error	t-value	p-value	95% CI
Intercept	45.6957	1.678	27.23	<0.001	[42.156, 49.236]
Gov_Effectiveness	34.7720	7.829	4.44	<0.001	[18.254, 51.290]
Control_Corruption	3.2656	13.219	0.25	0.808	[-24.623, 31.154]
GDP_per_Capita	0.0020	0.001	2.32	0.033	[0.000, 0.004]

Government Effectiveness has a strong and statistically significant positive impact on performance. A 0.1 increase in effectiveness results in a 3.48-point increase in the index. Control of Corruption does not have a statistically significant effect in this model, possibly due to multicollinearity or data limitations. GDP per Capita has a small but statistically significant positive effect, supporting the hypothesis that economic development contributes to performance outcomes.

This model confirms that government effectiveness and economic development are key determinants of public sector performance in Serbia. Although control of corruption is theoretically relevant, its statistical insignificance here suggests a need for either better-quality indicators or a more refined model. Further research could expand the model with qualitative data or longitudinal institutional analysis.

5. CONCLUSION

This study has explored the dynamics of performance management in Serbia's public sector through a quantitative lens, highlighting the critical role of government effectiveness, corruption control, and economic development as key determinants. The regression analysis demonstrated a strong positive association between government effectiveness and public sector performance, confirming that institutional quality significantly influences the capacity of public organizations to deliver efficient and transparent services.

Despite ongoing reforms aligned with OECD and EU standards, challenges such as persistent corruption and partial implementation of human resource management reforms continue to hinder optimal performance outcomes. The limited statistical significance of corruption control in the model underscores the complexity of governance issues that transcend quantitative measurement and call for integrated qualitative insights. Digital transformation initiatives, including e-government platforms and the development of public data ecosystems, present promising avenues for enhancing administrative efficiency and accountability. These technological advancements can serve as catalysts for improving data-driven decision-making and more responsive public service delivery.

Future research should aim to integrate longitudinal qualitative analyses with advanced modeling techniques, such as machine learning-based KPI selection, to better capture the multidimensional nature of performance management. Moreover, expanding the scope to include citizen satisfaction and service outcome measures could provide a more comprehensive understanding of public sector effectiveness in Serbia. In conclusion, while Serbia has made measurable progress in modernizing its public administration, sustained efforts in institutional reform, anti-corruption measures, and digital governance are essential to unlock the full potential of performance management systems and foster trust in public institutions.

DRAFT VERSION

ACKNOWLEDGMENT

This research received no external funding..

LITERATURE

Arsovska Tomovska, M. (2021). *The future of government: Serbia's growing bet on digital transformation pays off*. World Bank Governance Global Practice Blog,

Garengo, P. & Sardi, A. (2020). Performance measurement and management in the public sector: State of the art and research opportunities. *International Journal of Productivity and Performance Management*, 70(7), 1629–1654. Abstract & metadata on RePEc/IDEAS: https://ideas.repec.org/a/eme/ijppmp/ijppm-03-2020-0102.html

https://blogs.worldbank.org/en/governance/future-government-serbias-growing-bet-digital-transformation-pays?utm_source=chatgpt.com

https://doi.org/10.1007/978-3-030-05348-2_21

https://doi.org/10.1108/09513550510624059

Janković, S. (2019). Human Resource Management Reforms in Serbian Public Administration: Challenges and Opportunities. *Public Administration Review*, 79(3), 421-432, doi: https://doi.org/10.1111/puar.13025

Mauro, S. G., Cinquini, L. & Grossi, G. (2017). Insights into performance-based budgeting in the public sector: A literature review and research agenda. *Public Management Review*, 19(7), 911–931. Abstract metadata & doi: https://iris.unimore.it/handle/11380/1264611 and https://doi.org/10.1080/14719037.2016.1243810

Milojević, S., Domanović, V. & Knežević, S. (2024). Key factors of healthcare performance management: A case study from Serbia. *Journal of Infrastructure, Policy and Development*, 8(16), Article 10199, doi: https://doi.org/10.24294/jipd10199

Nikiforova, A., Lnenicka, M., Milić, P., Luterek, M., Rodríguez Bolívar, M. P. (2024). The Evolution of Public Data Ecosystems: An Empirical Study. *Government Information Quarterly*, Available at: https://arxiv.org/abs/2405.13606

OECD (2022). Public Administration Reform in Serbia: Progress and Challenges. OECD Public Governance Reviews, Available at: https://www.oecd.org/gov/public-administration-reform-serbia.pdf

OECD. (2025). *Public Administration in Serbia 2024: Assessment against the Principles of Public Administration*. SIGMA Monitoring Reports, OECD Publishing, Paris, doi:

Popović, K., Zivkovic, M. & Radenkovic, M. (2021). E-Government Development and Its Impact on Public Sector Performance in Serbia. *Government Information Quarterly*, 38(1), 101561, doi: https://doi.org/10.1016/j.giq.2020.101561

Rana, T., Steccolini, I. & Bracci, E. (2022). Performance auditing in the public sector: A systematic literature review and future research avenues. *Financial Accountability & Management*, 38(3), 337–359, open-access PDF (author accepted manuscript available): https://hdl.handle.net/11585/899066

Rose-Ackerman, S., & Palifka, B. J. (2016). Corruption and Government: Causes, Consequences, and Reform. *Cambridge University Press*, ISBN: 9781107441095, doi: 10.1017/CBO9781139962933

Šević, Ž. (2005). Measuring performance on a local government level in a transitional country: the case of Serbia. *International Journal of Public Sector Management*, 18(7), 582-603, doi:

Transparency International (2023). Corruption Perceptions Index 2023. Available at: https://www.transparency.org/en/cpi/2023

Van Helden, J. & Reichard, C. (2016). Commonalities and differences in public and private sector performance management practices: A literature review. In M. J. Epstein, F. Verbeeten, & S. K. Widener (Eds.), *Performance Measurement and Management Control: Contemporary Issues* (Vol. 31, pp. 309–351). Emerald Group Publishing. Abstract available via PSAAR & Emerald: https://psaar.net/abstracts/296-commonalities-and-differences-in-public-and-private-sector-performance-management-practices-a-litarature-review

DRAFT VERSION

Veledar, B., Bašić, M. & Kapić, J. (2014). Performance measurement in public sector of transition countries: The case of Bosnia and Herzegovina. *Business Systems Research Journal*, 5(2), 72–83. Full open-access article: https://hrcak.srce.hr/ojs/index.php/bsr/article/view/12528

Vujko, A., Vuković, D., Demirović Bajrami, D., Zečević Stanojević, O. & Zečević, L. (2022). The Nexus between Employee Engagement and Performance Management Processes – Fruška Gora National Park (Serbia) Case Study. *Sustainability*, 14(18), 11489, doi: https://doi.org/10.3390/su141811489

Vukonjanski, I. (2019). Characteristics of human resources management model in public administration of Republic of Serbia: Contemporary trends and challenges. *Facta Universitatis Series: Law and Politics*, 17(1), 33-45, doi: https://doi.org/10.22190/FULP1901033V

DRAFT VERSION

RESEARCH & DEVELOPMENT

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

100+ YEARS OF PATENTING IN SERBIA

ĐURO KUTLAČA[1]

[1] University of Belgrade, Mihajlo Pupin Institute, Belgrade, djuro.kutlaca@imp.bg.ac.rs, 0000-0001-6439-5660
[1] Belgrade, kutlacadjuro@gmail.com, 0000-0001-6439-5660

Abstract: *The paper analyses patenting in Serbia from the establishment of the Office for the Protection of Industrial Property of the Kingdom of Yugoslavia in 1921 to 2023. In a period of 102 years, domestic inventors were exposed to major political changes, wars, changes in the economic system and conditions for inventive work. For foreign inventors, the market in Serbia was a challenge to conquer, in which one of the first instruments was patent protection. The long time series of data on domestic and foreign patent applications represents a rich information fund for the analysis of one of the most important indicators of the functioning of the national innovation system. One of the conclusions that imposes itself on the author is that Serbia, after 102 years of patent activity, is somewhere again at the beginning of the path to the development of the economy based on its own knowledge and technologies.*

Keywords: *Patent activity, Serbia, Longitudinal research.*

1. INTRODUCTION

Patent statistics are one of the very few measurable indicators of research and development activities aimed at potential market exploitation. Moreover, patent activity is still a phase that precedes the development of innovations offered on the market of goods and services. Namely, many patents will never reach that phase of the innovation cycle, and some patents can be used to develop multiple innovations (so-called "generic" patents). Therefore, in the literature, authors have reached a consensus that patent statistics are important, but imperfect indicator of innovative activities (Pavitt, 1988; Griliches, 1990). Also, the literature proves that the number of patented inventions provides a fairly good, although not perfect, representation of innovative activity, thus supporting the use of the number of patents in studies examining technological change (Acs et all, 2002). Furthermore, a methodology is given that combines patent content clustering and technology life cycle forecasting to find a niche in the industry, as part of technology forecasting research (Trappey et all, 2011), that is, a model is developed to facilitate governments, inventors and manufacturing organizations to identify new research directions, which includes steps such as knowledge extraction on patents, presentation of knowledge and identification of research trends (Abbas et all, 2014)

Given the importance that patents can have for the development of new and/or improved products and services, and for innovation activity as a whole in general, the likelihood of success in directing patent activity towards the market realization of innovations largely depends on the model of scientific research and development (R&D) work that is represented in an economy, that is, the model on which the management and financing of research and innovation in a country depends. The classification of models and a broader discussion of their application in the governance of the research and innovation in Serbia is given in the author's article [Kutlača, 2022], which also provides a chronology of the development of innovation policies: (1) The first generation of innovation policy is based on the so-called "linear model of innovation activities", starting from basic sciences and moving in successive phases until new knowledge is incorporated into applications acceptable for market valorisation; (2) The second generation of innovation policy is based on the so-called "chain model", which characterizes feedback loops between different phases of the innovation process [Klein and Rosenberg, 1986]; (3) The third generation of innovation policy places innovation activities as a key area of every policy and represents the so-called "innovation activities integration model".

DRAFT VERSION

The evolution of innovation policies from a linear model, through a chain model, has led to a model of integration of innovation activities. The chain model is particularly important because of the feedback loops, which, in a theoretical sense, it has led to the concept of national, regional, sectoral and other types of innovation systems [Freeman 1987]. According to this model, innovation policy aims to improve two-way communication between different actors in the chain of the innovation activities and improve all types of innovation systems in order to ensure the establishment of an information-based decision-making system, i.e., to inform decision-makers on research and development, commercialization, adoption and implementation of technology, etc. It can be said that this model is still dominant in the organization of public governance of research and development and other innovation activities, and the third generation of innovation policies manages the networking process between all actors in the *national innovation system* (NIS). Innovation can be considered as the result of a learning process, which is interactive and cumulative in nature. Interactions within the economy, which may result in the combination of existing knowledge or, in some cases, new knowledge, take place between companies, financial, educational, research and development and other organizations, government agencies and other public and/or private institutions [Galli and Teubal, 1997]. Technological development is the result of complex interactions between the aforementioned organizations, which Richard Nelson first called the "*capitalist engine of growth*" [according to: Albuquerque, 1997], and later the "*national innovation system*" (NIS) [Nelson, 1993]. The NIS scheme is given in Figure 1, according to the modified concept of the author S. Kuhlmann [Kuhlmann, 2003].

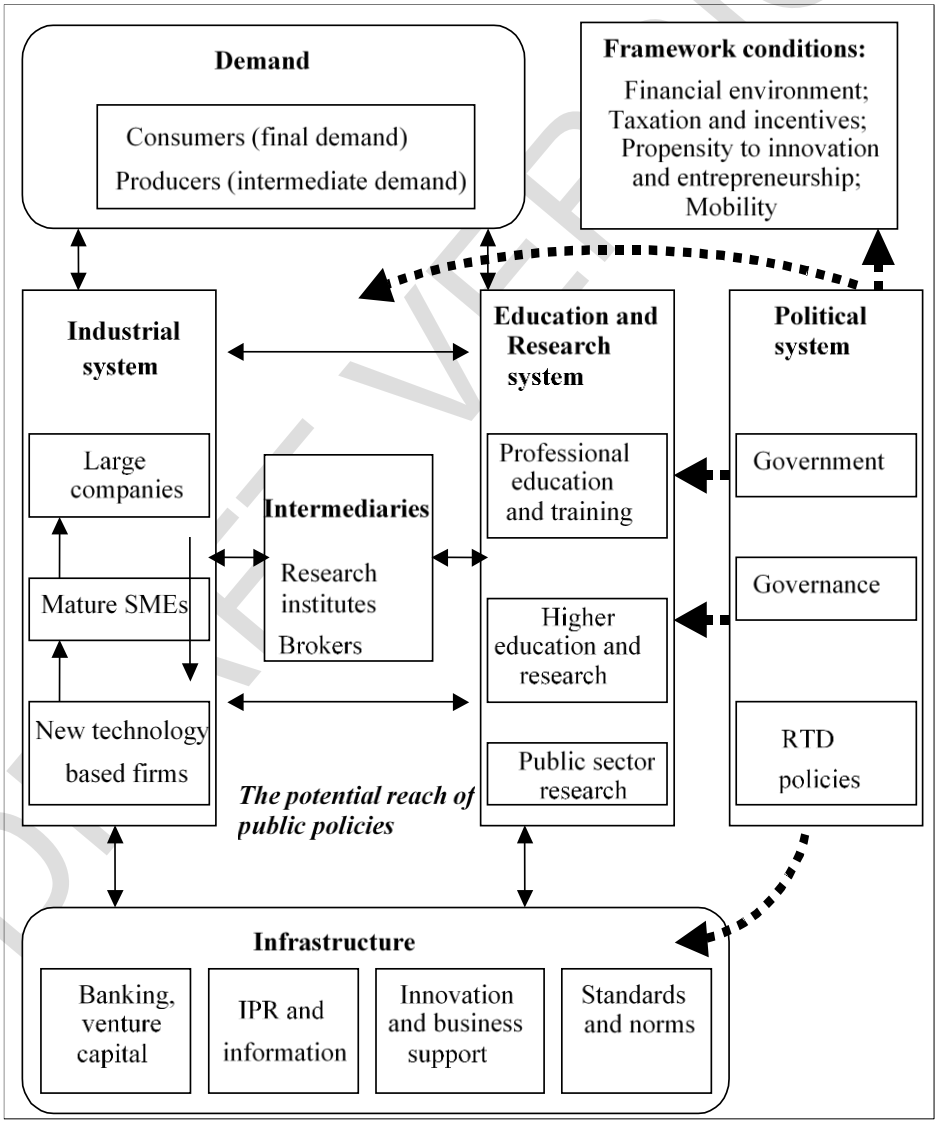


Figure 1: NIS – Concept modified and extended by S. Kuhlmann [Kuhlmann, 2003]

Research on the governance of science and technology development indicates the subject of interest of the relevant ministries in building a national innovation system. Thus, four key areas of management actions in the EU member states have been identified [Kuhlmann, 2003]: (1) In the 1970s, research and development (R&D) activities were stimulated as the main objective and main priority of innovation policy (IP). Subsidies

DRAFT VERSION

for R&D work were the content of IP, and financing was the main instrument of IP, both in the member states and at the level of the European Union; (2) In the 1980s, the transfer of knowledge and/or technological competence was stimulated as a priority and main objective, and research and development activities were stimulated as the content of innovation policy. The main instrument of EU innovation policy was support for the diffusion of knowledge and technologies; (3) In the 1980s, support for business start-ups was stimulated as the main objective of innovation policy. Social sciences, formal and tacit knowledge were the priorities and content of IP. The main instrument of EU innovation policy was support for management in research and development; (4) Since the late 1990s, and continuing to this day, facilitating change has been promoted as a priority of EU innovation policy. The content of innovation policy consists of scientific areas, formal and tacit knowledge, as well as the provision of strategic information, necessary for both current innovation activities and planning for future ones. System instruments are the main instruments of IP, e.g., supporting and reorganizing innovation networks, providing initiatives for organizational change and for the implementation of activities for predicting NT development ("foresight" in science and technology).

This introduction, which outlines the evolution of innovation policies, with the main elements of the governance mechanisms of administrations responsible for research and innovation, aims to highlight the importance that science and innovation governance has on the scientific research community and potential inventors and innovators. The results or consequences of science and innovation governance in a country are best illustrated by two sets of indicators: (1) Indicators of publication of papers in journals registered in the WoS and Scopus databases; (2) Indicators of intellectual property protection, especially patents, in national intellectual property offices, as well as in three specific offices: the European Patent Office (EPO), the US Patent Office (USPTO) and the Japanese Patent Office (JPO). The bottom of Figure 1 illustrates the infrastructure for NIS, without which NIS cannot be formed and function. One of the elements of this infrastructure is intellectual property rights (IPR). It is precisely the knowledge codified in patent documentation that represents a resource for further development into marketable products/processes/services, which must be available to all NIS actors. Reform of legislation in the field of intellectual property rights (IPR) in order to harmonize with EU standards in the countries of Central and Eastern Europe is an integral part of transitional changes [WDR, 1996] and one of the indicators of success in the implementation of negotiation chapters in the EU accession process (Chapter 7: Intellectual property law) in a particular country [EC, 2024]. Harmonization of legislation in the field of IPR with EU standards is one of the main conditions for the entry of foreign direct investment, because before entering the market of a country, foreign companies first protect their patents and other intellectual property rights, and only then bring their technologies, processes and products. Again, through the indicators of patent activity in a country, governance mechanisms for supporting research and innovation can be analysed, because these indicators indicate the interest of foreign inventors in the protection of IPR in a country. It has been shown in the literature that the number of patents protected by non-residents of a country in the countries of Central and Eastern Europe from 1945 to 1989 was marginal, to some extent with the exception of the former Yugoslavia [Radosevic and Kutlaca, 1998]. It singles out a set of IPR indicators as a result of research and innovation for the subject of this paper.

The paper analyses patenting in Serbia from the establishment of the Office for the Protection of Industrial Property of the Kingdom of Serbia in 1921 to 2023. In a period of 102 years, domestic inventors were exposed to major political changes, wars, changes in the economic system and conditions for inventive work. For foreign inventors, in some periods the Serbian market was a challenge to conquer, in which one of the first instruments was patent protection. The long time series of data on domestic and foreign patent applications represents a rich information fund for the analysis of one of the most important indicators of the functioning of the NIS. The analysis itself aims to indicate the effects of the governance mechanisms that the responsible administrations established in that period to support R&I. At the end of this extensive introduction, one of the conclusions that imposes itself on the author is that after 102 years of patent activity, Serbia is once again at the beginning of the path to economic development based on its own knowledge and technologies.

2. PATENTING IN SERBIA

In a paper from 1998, the author analyses the influence and effects of the decisions of the responsible administration on patenting in Serbia from 1921 to 1995 [Kutlaca, 1998]. This paper extends the time series of data on patent applications by residents and non-residents in Serbia until 2023. The concluding remarks comment on the governance of research and innovation in Serbia in the last three decades of patenting activity.

DRAFT VERSION

2.1. Patent applications of domestic authors

The Kingdom of Serbia is one of the signatory states - founders of the Paris Convention on the Protection of Industrial Property on March 20, 1883, together with Belgium, France, Italy and the Netherlands. Great Britain acceded to the Convention in 1884, the USA in 1887, and Germany in 1903. The Kingdom of Yugoslavia inherited the status of Serbia and acceded to the Convention in 1921. The Paris Convention on the Protection of Industrial Property from 1883 is still a valid regulation and standard for relations between countries in the field of intellectual property protection.

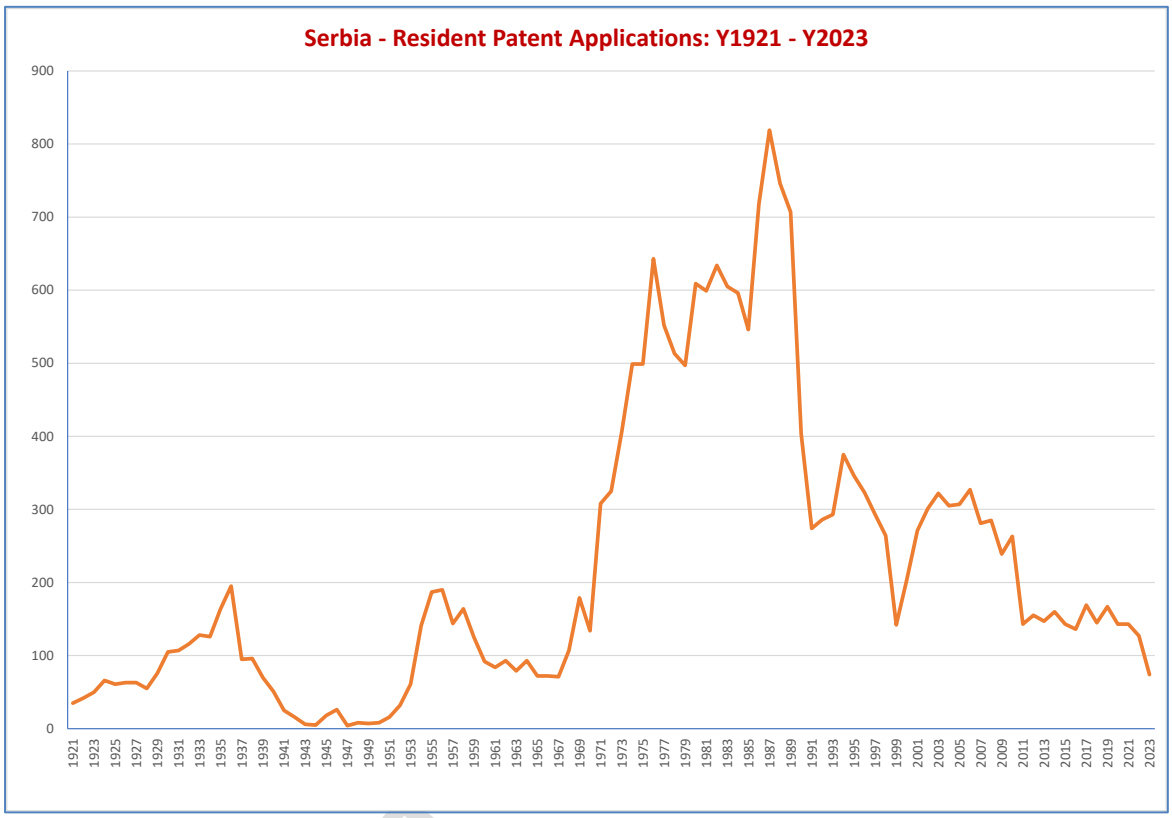


Figure 2: The patenting of domestic authors from Serbia in the observed period, from 1921 to 2023

Having established an adequate relationship with the ownership of intellectual property, Serbia developed its capitalist economy at the end of the 19th and the beginning of the 20th century by encouraging inventive work, formal protection of intellectual property rights, and then continued innovative activities aimed at the development of products, processes and services. Figure 2 illustrates the patenting of domestic authors from Serbia in the entire observed period, from 1921 to 2023. *The patent activity of domestic inventors* grew from 1921 until the beginning of the WW2. After the end of the war, greater involvement of inventors begins again, and reaches its culmination in the period from 1971 to 1991. War events in the former Yugoslavia drastically reduced the number of patent applications by domestic authors, with a slight halt in the decline in that number in the period from 2001 to 2010. Then there is a decline and stagnation again, only to drop drastically again with the corona virus epidemic. At the end of the observed period of 103 years, shortly before the epidemic, the intensity of patenting was at the level reached before the WW2.

2.2. Patent applications of foreign authors

The conditions for the protection of intellectual property in the Kingdom of Yugoslavia represented an acceptable environment for foreign companies and inventors, which brought their technologies to Yugoslavia / Serbia with prior protection in the Office for the Protection of Industrial Property of the Kingdom of Yugoslavia. By the beginning of WWII, the number of patent applications by both domestic and foreign authors reached a respectable level, and the protected technologies were the driver of the country's industrialization and a source of knowledge for the educated population. Figure 3 illustrates the patenting of foreign authors in the entire observed period, from 1921 to 2023.

DRAFT

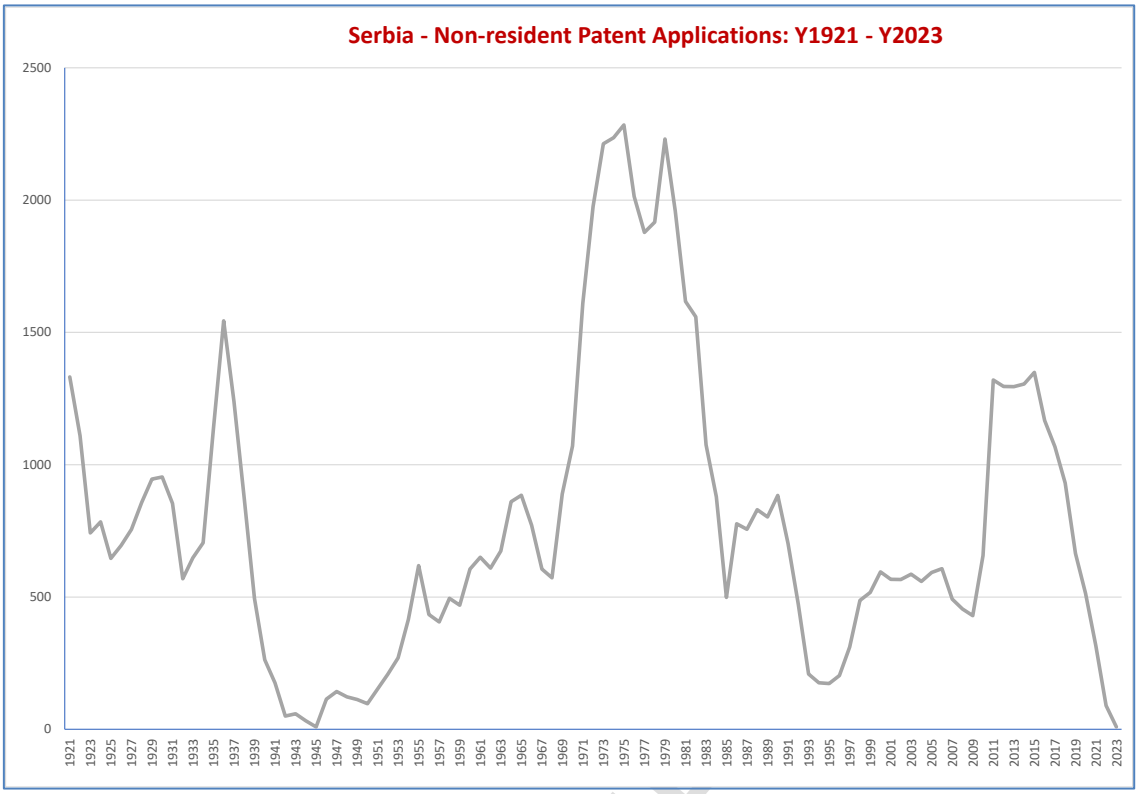


Figure 3: The patenting of foreign authors in the observed period, from 1921 to 2023

After WW2, *foreign authors* applied for more and more patents, and significant growth was recorded in the period from 1971 to 1985, similar to the case of domestic inventors. It is indicative, however, that the decline in the number of applications by foreign authors began a few years before the beginning of the disintegration of the former Yugoslavia, because the deterioration of the political situation in Yugoslavia had the effect of reducing the interest of foreigners in entering that market. The relative increase in the interest of foreign inventors in patent protection in Serbia can be seen again in the period from 2001 to 2017, and then a drastic drop in the number of patent applications by foreign authors begins. The growth of the economy that is said to be happening after the corona virus epidemic is obviously not based on technologies that need protection, that is, it is driven by labour-intensive technologies that were launched (and protected) on the market many years ago. Obviously, foreign inventors are waiting for the political situation in the country to change and stabilize in order to re-enter the Serbian market. The *total number of patent applications* by domestic and foreign authors together in the observed period from 1921 to 2023 follows the behaviour of foreign inventors, which is a consequence of a significantly higher number of patent applications by foreign authors compared to the number of applications by domestic authors. Hence, everything that was stated in the discussion of patenting by foreign authors, mostly applies to the overall patent activity in Serbia.

3. CONCLUSION

In the aforementioned paper from 1998, the author discusses the connection between changes in patent activity and political decisions, that is, laws and by-laws regulating the governance of science and research in the state, which greatly influenced the intensity of patenting [Kutlaca, 1998]. An additional analysis is provided here for the additional period covered by this paper, i.e. for the years from 1995 to 2023. In that period, the following critical periods are significant, which are reflected in Fig. 2 in a large change (mainly decline) in the number of patent applications by domestic authors: (1) The first turning point is precisely 1995-1996, because the end of the wars in the former Yugoslavia is an excuse for the end of the state's support for domestic inventive and innovative efforts, since with the lifting of sanctions, the missing technologies can be bought abroad; (2) The second turning point is 2000-2001, because with the change of regime in Serbia, new perspectives of development based on own knowledge are opened; (3) The third breakthrough period will be 2010-2011, with the beginning of the new project cycle of financing scientific research projects, in which the main emphasis is placed on the publication of articles in journals registered in the WoS and Scopus databases. The laws regulating science and innovation were changed, with the formal recognition of the importance of

DRAFT VERSION

patenting, but the mentioned publication, as the main condition for advancement in the academic and scientific career, completely prevailed in the activities of the higher education and scientific community; (4) The final collapse of patent activity began with the corona virus epidemic in 2019, with negative effects lasting until the end of the observed period.

Due to length limitations, the paper presents only a few fragments of the analysis provided by patent statistics for 103 years of patent activity in Serbia. The rich information base represents the basis for a better understanding of the process of building innovation capacities and creating a national innovation system, within and according to the rules of knowledge-based development in the 21st century.

LITERATURE

Abbas A, et all (2014): A literature review on the state-of-the-art in patent analysis, *World Patent Information*, http://dx.doi.org/10.1016/j.wpi.2013.12.006

Acs Zoltan J, Anselin Luc, Varga Attila (2002): Patents and Innovation Counts as Measures of Regional Production of New Knowledge, February 2002, *Research Policy* 31(7):1069-1085, DOI:10.1016/S0048-7333(01)00184-6

Albuquerque, Eduardo da Motta e (1997): "*National systems of innovation: notes about a rudimentary and tentative "typology*"", SPRU, Sussex University, Brighton.

EC (2024): *Serbia 2024 Report*, European Commission 2024 Communication on EU enlargement policy, SWD(2024) 695 final, Brussels, 30.10.2024

Freeman, Christopher (1987): "*Technology and Economic Performance: Lessons from Japan*", Pinter, London, 1987.

Galli, Riccardo and Teubal, Morris (1997): "*Paradigmatic Shifts in National Innovation Systems"*, Chapter in: Systems of Innovation-Technologies, Institutions and Organizations, ed. by C. Edquist, Pinter, London.

Griliches, Z. (1990): "Patent statistics as economic indicators: a survey", *Journal of Economic Literature*, Vol. 28, pp. 1661-1707.

Kuhlmann, S. (2003), "*Future governance of innovation policy in Europe*", Chapter in: "Future directions of innovation policy in Europe", Innovation Papers No 31, pp. 40-48, European Commission, 2003

Kutlaca, Djuro (1998): "Patent-Related Activities in Serbia from 1921 to 1995", *Scientometrics*, Vol. 42, No. 2, June 1998, pp.171-193.

Kutlača, Đ. (2022): "*Systematic Study of Science*", Scientific and Professional Conference "Science: State and Perspectives", held on September 23 and 24, 2022, Proceedings, pp. 185-222, Publisher: Serbian Academy of Sciences and Arts, Belgrade. ISBN 978-86-6184-025-8, COBISS.SR-ID 161342217

Nelson, R. (1993): "*National Innovation Systems: a Comparative Analysis*", Oxford University Press.

Pavitt, Keith (1988): "*Uses and abuses of patent statistics*", in: Van Raan, A.F.J. (ed.): "Handbook of Quantitative Studies of Science and Technology", North Holland, Amsterdam.

Radosevic, Slavo and Kutlaca, Djuro (1998): "*Assessing the Basis for 'Catching-up' of Eastern Europe: an Analysis Based on US Foreign Patenting Data*", STEEP Discussion Paper No 42, SPRU, University of Sussex, Brighton, UK, March 1998.

Trappey Charles V., Hsin-Ying Wu, Fataneh Taghaboni-Dutta, Trappey Amy J.C., (2011): Using patent data for technology forecasting: China RFID patent analysis, *Advanced Engineering Informatics* 25, pp. 53–64, 2010 Elsevier Ltd.

WDR (1996): "*From plan to market*", World Development Report 1996, The World Bank, Published by Oxford University Press, Inc., New York.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

OPERATIONS RESEARCH IN CIVIL ENGINEERING AND GEODESY ACADEMIC STUDIES IN FORMER YUGOSLAV REPUBLICS

JOVANA TOPALIĆ[1], TIANA MILOVIĆ[2], VLADIMIR MUČENSKI[3], MEHMED BATILOVIĆ[4]
[1] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, jovanatopalic90@uns.ac.rs, 0000-0001-7854-3257
2 University of Novi Sad, Faculty of Technical Sciences, Novi Sad,tiana.milovic@uns.ac.rs 0000-0002-3905-7018
[3] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, mucenskiv@uns.ac.rs, 0000-0001-9830-4747
[4] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, mehmed@uns.ac.rs, 0000-0001-5605-1112

Abstract: *The research presented in this paper is an extended research on the implementation of operations research in the subjects of undergraduate (bachelor) and master academic studies at the faculties of civil engineering and geodesy. For the purposes of this research, besides the Republic of Serbia, five more countries in the region were included: Bosnia and Herzegovina, Croatia, Slovenia, Montenegro, and the Republic of North Macedonia. A total of fourteen faculties from former Yugoslav countries were analyzed, and all of them are public, except UNION Nikola Tesla. This topic emphasizes the importance of implementing operations research in civil engineering and geodesy subjects as a mathematical method of decision-making.*

Keywords: *Operations research, civil engineering, geodesy, bachelor programmes, master studies.*

1. INTRODUCTION

Operations research is a field of applied mathematics that integrates advanced analytics methods to improve management and decision-making (Medium, 2021). Moreover, mathematical tools are used not only to investigate mathematics further, but rather to build mathematical models, analyze complex situations, and solve problems within the operations research domains (Stonybrook, 2025). Its historical origins are in the 17th century (Medium, 2021) when mathematicians Blaise Pascal and Christiaan Huygens solved problems involving complex decisions (the problem of points - a classical problem in probability theory) by using game-theoretic ideas and expected values, while Pierre de Fermat and Jacob Bernoulli solved these types of problems using combinatorial reasoning instead. The modern version of operations research dates back to World War II, when sciencists and mathematicians were recruited to solve complex military problems such as critical logistics and supply chain issues. In 1938, the term operations research was used for the first time by a group of British Royal Air Force officers and civilian scientists who were asked to determine how recently developed radar technology could be integrated into the wider air defence network (Assad, 2011). In 1947, an American mathematician George B. Dantzig created the first mathematical technique of operations research, called the Simplex Method, which greatly simplified the solution of linear programming problems (University, 2025). According to C.Goodeve (Goodeve, 1948) the operations research represents a scientific method for providing executive departments with a quantitative basis for decisions regarding the operations under their control. The purpose of operations research is to translate complicated, but important real-world problems into precise, quantitative language that can be clearly and unambiguously analysed (Rajgopal, 2025). According to (Gupta, 2019), the operations research development stages consist of the following six sequential steps: (i) Observation of the problematic setting and the problem formulation; (ii) Creating a model; (iii) Deriving the solution from the model; (iv) Testing the model and its solution as well as updating the model; (v) Establishing control over the solution, and (vi) Implementation of the final results. In order to optimize and improve decision-making processes different techniques can be used (Kapadi, 2022), such as: linear programming, non-linear programming, integer programming, decision theory, queuing theory, game theory, network scheduling, information theory, Markov Process, simulation, etc. Today, operations research has a wide range of applications such as in industries, from telecommunication to health care to financial services, facilities planning, manufacturing, the construction sector, marketing, human resources, etc. Therefore, Alex Elkjær Vasegaard (Medium, 2021) tried to illustrate the complexity of operations research in terms of the related fields, subfields, and the addressed problems, although he was limited to a 2D representation as there were many other connections between disciplines, e.g. probability theory and statistics are integral to machine

DRAFT VERSION

learning (ML), Figure 1. Furthermore, artificial intelligence (AI), defined as machine capacity to carry out operations that ordinarily require human intellect, and ML, as a branch of AI that focuses on developing algorithms that learn from data and experience, have become essential tools for advancing operations research (Soori, 2023). AI techniques have shown great potential in improving every stage of the operations research process.

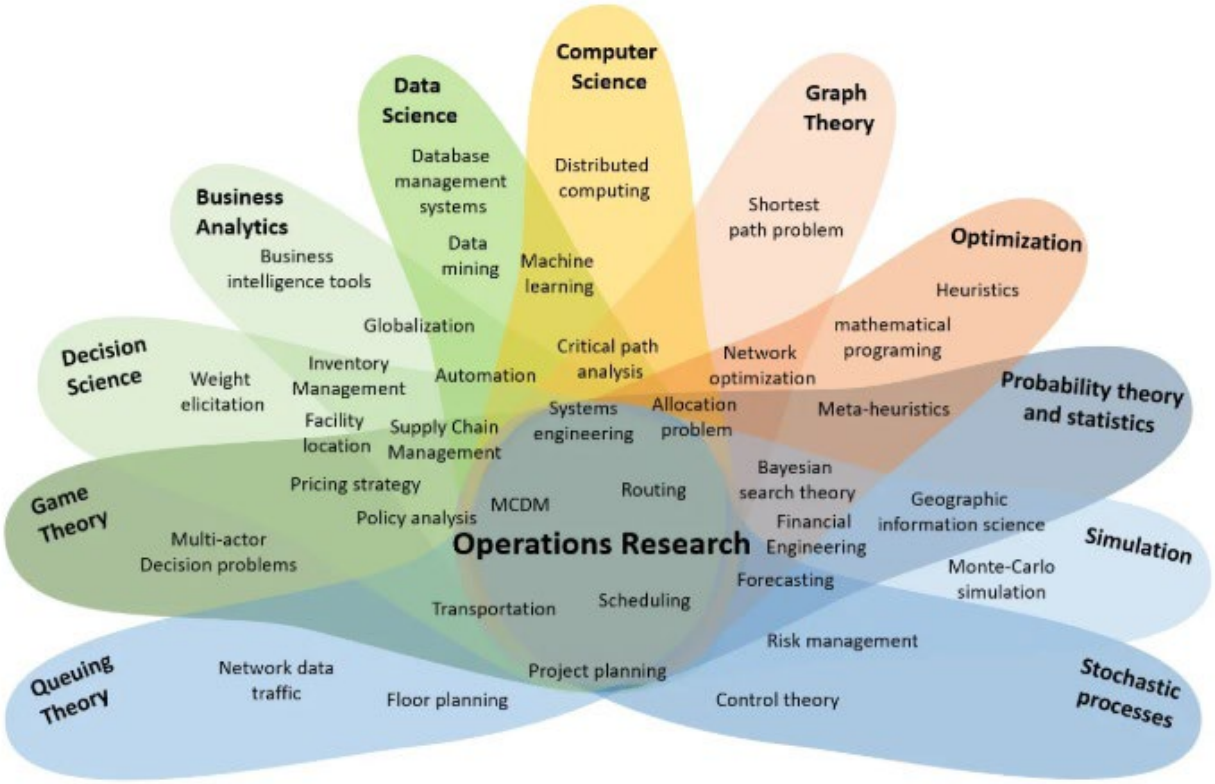


Figure 1: 2D holistic illustration of the disciplines and problems related to operations research (Medium, 2021)

In this paper are presented the research about the implementation of operations research in subjects at civil engineering and geodesy undergraduate and master academic studies in former Yugoslav republics.

2. THE IMPLEMENTATION OF OPERATIONS RESEARCH AT FACULTIES OF CIVIL ENGINEERING AND GEODESY

Operations research techniques are frequently required in engineering design, where formulating a design task as an optimization problem—without imposed economic or technical constraints—often results in multiple feasible solutions. Although these methods were initially created to address managerial and operational challenges in established systems, they are also well-suited for application in various engineering design scenarios. This is the reason why it is important to implement operations research in the engineering study programmes.

Operations research, also referred to as Management Science is the application of a scientific approach to solving difficult resource allocation, management and engineering problems in order to make better decisions. Operations research has applications in many fields, including civil and environmental engineering (Jaynes, 1957)]. There are ways to combat modeling limitations and data quality issues, but it is the responsibility of the operations research analyst to incorporate any limitations or assumptions into their analysis. Operations research tools should not be treated as a black box from which results are implemented without question, interpretation and context are vital to their successful application.

There is established EURO Working Group “Operations Research in Sustainable Development and Civil Engineering”. This idea was taken (Lownes, 2023) after a long international collaboration between researchers from several European countries (Lithuania, Germany, Poland, United Kingdom, Belgium, Denmark, Latvia,

DRAFT

Estonia, Czech Republic, Slovenia). The working group was created at EURO Conference XXIII in 2009, Bonn (Germany). In the 2013. in this group were more than 100 scientists from 20 countries (Lithuania, Germany, Poland, United Kingdom, Belgium, Denmark, Netherlands, Portugal, Latvia, Estonia, Czech Republic, Slovenia, Russia, Ukraine, Australia, Iran, USA, Taiwan and others), but at the early beginning research group consisted of the professors from the only Lithuanian, German and Polish academic and research centres. Marking the importance and the uniqueness of the beginning of collaboration this research group was later in articles referred to as the "scientific triangle".

This article is written on the foundation of former research. The former research was all about civil engineering faculties and investigation of operations research subjects at the Faculties. Because of the importance of operations research in engineering studies, the research is updated with analysis of civil engineering curriculums for any changes and geodesy curriculums at the same Universities.

The implementation of operations research in civil engineering and geodesy is a need for the future, and because of that, the scope of this research is defined. The scope is to find out more about the inclusion of operations analysis and research in subjects at the Faculty of Civil Engineering and Geodesy in the former Yugoslav republics. The selection of public faculties includes one private faculty. The faculties of civil engineering and geodesy in six countries are targeted, and the curriculum of every faculty is analyzed.

In the Republic of Serbia, faculties of civil engineering in Novi Sad, Belgrade, Subotica and Niš are selected. Geodesy faculties in the Republic of Serbia are in Novi Sad, Belgrade and Subotica. There is no geodesy study programme in Niš and Novi Pazar. The one private Faculty in Belgrade is also analyzed for civil engineering studies. In Slovenia the Faculty of Civil Engineering and Geodesy in Ljubljana is analyzed. In Montenegro the Faculty of Civil Engineering in Podgorica is analyzed and geodesy study programme at the private Faculty of Applied Sciences. In Croatia the Faculty of Civil Engineering and Geodesy Faculty in Zagreb is selected. In the Republic of North Macedonia, the Faculty of Civil Engineering in Skopje, is also selected. After detailed analysis of the curriculum of undergraduate (bachelor) studies and master studies at the selected Universities the results are summarized in the Table 1.

Table 1: Subjects with implementation of operations research at the Civil Engineering and Geodesy faculties

	Year	FACULTY	CITY	STUDY PROGRAMME	SUBJECT	SCORE (ECTS)
1	5th	Faculty of Technical Sciences (Department of Civil Engineering and Geodesy)	Novi Sad	Civil engineering	The modelling of the processes in the civil engineering	3
	3rd			Geodesy	The methods of optimization	6
2	3rd	Faculty of Civil Engineering	Belgrade	Civil engineering	Informational modelling of civil engineering objects	4
	5th				Operations research	4
				Geodesy	No subject	
3	5th	UNION NIKOLA TESLA	Belgrade	Civil engineering	Operational research	5
4		Faculty of Civil Engineering and Architecture	Niš	Civil engineering	No subject	
5	2nd	Faculty of Civil Engineering	Subotica	Civil engineering	The basics of programming	7
				Geodesy	No subject	
6		Faculty of Civil Engineering	Novi Pazar	Civil engineering	No subject	
7		Faculty of Civil Engineering	Podgorica	Civil engineering	No subject	
8		Faculty of applied science	Podgorica	Geodesy	No subject	

DRAFT VERSION

9	5th	Faculty of Architecture, Civil Engineering and Geodesy	Banja Luka	Civil engineering	The operational research in civil engineering	4
				Geodesy	No subject	
10		Faculty of Civil Engineering	Sarajevo	Civil engineering	No subject	
				Geodesy	No subject	
11	3rd	Faculty of Civil and Geodetic Engineering	Ljubljana	Civil engineering	Project Management	4
	1st			Geodesy	Observations analysis in surveying	4
12		Faculty of Civil Engineering	Zagreb	Civil engineering	No subject	
13		Faculty of Geodesy		Geodesy	No subject	
14	3rd	Faculty of Civil Engineering	Skopje	Civil engineering	Management in civil engineering	5

At the Department of Civil Engineering and Geodesy **in Novi Sad**, there is one subject, at the 5th year of studies (1st year od master studies) called *The modelling of the processes in the civil engineering.* This subject is mandatory and carries three ECTS. Also, at the geodesy study programme there is one subject called *The methods of optimization*. This subject is mandatory and carries six ECTS.

At the Faculty of Civil Engineering **in Belgrade**, there is one subject at the 3rd year of undergraduate (bachelor) studies called *Informational modelling of civil engineering objects* and one at the 5th year called *Operations research.* These subjects are optional and carries four ECTS.

At the Faculty of Civil Engineering **UNION Nikola Tesla**, there is one subject at the 5th year of studies (1st year of master studies) called *Operational Research.* This subject is optional and carries five ECTS.

At the Faculty of Civil Engineering and Architecture **in Niš**, there is no subject in curriculum with operations research involved.

At the Faculty of Civil Engineering **in Subotica**, there is one subject at the 2nd year of undergraduate (bachelor) studies called *The basics of programming.* This subject is optional and carries seven ECTS.

At the Faculty of Civil Engineering **in Novi Pazar**, there is no subject in curriculum with operations research involved.

At the Faculty of Civil Engineering **in Podgorica**, there is no subject in curriculum with operations research involved.

At the Faculty of Architecture, Civil Engineering and Geodesy **in Banja Luka**, there is one subject at the 5th year of studies (1st year of master studies) called *The operational research in civil engineering*. This subject is optional and wears four ECTS.

At the Faculty of Civil Engineering **in Sarajevo**, there is no subject in curriculum with operations research involved.

At the Faculty of Civil and Geodetic Engineering **in Ljubljana**, there is one subject at the 3rd year of undergraduate (bachelor) studies called *Project Management* and also *Observations analysis in surveying.* The first subject is optional and carries four ECTS. The second subject is mandatory and carries four ECTS.

At the Faculty of Civil Engineering **in Zagreb**, there is no subject in curriculum with operations research involved.

At the Faculty of Civil Engineering **in Skopje**, there is one subject at the 3rd year of studies (1st year of master studies) called *Management in civil engineering.* This subject is optional and carries five ECTS.

Except at the the Department for Civil Engineering and Geodesy (Faculty of Technical Sciences, University of Novi Sad), there are no subjects about operations research at the other faculties.

After analyzing the data at the Table 1, the conclusion is that at seven faculties of civil engineering in the former Yugoslav republics, subjects connected with operations research are implemented. Five faculties of Civil Engineering do not have subjects on the implementation of operations research. Faculties that have subjects are from: Novi Sad, Belgrade, Subotica, Ljubljana, Skopje, and Banja Luka. Two geodesy Faculties which have subjects are the Department of Civil Engineering and Geodesy in Novi Sad and the Faculty of Civil and Geodetic Engineering in Ljubljana.

Subjects are mostly optional for Civil Engineering, except for the faculty in Novi Sad. The years in which these subjects are included are the second and third in Skopje, Ljubljana, Belgrade, and Subotica. In Novi Sad,

DRAFT VERSION

Banja Luka, Belgrade and UNION Nikola Tesla, there are subjects at the master's studies. For geodesy studies at the Department of Civil Engineering and Geodesy in Novi Sad, the subject is obligatory and it is in the third year of studies, first year at the Faculty of Civil and Geodetic Engineering in Ljubljana.

3. CONCLUSION

Operations Research, also referred to as Management Science, is the application of a scientific approach to solving difficult resource allocation, management and engineering problems in order to make better decisions.

In this work the inclusion of operations research in curriculum of civil engineering and geodesy faculties is analyzed. Six countries from region are selected: Serbia, Bosnia and Herzegovina, Croatia, Montenegro, Republic of North Macedonia and Slovenia. In that six countries fourteen faculties are selected for research. After detailed analyze of curriculums the results shows that at five faculties of Civil Engineering do not have subjects on the implementation of operations research. Faculties that have the subjects are from: Novi Sad, Belgrade, Subotica, Ljubljana, Skopje and Banja Luka. Subjects are mostly optional, except for the Faculty of Technical Sciences in Novi Sad. Two geodesy Faculties which have subjects are the Department of Civil Engineering and Geodesy in Novi Sad and the Faculty of Civil and Geodetic Engineering in Ljubljana.

The years in which these subjects are included are the second and third in Skopje, Ljubljana, Belgrade, and Subotica. In Novi Sad, Banja Luka, Belgrade and UNION Nikola Tesla, there are the subjects at the master's studies. For geodesy studies at the Department of Civil Engineering and Geodesy in Novi Sad, the subject is obligatory and it is in the third year of studies, first at the Faculty of Civil and Geodetic Engineering in Ljubljana.

ACKNOWLEDGMENT

This research has been supported by the Ministry of Science, Technological Development and Innovation through Contract No. 451-03-136/2025-03/200156. (J.T., T.M.)

This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-137/2025-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project “Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad 2025” (No. 01-50/295). (V.M., M.B.)

LITERATURE

(2021). Retrieved May 31, 2025, from Medium: https://medium.com/data-science/why-operations-research-is-awesome-an-introduction-7a0b9e62b405

(2025). Retrieved May 31, 2025, from Stonybrook: https://www.stonybrook.edu/commcms/ams/graduate/or/

Assad, A. A. (2011). Profiles in Operations Research - Pioneers and Innovators. *International Series in Operations Research & Management Science*, 1-867.

Goodeve, C. (1948). Operational Research. *Nature, 161, (4089)*, 377-384.

Gupta, S. (2019). Retrieved May 31, 2025, from https://plopdo.com/2019/03/02/what-is-operations-research

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review, 106*(4), 620-630.

Kapadi, O. (2022). A Brief study of operation research. Retrieved May 31, 2025, from https://www.researchgate.net/publication/360964343_A_Brief_study_of_operation_research

Lownes, N. (2023). Operations Research in Civil and Environmental Engineering,. (O. E. resource, Ed.)

Rajgopal, J. (2025). (U. o. Pittsburgh, Ed.) Retrieved May 31, 2025, from https://sites.pitt.edu/~jrclass/or/webnotes/or-intro.html

Soori, M. A. (2023). Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics, 3*, 54-70. Retrieved from https://arxiv.org/html/2401.03244v1/#S2

University, M. (2025, May). Retrieved from https://uomustansiriyah.edu.iq/media/lectures/6/6_2020_02_04!07_54_32_AM.pdf

DRAFT VERSION

SIMULATION AND STOCHASTIC MODELS

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

SIMULATION MODEL OF ROBOT MOVEMENT ALONG A DEFINED PATH DURING PRE-FLIGHT INSPECTION OF AN A320 AIRCRAFT

DUNJA SAVIĆ[1], MARKO ĐOGATOVIĆ[2], LJUBIŠA VASOV[3]

[1] University of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, dunja.savic.student.23d005@sf.bg.ac.rs, 0009-0009-2022-7230
[2] University of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, m.djogatovic@sf.bg.ac.rs, 0000-0003-4419-0021
[3] University of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, lj.vasov@sf.bg.ac.rs, 0000-0001-5711-1471

Abstract: *This paper presents a simulation of a robot's movement along a predefined path during pre-flight aircraft inspection. The objective of the study is to assess the robot's ability to accurately follow the planned trajectory. The simulation model was developed in MATLAB/Simulink, and the robot's motion was evaluated based on parameters such as position deviation, orientation angle, and linear velocity. The results demonstrate that the integration of the applied technologies enables high-precision path tracking, highlighting their potential for autonomous inspection tasks in aviation and related fields.*

Keywords: *pre-flight inspection, mobile robot, path tracking, simulation modeling.*

1. INTRODUCTION

Pre-flight inspection (Yasuda, Cappabianco, Martins, & Gripp, 2022) represents a fundamental aspect of aircraft maintenance, carried out immediately prior to each flight to ensure that the aircraft is in a safe and airworthy condition. The inspection procedure consists of four main stages. The first step involves a detailed visual examination of the aircraft's exterior, aiming to identify any damage, deformations, or other visible issues affecting the aircraft structure. The condition of control surfaces located on the wings and tail is also inspected for potential damage, dirt accumulation, or ice.

The second step entails a system check, including testing of the aircraft's essential systems such as the hydraulic and pneumatic systems. The functionality of cockpit instruments and communication devices is also verified. The third step is an inspection of the aircraft's interior, which includes checking safety equipment (seat belts, life vests), inspecting the cabin, and securing cargo. The fourth step involves reviewing and verifying the required documentation, such as the flight plan, technical data, and checklists. Pre-flight inspections are typically performed by flight crew members or qualified maintenance technicians.

Since the pre-flight inspection is carried out by humans, the human factor is an unavoidable element. Fatigue and stress can lead to decreased concentration and reduced attention, potentially resulting in oversights. Human errors in aviation can cause serious consequences, endangering the safety of passengers, crew, and cargo. Additionally, such errors can lead to substantial financial costs due to physical damage to the aircraft.

The application of modern technologies helps reduce the probability of errors and increases safety levels. Automation aims to simplify the performance of specific tasks for personnel, enhancing accuracy and precision. The use of robots can further improve precision and reliability through automation processes. A robot can perform a pre-flight inspection of an aircraft. The basic concept involves a robot moving along a predefined path around the aircraft, capturing images and videos of critical points on the aircraft's fuselage. These critical points include various sensors, pitot-static system openings, control surfaces, and more. The captain, first officer, and maintenance engineers would be able to monitor the photos and videos taken by the robot via a display screen. Software installation would enable automatic recognition and detection of any changes on the aircraft's fuselage. All images and videos would be stored within the software's memory, making them readily accessible at any time.

This paper presents a simulation of a robot's movement along a predefined path during the pre-flight inspection of an Airbus A320-200 aircraft. The purpose of the simulation is to ensure precise tracking of the robot's predefined path. To address the problem of robot path-following, the programs Matlab and Simulink were used.

DRAFT VERSION

2. PROBLEM FORMULATION

Mobile robots are advanced autonomous systems equipped with servo motors and sensors, capable of operating in various environments. One of the primary challenges they face is avoiding collisions with obstacles in the workspace. This issue is addressed through various navigation methods, including wheel rotation measurements, accelerometers, gyroscopes, electromagnetic beacons, and optical or magnetic signal tracking. Also, the robot is equipped with software based on the Robot Operating System (ROS), which enables remote control, navigation along a predefined path, and fully autonomous operation. Automating the pre-flight inspection process with robotic systems significantly reduces the need for human intervention and accelerates the inspection procedure. The robot is outfitted with various damage detection sensors and high-resolution cameras to ensure precise analysis of the aircraft's surface. Additionally, it incorporates autonomous navigation technologies that allow it to move around the aircraft and inspect all critical areas.

The data collected during the inspection is analyzed to detect potential issues, enabling prompt decision-making and timely execution of necessary repairs. The algorithm used to simulate the robot's movement along a predefined path consists of several key steps. The objective of the simulation is to ensure accurate tracking of the robot's designated path. The robot is equipped with a range of sensors and guidance systems, along with cameras that scan the aircraft's surface and detect potential anomalies.

Differential drive is the most common motion system in mobile robotics. It is commonly applied in robots with two wheels mounted on a shared rotating axis, each driven by its own servomotor (Figure 1). The identical methodology remains applicable to four-wheeled robotic platforms. Kinematic model for a moving robot following a path that moves with constant velocity is given by:

$$\begin{bmatrix} \dot{x}(t) \\ \dot{y}(t) \\ \dot{\varphi}(t) \end{bmatrix} = \begin{bmatrix} v \cdot \cos\varphi(t) \\ v \cdot \sin\varphi(t) \\ \omega(t) \end{bmatrix}, \tag{1}$$

where φ is the steering direction (rotation angle), ω is angular velocity and v is constant linear velocity (Crenganis, Biris, & Girjob, 2021). We assume that the robot is rotating in place, with its wheels spinning in opposite directions at equal speeds. As a result, the instantaneous center of rotation (ICR) lies at the midpoint between the two wheels. The angular velocity (ω) of the robot is determined by the wheel speeds and the distance between the wheels.

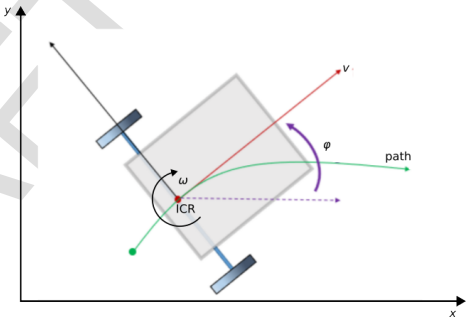


Figure 1: Moving robot with a differential drive

In the simulation model robot uses a Proportional-Integral-Derivative (PID) (Ang, Chong, & Li, 2005) controller for control of rotation angle and a Pure Pursuit controller for trajectory tracking. Block diagram of PID controller is given on Figure 2.

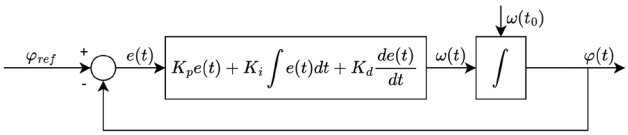


Figure 2: Block diagram of PID controller

The reference heading angle φ_{ref} indicates the direction toward the next waypoint in the robot's trajectory.

DRAFT VERSION

The Pure Pursuit controller (Macenski, Singh, Martin, & Gines, 2023) is implemented as part of the robot's control system to manage its movement along a predefined path. This controller dynamically adjusts the robot's speed and steering based on its deviation from the path and the distance to the next target point.

The simulation also involves defining the exact path of movement, including the coordinates of the key positions where the robot pauses to inspect the aircraft's fuselage and plating for possible damage. For tracking robot movement, two important sensors are: position and velocity sensors. Modern solutions for positioning (satellite, Ultra-Wideband, LiDAR etc.) and velocity sensing (encoders) provide highly accurate position (up to cm to dm-level resolution) and speed (up to cm/s-level resolution) estimation with very low noise levels. Also, latency for those sensors is low. The Airbus A320-200 model is used as a reference for the simulation. Inspection checkpoints are defined according to the aircraft maintenance manual, and based on this documentation, the robot's trajectory and corresponding coordinates as illustrated in Figure 3 were established. The aircraft is considered to be stationed at the gate at the airport. In the paper, we assume that the aircraft inspection lighting is optimal, providing sufficient intensity, appropriate diffusion, and glare control to ensure visibility of details on the aircraft.

The performance of the system is evaluated by assessing tracking accuracy, the time required to complete the inspection, and any deviations or issues encountered during movement. The analysis of simulation results enables further improvements by tuning the control system parameters based on the collected data, ultimately enhancing the robot's efficiency and precision. Simulation results are generated to present the outcomes and highlight potential areas for improvement.

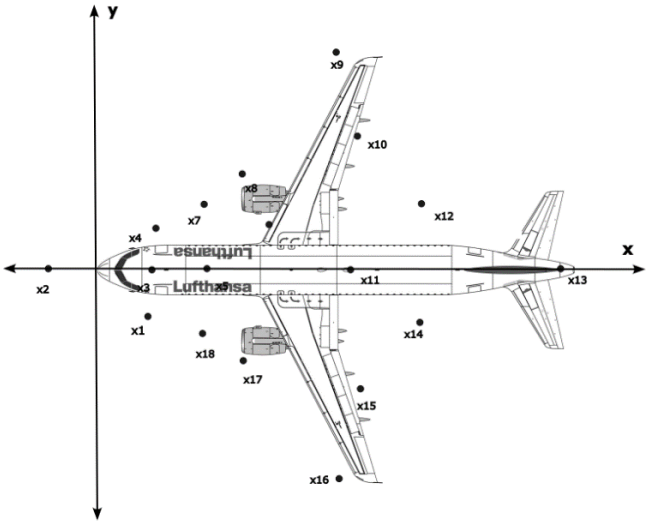


Figure 3: Top-down airplane projection with path of the robot in the coordinate system

3. SIMULATION MODEL

Figure 4 illustrates the Simulink-based simulation model used for testing and analyzing the robot's path-following performance. The implemented model uses a hybrid discrete-event/continuous simulation approach.

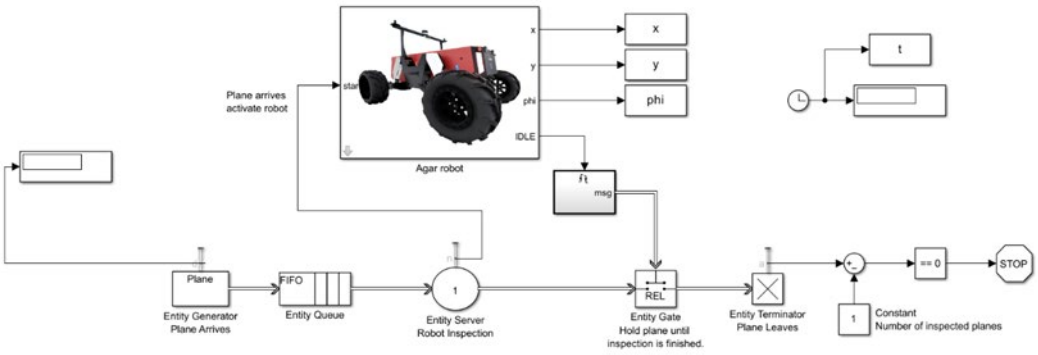


Figure 4: Simulink-based simulation model

The input parameters include the coordinates that define the robot's movement path, with the starting point marked as 1 and an initial orientation angle of 180°. An entity generator simulates the arrival of an aircraft, and the arrived entity serves as the trigger for initiating the robot's operation. The simulation incorporates an Entity Queue block, used to model queuing behavior. The queue operates on a First-In, First-Out (FIFO) principle. An Entity Server block is implemented to simulate the processing of each entity. It receives aircraft entities, processes them, and forwards them through the system. The robot is configured to inspect one aircraft at a time. Robot movement is modeled using a state machine, a mathematical model that represents system behavior through a finite number of defined states and is shown in Figure 5. The system exists in one state at any given moment and can transition to another based on incoming signals or events. Each transition may be subject to specific conditions and can trigger actions either during the transition or while the system remains in a particular state. The inputs to the state machine are the x and y coordinates that define the robot's linear motion, the angle φ (phi) that indicates the robot's orientation toward the next target point, and the start signal, which initiates robot operation. The outputs include the current state of the robot, the value n that represents the current point of inspection, rotate and move signals which correspond to rotational and linear movements between points, and the idle signal indicating that the inspection task is complete. The rotate signal is directed to the rotation block, while the move signal is passed to the motion block to control the robot's physical movement.

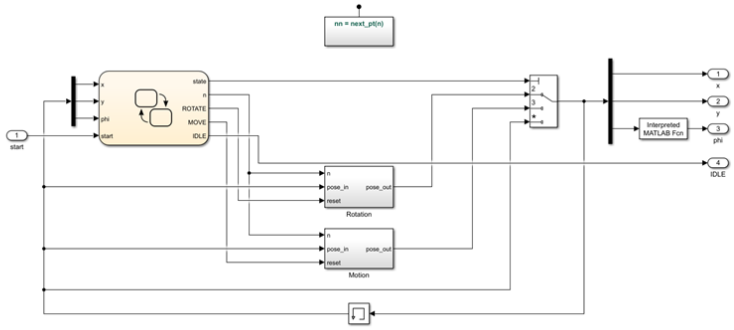


Figure 5: Robot movement modeling block

The rotation block in Figure 6 receives three inputs (n, pose in, and reset) and produces one output (pose out). Here, n represents the current point, pose in is the robot's current position, and reset restarts the simulation when set to 1. The pose out represents the robot's position after the rotation has been completed. Rotation is based on the current point (pt1) and the next point (pt2), whose coordinates are retrieved from memory. Using this data, the system calculates the desired heading angle. A PID controller regulates the wheel rotation speed, enabling the robot to align correctly with the next waypoint. This control is essential since the robot steers by adjusting wheel torque. The controller minimizes the error between the current and target orientation, driving the wheels accordingly. The angular velocity is integrated to compute the new position. In this closed-loop system, pose out is fed back into the controller, ensuring continuous adjustment and accurate path tracking.

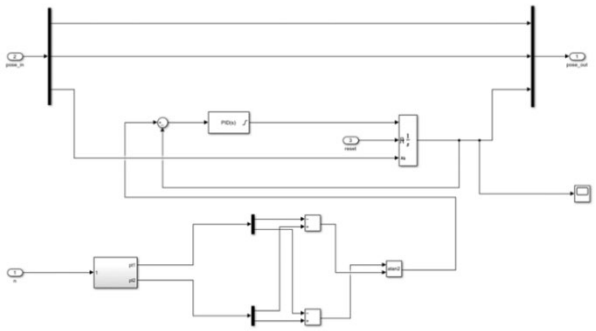


Figure 6: Rotation block

The motion block, shown in Figure 7, has the same inputs and output as the rotation block: n, pose in, reset, and pose out. It receives the current point as input, and both the current and next point coordinates are retrieved from memory. These are sent to the Pure Pursuit controller, an advanced control algorithm implemented in Simulink. The controller uses parameters such as angular velocity (ω) and linear speed (v) to steer the robot from its current to the next point. It ensures precise motion execution according to the instructions from the

DRAFT

state machine. When the robot returns to the starting point, an idle signal is triggered, indicating that the task is complete.

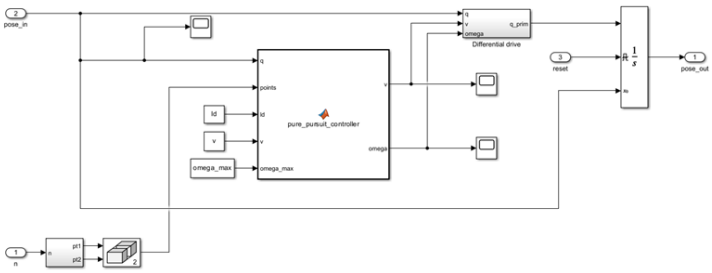


Figure 7: Motion block

4. SIMULATION RESULTS

This paper presents a simulation of a robot's movement along a predefined path during the pre-flight inspection of an aircraft. The simulation is based on the inspection trajectory around an Airbus A320, illustrated in Figure 2. Table 1 presents the input parameters of the applied model. Some of the input parameters are taken from the website (AGAR, n.d.). Distance tolerance refers to the maximum permissible deviation from the actual position of a point along the predefined path. Angle tolerance defines the acceptable angular deviation when rotating towards the next point. Lookahead distance indicates how far ahead the robot "looks" along the path in order to determine the appropriate control command.

Table 1: Input parameters

Parameter group	Parameter name	Value
Path, starting position and orientation	Initial angle (°)	180
	Distance tolerance (m)	0.05
	Angle tolerance (°)	0.1
Robot parameters	Linear velocity (m/s)	1.8
	Distance between center of wheels (m)	1.219
	Scanning time (s)	2
Pure pursuit controller	Lookahead distance(m)	0.5
PID controller	P gain (K_p)	10
	I gain (K_i)	20
	D gain (K_d)	-0.2

Using the input parameters defined in the previous section, the robot's movement was simulated in Simulink, resulting in a trajectory consisting of 18 waypoints. Since the robot rotates in place to orient itself toward the next waypoint, its deviation from the linear path remains minimal. The robot's arrival at a waypoint is determined by the maximal Euclidean distance from the waypoint. The complete path followed by the robot is visualized in the first graph, shown in Figure 8. Figure 9 shows the robot's motion dynamics. Graph a) presents the linear velocity along the x and y axes, while graph b) illustrates changes in orientation angle as the robot follows the predefined path. Since we are simulating inspection process for only one aircraft simulation execution time and time for inspection is same with value of 120.6 s.

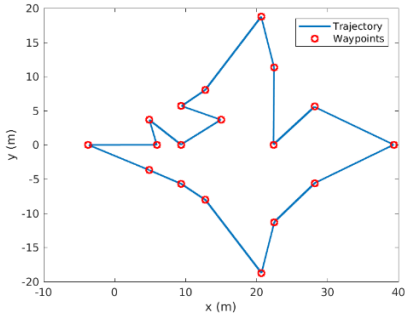


Figure 8: Simulated robot trajectory

DRAFT VERSION

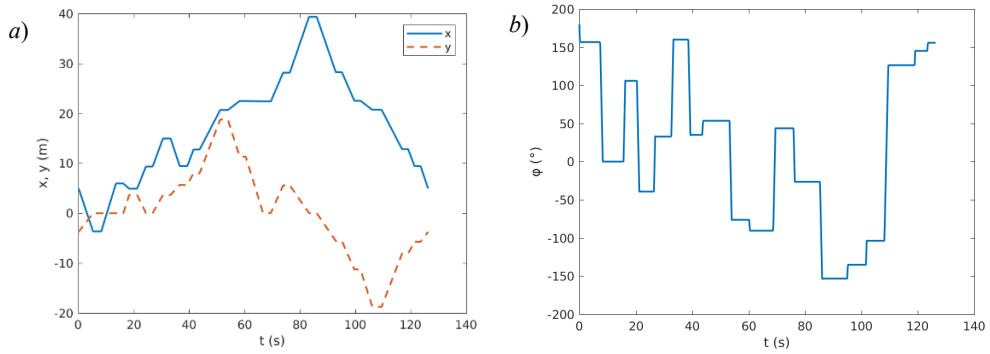


Figure 9: Simulated robot a) position (x, y), and b) rotation angle (φ)

This simulation not only demonstrates the robot's ability to follow a precise path but also allows for the evaluation of its performance in terms of path accuracy, efficiency, and the time required to complete the inspection. The results indicate that the robot can effectively navigate the defined trajectory, ensuring thorough, timely, and efficient pre-flight inspections.

5. CONCLUSION

This paper presented a simulation of autonomous robot movement along a predefined path during pre-flight aircraft inspection, using the Airbus A320 as a reference. The study highlights the crucial role of pre-flight inspections in ensuring flight safety and operational efficiency by enabling early detection of potential technical issues. The simulation demonstrated that the implemented control strategy significantly enhances path-tracking precision and reduces inspection time. These improvements directly contribute to increased safety and reduced operational costs. Each aircraft type follows a predefined sequence of visual inspection points, determined by its specific design and equipment. These points can be appropriate coordinates, enabling accurate camera positioning by the robotic system, whose camera can also rotate to capture various angles of the aircraft. While comparative data are currently lacking, using traditional manual inspections—which typically take between 15 and 30 minutes—as a benchmark would help validate this approach. Beyond time savings, automation ensures procedural consistency, enables defect detection, and allows systematic data recording. To further improve system performance, future work should focus on fine-tuning control parameters and integrating advanced control methods such as Model Predictive Control (MPC) or adaptive algorithms. Enhancing sensor accuracy and update frequency, applying sensor fusion techniques, and using advanced path-planning algorithms can further improve accuracy and efficiency. Additionally, real-time speed adjustment based on environmental complexity may reduce task duration. Continuous simulation and validation remain essential for iterative improvement, ensuring the robot consistently meets the demands of high-precision and time-efficient aircraft inspections.

LITERATURE

Ang, K. H., Chong, G., & Li, Y. (2005). PID control system analysis, design, and technology. *IEEE Transactions on Control Systems Technology, 13(4), 559-579.* https://doi.org/10.1109/TCST.2005.847331

Crenganis, M., Biris, C., & Girjob, C. (2021). Mechatronic Design of a Four-Wheel drive mobile robot and differential steering. *MATEC web of conferences, 343, 08003.* https://doi.org/10.1051/matecconf/202134308003

Macenski, S., Singh, S., Ginés, J., & Martín, F. (2023). Regulated pure pursuit for robot path tracking. *Autonomous Robots,47, 685–694.* https://doi.org/10.1007/s10514-023-10097-6

Yasuda, Y. D., Cappabianco, F. A., Martins, L. E., & Gripp, J. A. (2022). Aircraft visual inspection: A systematic literature review. *Computers in Industry,141, 103695.* https://doi.org/10.1016/j.compind.2022.103695

AGAR. (n.d.). *Features*. Retrieved June 20, 2024, from https://agarrobot.com/features

DRAFT VERSION

SOFT COMPUTING

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

VARIANTS OF THE MINIMUM ERROR THRESHOLDING METHOD FOR IMAGE SEGMENTATION

DEJAN ĆEBIĆ[1], NEBOJŠA M. RALEVIĆ[2], NATAŠA MILOSAVLJEVIĆ[3]

[1] University of Belgrade, Faculty of Mining and Geology, Belgrade, dejan.cebic@rgf.bg.ac.rs, 0000-0003-0261-0681
[2] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, nralevic@uns.ac.rs, 0000-0002-3825-9822
[3] University of Belgrade, Faculty of Agriculture, Belgrade, natasam@agrif.bg.ac.rs, 0000-0003-4056-089X

Abstract: *Image segmentation is an important problem because it enables meaningful partitioning of an image and extraction of relevant structures and features, facilitating further processing and decision-making. Various approaches have been proposed to address this problem, with one of the simplest and most commonly used being segmentation based on thresholding. This research relies on the classic and well-known Minimum Error Thresholding (MET) method, specifically employing a simple idea of linearly combining the MET method with its median-based extension, creating three modifications of MET method in order to improve thresholding results in terms of quality coefficients such as Mean Square Error (MSE) and Structural Similarity Index Measure (SSIM). Several experiments on some standard gray-level images have demonstrated that the proposed modifications improve segmentation performance, as confirmed by MSE and SSIM indices.*

Keywords: *image segmentation, image thresholding, minimum error thresholding (MET) method, median-based methods*

DRAFT VERSION

1. INTRODUCTION

Image segmentation is one of the most important problems in image analysis since it is required preprocessing step for important computer vision tasks such as feature extraction and object recognition. Good segmentation enables more precise introspection of specific regions within an image. It can also enhance algorithm performance by reducing the amount of data to be processed.

A numerous techniques have been developed for image segmentation. Some of the most commonly used methods are traditional approaches such as threshold-based methods, edge-based methods, clustering-based methods, graph-based methods, and region-based methods, etc. In addition to these, more recent and advanced techniques have been developed, including deep learning-based methods (e.g., U-Net, DeepLab), level set methods, superpixel-based approaches, etc. (see, for example (Brar et al., 2025; Minaee et al., 2021)). Although the latter techniques have become dominant in recent years, thresholding remains popular due to its simplicity and effectiveness, and because of these properties, it is often incorporated into more complex algorithms (Castiglione et al., 2025; Harrak et al., 2025; S. W. Kim et al., 2025).

Threshold methods were initially created to convert grayscale images into binary images, with the aim of identification of the background and foreground (object) in the image. This is accomplished by setting a threshold value, with each pixel being assigned to one of the two segments based on whether its intensity is greater or less than the threshold value. Such segmentation process where a single threshold value is determined is referred to as binary thresholding. When the segmentation task uses two or more threshold values, it is known as multilevel thresholding. Although simple and fast, thresholding methods have several limitations, as they often perform poorly on noisy images, low contrast and uneven illumination.

The primary focus for many researches is to find optimal threshold values that accurately separates images into two or multiple segments. Perhaps the best-known thresholding method is the Otsu method (Otsu et al., 1975), which is implemented in many image processing software and computer vision applications. Some other widely used thresholding methods include: Kapur's method (Kapur et al., 1985), method based on Tsallis entropies (De Albuquerque et al., 2004), minimum cross entropy thresholding method (Li and C. Lee, 1993), Kittler and Illingworth's minimum error thresholding method (MET) (Kittler and Illingworth, 1986), etc. In the case of multiple thresholding determination, as the number of thresholds increases, the mentioned methods become more complicated in the sense of increasing the computational time for threshold values calculation. To overcome this difficulty, many modifications and hybrid variants of these methods have been combined with the

various optimization techniques such as nature-inspired algorithms (see for example (Abd El Aziz et al., 2017; Agrawal et al., 2013; Hancer, 2020; Horng, 2011; Houssein et al., 2021)).

It is hard to declare some of the mentioned thresholding methods as the best since the goodness of thresholding depends on specific properties of the image and the application requirements. However, Sezgin and Sankur conducted an extensive study (Sezgin and Sankur, 2004) and ranked the MET method as the best among the methods developed up to that point, based on several evaluation parameters. Therefore, considering that the MET method is still used in its basic form in recent studies (Harrak et al., 2025; S. W. Kim et al., 2025), this paper will primarily focus on the MET method and exploring ways to enhance its performance. The next section provides a brief recall of the MET method and its median-based modification. Section 3 introduces three new variants of MET-type thresholding methods. The experimental results based on the test images and the comparison of the new variants of MET method are presented in Section 4. Finally, the last section contains some concluding remarks and comments.

2. THE MET METHOD AND ITS MEDIAN-BASED MODIFICATION

Suppose that a given gray-level image contains N pixels, and each pixel is represented by its intensity x_j, $j = 1,...,N$, where 0 and T are the smallest and the largest possible intensities, respectively ($T = 255$ for an $8-$bit gray-level image). The associated histogram can be created by counting the number of pixels at each intensity level and plot these counts on a graph.

Let f_i denote the frequency of occurrence of level i in the image, and $p_i = f_i/N$ denote the probability of occurrence of level i. It is clear that

$$\sum_{i=0}^{T} f_i = N \quad \text{and} \quad \sum_{i=0}^{T} p_i = 1.$$

Binary thresholding methods generate a threshold value t that splits the set of pixels into two classes $C_1(t)$ and $C_2(t)$ such that $C_1(t) = \{j|\ 0 \le x_j \le t,\ 1 \le j \le N\}$ and $C_2(t) = \{j|\ t < x_j \le T,\ 1 \le j \le N\}$.

The MET method selects an optimal threshold value t^* that minimizes the objective function

$$J(t) = \omega_1(t) \cdot \log \frac{\sigma_1(t)}{\omega_1(t)} + \omega_2(t) \cdot \log \frac{\sigma_2(t)}{\omega_2(t)}, \tag{1}$$

where $\omega_1(t)$ and $\omega_2(t)$ denote cumulative probabilities, while $\sigma_1(t)$ and $\sigma_2(t)$ are variances of the classes $C_1(t)$ and $C_2(t)$, which means that

$$\omega_1(t) = \sum_{i=0}^{t} p_i, \quad \omega_2(t) = \sum_{i=t+1}^{T} p_i = 1 - \omega_1(t),$$

$$\sigma_1^2(t) = \frac{\sum_{i=0}^{t} f_i (i-\mu_1(t))^2}{\sum_{i=0}^{t} f_i} = \frac{1}{\omega_1(t)} \sum_{i=0}^{t} p_i (i-\mu_1(t))^2, \tag{2}$$

$$\sigma_2^2(t) = \frac{\sum_{i=t+1}^{T} f_i (i-\mu_2(t))^2}{\sum_{i=t+1}^{T} f_i} = \frac{1}{\omega_2(t)} \sum_{i=t+1}^{T} p_i (i-\mu_2(t))^2, \tag{3}$$

where $\mu_1(t)$ and $\mu_2(t)$ are the means of $C_1(t)$ and $C_2(t)$, i.e.

$$\mu_1(t) = \frac{\sum_{i=0}^{t} f_i \cdot i}{\sum_{i=0}^{t} f_i} = \frac{1}{\omega_1(t)} \sum_{i=0}^{t} p_i \cdot i, \tag{4}$$

$$\mu_2(t) = \frac{\sum_{i=t+1}^{T} f_i \cdot i}{\sum_{i=t+1}^{T} f_i} = \frac{1}{\omega_2(t)} \sum_{i=t+1}^{T} p_i \cdot i. \tag{5}$$

Kurita et al. showed in (Kurita et al., 1992) that the MET method can be derived from maximisation of log-likelihoods based on mixtures of Gaussian distributions.

In the case of multi-level thresholding, the MET method can be adapted to find certain number of classes. For example, if the number of classes is $K > 2$, then the aim of every method is to find $K-1$ thresholding

DRAFT VERSION

values $\mathbf{t} = (t_1, t_2, ..., t_{K-1})$ for defining $C_1, C_2, ..., C_K$. Therefore, the objective function (1) can be rewritten and the appropriate optimization problem can be viewed as

$$\mathbf{t}^* = \underset{\mathbf{t}}{\operatorname{argmin}} \sum_{k=1}^{K} \left\{ \omega_k(\mathbf{t}) \log \frac{\sigma_k(\mathbf{t})}{\omega_k(\mathbf{t})} \right\}. \tag{6}$$

Xue and Titterington proposed in (Xue and Titterington, 2011) an alternative option, namely a modified MET method, which uses the median for location parameters instead of the means defined by (4) and (5). Analogously to the MET method, they presented a similar theoretical median-based approach using the mixtures of Laplace instead of Gaussian distribution. Accordingly, when the median is used instead of the mean, equations (2) and (3) for dispersion parameters are typically not used; instead, the mean absolute deviation from the median (MAD) is employed. Hence, in the case of the binary thresholding, MAD values are given by

$$\mathrm{MAD}_1(t) = \frac{\sum_{i=0}^{t} f_i |i - m_1(t)|}{\sum_{i=0}^{t} f_i} = \frac{1}{\omega_1(t)} \sum_{i=0}^{t} p_i |i - m_1(t)|,$$

$$\mathrm{MAD}_2(t) = \frac{\sum_{i=t+1}^{T} f_i |i - m_2(t)|}{\sum_{i=t+1}^{T} f_i} = \frac{1}{\omega_2(t)} \sum_{i=t+1}^{T} p_i |i - m_2(t)|,$$

where $m_1(t)$ and $m_2(t)$ are the sample medians of $C_1(t)$ and $C_2(t)$. Therefore, the median-based MET method (denoted by medMET) gets the following form

$$t^* = \underset{t}{\operatorname{argmin}} \Big(\omega_1(t) \log \frac{\mathrm{MAD}_1(t)}{\omega_1(t)} + \omega_2(t) \log \frac{\mathrm{MAD}_2(t)}{\omega_2(t)} \Big), \tag{7}$$

The corresponding multi-level thresholding median-based MET variant is then

$$\mathbf{t}^* = \underset{\mathbf{t}}{\operatorname{argmin}} \sum_{k=1}^{K} \left\{ \omega_k(\mathbf{t}) \log \frac{\mathrm{MAD}_k(\mathbf{t})}{\omega_k(\mathbf{t})} \right\}. \tag{8}$$

The motivation for introducing the median variant of the MET method was its better performance when the distributions of the classes C_k were skew or heavy-tailed. In such cases, it has been shown that the median is more robust location estimator than the mean, so then the median-based modification of the MET method is more suitable for the thresholds determination.

Figure 1: Image *Girlface*

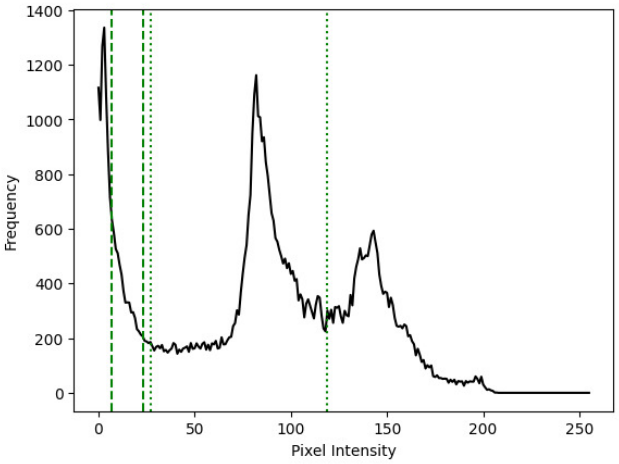


Figure 2: Histogram of *Girlface* and threshold values

For instance, by applying the three-level thresholding MET and the modified median-based MET method on the 256×256 grey-level image given as Fig. 1, according to the results displayed in the associated histogram (Fig. 2), median-based MET method acts better than MET. Namely, for $K = 3$, the optimization rules given by (6) and (8) select the following threshold values: $\mathbf{t}^* = (t_1^*, t_2^*) = (7, 23)$ for MET method and $\mathbf{t}^* = (t_1^*, t_2^*) = (27, 119)$ for medMET, and these threshold values are displayed within histogram as dashed lines and dotted lines, respectively. Obviously, the histogram has three peaks and the MET thresholds lie close to the same valley, while medMET thresholds are better distributed between three peaks (medMET finds the valleys between the peaks) which indicate better separation of classes.

DRAFT VERSION

3. NEW VARIANTS OF MET-TYPE THRESHODLING METHODS

In this section, three new modified versions of the MET method are introduced. Actually, the new variants are combinations of the MET and medMET methods, aiming to preserve the good properties identified in (Sezgin and Sankur, 2004) nad (Xue and Titterington, 2011) while potentially improving the performance of the initial approaches. All three variants are based on the arithmetic means of the components of the classical MET (1) and its median modification (7). In each of the three cases, the main idea is to take both the mean and the median into consideration simultaneously.

The first variant (denoted by METvar1) is a simple arithmetic mean of the objective functions (1) and (7)

$$t^* = \underset{t}{\operatorname{argmin}}\Big(\frac{1}{2}\Big[\omega_1(t)\log\frac{\sigma_1(t)}{\omega_1(t)} + \omega_2(t)\log\frac{\sigma_2(t)}{\omega_2(t)}\Big] + \frac{1}{2}\Big[\omega_1(t)\log\frac{\mathrm{MAD}_1(t)}{\omega_1(t)} + \omega_2(t)\log\frac{\mathrm{MAD}_2(t)}{\omega_2(t)}\Big]\Big).$$

Note that the objective function can be rewritten as

$$J(t) = \omega_1(t)\log\frac{\sqrt{\sigma_1(t)\mathrm{MAD}_1(t)}}{\omega_1(t)} + \omega_2(t)\log\frac{\sqrt{\sigma_2(t)\mathrm{MAD}_2(t)}}{\omega_2(t)} \tag{9}$$

where essentially the numerator represents the geometric mean of the standard deviation and the mean absolute deviation from the median.

If the arithmetic mean is used instead, we get another method (METver2) and the corresponding objective function

$$J(t) = \omega_1(t)\log\frac{s_1(t)}{\omega_1(t)} + \omega_2(t)\log\frac{s_2(t)}{\omega_2(t)} \tag{10}$$

where $s_1(t) = 0.5(\sigma_1(t) + \mathrm{MAD}_1(t))$ and $s_2(t) = 0.5(\sigma_2(t) + \mathrm{MAD}_2(t))$.

The third version METver3 relies on the basic mean value with the aritmetic mean and the median substituted in $\mathrm{MAD}_1(t)$ and $\mathrm{MAD}_2(t)$. Hence, the objective function has a form

$$J(t) = \omega_1(t)\log\frac{\mathrm{MADmod}_1(t)}{\omega_1(t)} + \omega_2(t)\log\frac{\mathrm{MADmod}_2(t)}{\omega_2(t)}, \tag{11}$$

where $\mathrm{MADmod}_1(t)$ and $\mathrm{MADmod}_2(t)$ are the values calculated using modified $\mathrm{MAD}(t)$ expressions such that

$$\mathrm{MADmod}_1(t) = \frac{1}{\omega_1(t)}\sum_{i=0}^{t} p_i|i - r_1(t)| \quad \text{and} \quad \mathrm{MADmod}_2(t) = \frac{1}{\omega_2(t)}\sum_{i=t+1}^{T} p_i|i - r_2(t)|,$$

for $r_1(t)$ and $r_2(t)$ given by $r_1(t) = 0.5\cdot(\mu_1(t) + m_1(t))$ and $r_2(t) = 0.5\cdot(\mu_2(t) + m_2(t))$.

Presented formulas (9), (10) and (11) are designed for determining a single threshold, and they also have extended analogous versions for multi-level thresholding as follows

$$\mathbf{t}^* = \underset{\mathbf{t}}{\operatorname{argmin}}\sum_{k=1}^{K}\Big\{\omega_k(\mathbf{t})\log\frac{\sqrt{\sigma_k(\mathbf{t})\mathrm{MAD}_k(\mathbf{t})}}{\omega_k(\mathbf{t})}\Big\}, \quad \mathbf{t}^* = \underset{\mathbf{t}}{\operatorname{argmin}}\sum_{k=1}^{K}\Big\{\omega_k(\mathbf{t})\log\frac{s_k(\mathbf{t})}{\omega_k(\mathbf{t})}\Big\} \quad \text{and}$$

$$\mathbf{t}^* = \underset{\mathbf{t}}{\operatorname{argmin}}\sum_{k=1}^{K}\Big\{\omega_k(\mathbf{t})\log\frac{\mathrm{MADmod}_k(\mathbf{t})}{\omega_k(\mathbf{t})}\Big\},$$

respectively.

4. EXPERIMENTAL RESULTS AND COMPARISON

To illustrate the performance of the new variants of the MET method, three additional standard test gray-style 256×256 images are employed. These images are presented in Figures 3-5 along with their associated histograms.

Despite the fact that it is difficult to designate any of the existing methods as the best due to the effectiveness of thresholding depends on the segmentation purposes, there are several commonly used thresholding quality indices for comparing methods. Hence, to evaluate the performance of the described methods and the segmentation accuracy of the experimental results, this research relies on two measures.

The first one is the Mean Square Error (MSE) (Sara et al., 2019), given by

$$\mathrm{MSE} = \frac{1}{m\cdot n}\sum_{i=1}^{m}\sum_{j=1}^{n}\big(I(i,j) - S(i,j)\big)^2,$$

DRAFT VERSION

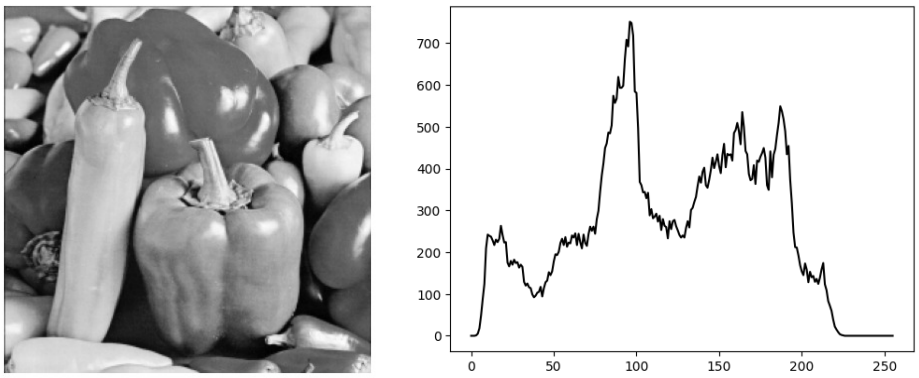


Figure 3: Image *Peppers* and its histogram

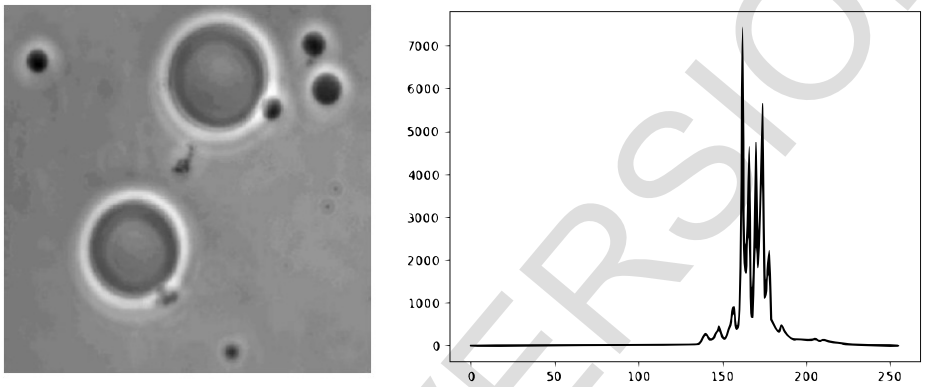


Figure 4: Image *Polymersomes* and its histogram

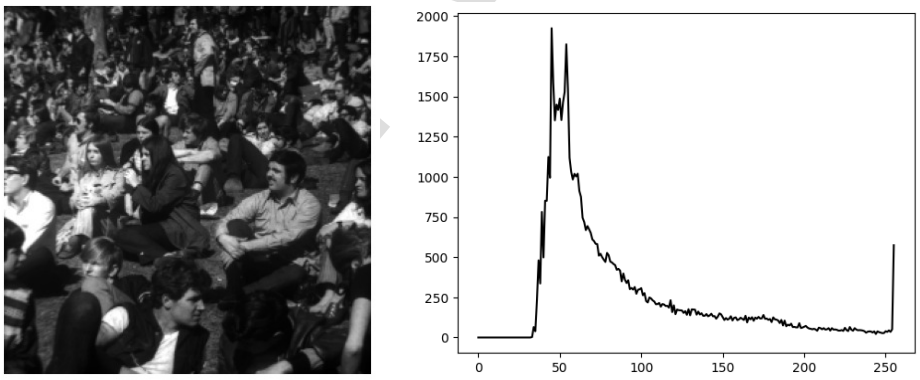


Figure 5: Image *Crowd* and its histogram

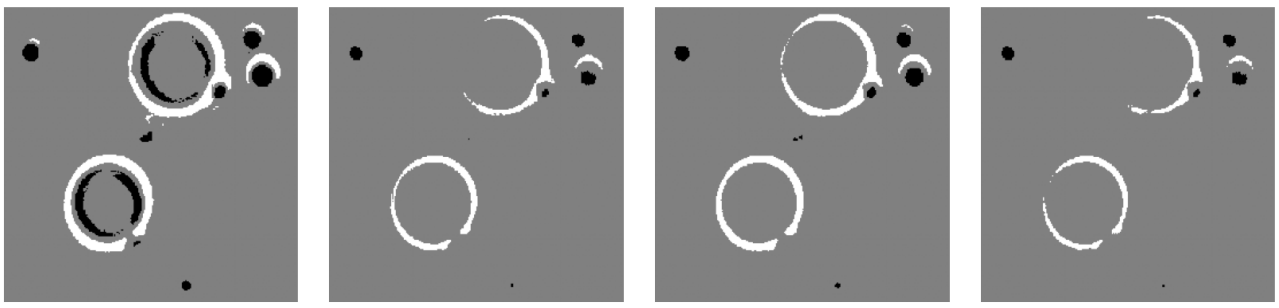


Figure 6: Examples of the segmented *Polymersomes* image using three-level thresholding methods (from left to right: MET, medMET, METver1, METver3)

DRAFT VERSION

where m and n represent the image dimensions, $I(i,j)$ and $S(i,j)$ are the intensities of pixel (i,j) in the original and segmented image, respectively. Lower MSE values indicate better segmentation results. Note that similar quality index PSNR (Peak Signal to Noise Ratio (Huynh-Thu and Ghanbari, 2008)) can be derived from MSE since $\text{PSNR} = 20\log_{10}(255/\sqrt{\text{MSE}})$.

Another quality measure is the Structural Similarity Index Measure (SSIM) (Wang et al., 2004), which quantifies the similarity between the original and segmented images in the following way:

$$\text{SSIM}(I,S) = \frac{(2\mu_I\mu_S + c_1)(2\sigma_{I,S} + c_2)}{(\mu_I^2 + \mu_S^2 + c_1)(\sigma_I^2 + \sigma_S^2 + c_2)}.$$

Here, μ_I and μ_S are the mean intensities of the original and segmented images, σ_I and σ_S are the standard deviations, while $\sigma_{I,S}$ denotes the covariance between I and S, and c_1 and c_2 are constants to stabilize the division when the denominator is weak (the constants are commonly set to $c_1 = 0.0065025$ and $c_2 = 0.0585225$).

Table 1: Girlface (two-level thresholding)

Method	Threshold t^*	MSE	SSIM
MET	19	19935.35	0.96341
medMET	14	21243.92	0.96079
METver1	17	20416.79	0.96246
METver2	17	20416.79	0.96246
METver3	15	20953.42	0.96139

Table 2: Peppers (two-level thresholding)

Method	Threshold t^*	MSE	SSIM
MET	26	17432.58	0.96727
medMET	135	7505.38	0.98831
METver1	132	7458.78	0.98827
METver2	132	7458.78	0.98827
METver3	138	7584.59	0.98823

Table 3: Polymersomes (two-level thresholding)

Method	Threshold t^*	MSE	SSIM
MET	198	27146.20	0.94178
medMET	205	27691.48	0.94017
METver1	200	27284.13	0.94137
METver2	200	27284.13	0.94137
METver3	210	28083.60	0.93895

Table 4: Crowd (two-level thresholding)

Method	Threshold t^*	MSE	SSIM
MET	91	6489.86	0.99349
medMET	96	6081.68	0.99364
METver1	92	6392.61	0.99354
METver2	92	6392.61	0.99354
METver3	92	6392.61	0.99354

Table 5: Girlface (three-level thresholding)

Method	(t_1^*, t_2^*)	MSE	SSIM
MET	$(7,23)$	20327.96	0.96104
medMET	$(27,119)$	4696.78	0.99160
METver1	$(7,21)$	20586.03	0.96058
METver2	$(7,21)$	20586.03	0.96058
METver3	$(29,120)$	4575.78	0.99184

Table 6: Peppers (three-level thresholding)

Method	(t_1^*, t_2^*)	MSE	SSIM
MET	$(38,125)$	4821.90	0.99137
medMET	$(35,126)$	4794.57	0.99139
METver1	$(37,125)$	4831.12	0.99135
METver2	$(37,125)$	4831.12	0.99135
METver3	$(36,127)$	4721.86	0.99155

Tables 1-8 illustrate the application of the described thresholding methods on the presented images. Specifically, the first four tables show the values of the MSE and SSIM coefficients for binary segmentation (two-level thresholding), while the next four tables present these values for multi-level segmentation (three-level thresholding). It is noticeable that the results vary depending on the image being analyzed and the number of threshold values being sought. For instance, for the binary thresholding applied on the girlface, the best MSE and SSIM values possesses MET method, while for the three-level case medMET provides better MSE and SSIM coefficients (as it could be expected according to the analysis of Fig.2). Note that these coefficients are the best for METver3 in this case (Table 5), as well as for the three-level thresholding in Table 6. METver1 and METver2 produce the same or very similar threshold values (and consequently MSE and SSIM indices) except in Table 8. Interestingly, in that case, the best MSE value corresponds to METver1, while the best SSIM value is associated with the METver2 method.

5. CONCLUDING REMARKS

Many thresholding techniques, such as the MET method, are very suitable for image segmentation due to their simplicity, effectiveness, and ease of application. Through numerous research, these methods are improved and adapted to meet specific requirements or image characteristics. In this research, we have proposed

DRAFT VERSION

Table 7: Polymersomes (three-level thresholding)

Method	(t_1^*, t_2^*)	MSE	SSIM
MET	$(145, 191)$	2316.06	0.99569
medMET	$(120, 204)$	1809.44	0.99648
METver1	$(130, 197)$	1821.08	0.99651
METver2	$(131, 197)$	1826.89	0.99650
METver3	$(119, 207)$	1830.20	0.99639

Table 8: Crowd (three-level thresholding)

Method	(t_1^*, t_2^*)	MSE	SSIM
MET	$(72, 110)$	4122.83	0.99520
medMET	$(96, 253)$	3301.68	0.99385
METver1	$(91, 253)$	3115.26	0.99446
METver2	$(73, 112)$	4009.87	0.99532
METver3	$(92, 253)$	3152.82	0.99433

three new methods for image thresholding, which are based on the MET and median MET approaches. The primary motivation for developing new methods is to combine the strengths of both approaches, resulting in new objective functions that carry information about both the means and medians of the classes. The presented experimental results suggest that the proposed methods could be valuable alternatives to the original methods.

However, to fully evaluate the usefulness and effectiveness of the presented methods, further research is necessary in the several areas: examining the effectiveness of the methods on a broader range of diverse images, analyzing their usability with noisy images, considering their application to color images, exploring potential improvements in computational efficiency, especially for multi-level thresholding, and potentially developing hybrid methods incorporating well-known techniques such as Otsu's and Kapur's methods.

ACKNOWLEDGMENT

This research is supported by the Science Fund of the Republic of Serbia, *#GRANT* No 7632, Project "Mathematical Methods in Image Processing under Uncertainty" - MaMIPU. This research is supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia by the Decision on the transfer of funds to finance the scientific research work of teaching staff at faculties in 2024, (Contract No. 451-03-136/2025-03/200126, No. 451-03-137/2025-03/200156 and No. 451-03-137/2025-03/200116).

LITERATURE

[1] K.K. Brar, B. Goyal, A. Dogra, M.A. Mustafa, R. Majumdar, A. Alkhayyat, V. Kukreja, "Image segmentation review: Theoretical background and recent advances," Information Fusion, vol. 114, 2025, pp. 102608

[2] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, "Image segmentation using deep learning: A survey," IEEE transactions on pattern analysis and machine intelligence, vol. 44, 2021, pp. 3523–3542

[3] M. Castiglione, F. Gallo, M. Pastorino, G. Moser, M. Nigro, N. Sacco, "Multisource methodology for traffic analysis zone definition based on the fusion of Remote Sensing, OpenStreetMap, and Floating Car Data," Journal of Transport Geography, vol. 128, 2025, pp. 104324

[4] S.W. Kim, A.H.L. Yuen, H.W. Kim, S. Lee, S.B. Lee, Y.M. Lee, W.J. Jung, C.T.C. Poon, D. Park, S. Kim, "Digital removal of dermal denticle layer using geometric AI from 3D CT scans of shark craniofacial structures enhances anatomical precision," Scientific Reports, vol. 15, 2025, pp. 19659

[5] Y. Harrak, A. Rachid, R. Aguejdad, "Evaluation of Spectral Indices and Global Thresholding Methods for the Automatic Extraction of Built-Up Areas: An Application to a Semi-Arid Climate Using Landsat 8 Imagery," Urban Science, vol. 9, 2025, pp. 78

[6] N. Otsu, "A threshold selection method from gray-level histograms," Automatica, vol. 11, 1975, pp. 23–27

[7] J.N. Kapur, P. Sahoo, A.K.C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," Computer vision, graphics, and image processing, vol. 29, 1985, pp. 273–285

[8] M.P. De Albuquerque, I. Esquef, A.R.G. Mello, "Image thresholding using Tsallis entropy," Pattern Recognition Letters, vol. 25, 2004, pp. 1059–1065

[9] C. Li, C.K. Lee, "Minimum cross entropy thresholding," Pattern recognition, vol. 26, 1993, pp. 617–625

[10] J. Kittler, J. Illingworth, "Minimum error thresholding," Pattern recognition, vol. 19, 1986, pp. 41–47

[11] M. Abd El Aziz, A. Ewees, A.E. Hassanien, "Whale optimization algorithm and moth-flame optimization for multilevel thresholding image segmentation," Expert Systems with Applications, vol. 83, 2017, pp. 242–256

[12] S. Agrawal, R. Panda, S. Bhuyan, B.K. Panigrahi,"Tsallis entropy based optimal multilevel thresholding using cuckoo search algorithm," Swarm and Evolutionary Computation, vol. 311, 2013, pp. 16–30

[13] E. Hancer, "Artificial bee colony: theory, literature review, and application in image segmentation," Recent advances on memetic algorithms and its applications in image processing, 2020, pp. 47–67

DRAFT VERSION

[14] M.H. Horng, “Multilevel thresholding selection based on the artificial bee colony algorithm for image segmentation,“ Expert Systems with Applications, vol. 38, 2011, pp. 13785–13791
[15] E. Houssein, E. Helmy, D. Oliva, A.A. Elngar, H. Shaban, “Multi-level thresholding image segmentation based on nature-inspired optimization algorithms: a comprehensive review,“ Metaheuristics in machine learning: theory and applications, 2021, pp. 239–265
[16] M. Sezgin, B. Sankur,“Survey over image thresholding techniques and quantitative performance evaluation,“ Journal of Electronic imaging, vol. 13, 2004, pp. 146–168
[17] T. Kurita, N. Otsu, N. Abdelmalek, “Maximum likelihood thresholding based on population mixture models,“ Pattern recognition, vol. 25, 1992, pp. 1231–1240
[18] J.H. Xue, D.M. Titterington, “Median-based image thresholding,“ Image and Vision Computing, vol. 29, 2011, pp. 631–637
[19] U. Sara, M. Akter, M.S. Uddin, “Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study,“ Journal of Computer and Communications, vol. 7, 2019, pp. 8–18
[20] Q. Huynh-Thu, M. Ghanbari, “Scope of validity of PSNR in image/video quality assessment,“ Electronics letters, vol. 44, 2008, pp. 800–801
[21] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,“ IEEE transactions on image processing, vol. 13, 2004, pp. 600–612

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

LOGIC AGGREGATION BY QUANTIFIER 'MOST OF' FOR ENTITIES EVALUATION

MIROSLAV HUDEC[1,2], FRANTIŠEK PAPALA[2], HANNA KRISTÍN SKAFTADÓTTIR[3]

[1] University of Economics in Bratislava, Faculty of Economic Informatics, Bratislava, Slovakia, miroslav.hudec@euba.sk 0000–0002–2868–0322
[2] VSB – Technical University of Ostrava, Faculty of Economics,Ostrava, Czech Republic, {frantisek.papala.st, miroslav.hudec}@vsb.cz,
[3] Bifröst University, Department of Business, Bifröst, Iceland, hannak@bifrost.is, 0000-0001-5228-8294

Abstract: *People often consider that when a number of satisfied elementary requirements increases, the entity should be more relevant. For this case, we need mixed aggregation function covering cases from nilpotent conjunction to nilpotent disjunction. In these cases, the aggregation of elementary requirements is formalised by the quantifier "most of" or variants thereof. However, this aggregation does not support mandatory or optional elementary requirements. Thus, we also explored how to aggregate mandatory and optional requirements with quantified aggregation. Next, we discussed how to augment business intelligence reporting by quantified aggregation and raised future research topics in assigning weights and recognising dependencies between elementary requirements.*

Keywords: *Fuzzy relative quantifier, Mixed aggregation function, Mandatory requirements, Business intelligence reporting.*

DRAFT VERSION

1. INTRODUCTION

In a business environment (or in everyday evaluation), people often consider that when a number of satisfied elementary requirements increases, the entity (product, customer, holiday destination) should be more relevant. On the contrary, when a number of satisfied elementary conditions decreases, the entity should be downplayed.

In this case, people evaluate somewhere between conjunctive aggregation (all elementary conditions should be satisfied) and disjunctive aggregation (at least one elementary condition should be satisfied). The former is restrictive, that is, an entity is rejected even though only one elementary condition is not met. This might lead to the empty answer problem (Bosc et al., 2008). The latter aggregation might be considered as too relaxed, that is, an entity is suitable if it meets at least one elementary condition, which leads to the overabundant answer problem (Bosc et al., 2008).

The main classification of aggregation functions applied in logical evaluation is due to (Dubois and Prade, 2004): conjunctive, averaging, disjunctive, and mixed. Generally, we can express conjunctive functions as $0 \leq A(\boldsymbol{x}) \leq \min(\boldsymbol{x})$, averaging functions as $\min(\boldsymbol{x}) \leq A(\boldsymbol{x}) \leq \max(\boldsymbol{x})$, disjunctive functions as $\max(\boldsymbol{x}) \leq A(\boldsymbol{x}) \leq 1$, and mixed aggregation functions (which can be adapted to data, that is, behave in different parts of domains differently), where $\mathbf{x}$ is a vector of elementary satisfaction degrees $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$.

Although the functions *min* and *max* might be considered too restrictive or too relaxed, respectively, the other conjunctive functions are even more restrictive (downward reinforcement), while the other disjunctive functions are even more relaxed (upward reinforcement) (Grabisch et al., 2009).

We can easily imagine a situation where two out of ten elementary satisfied conditions are not a good evaluation, while eight out of ten elementary satisfied conditions are a good evaluation. When about half atomic conditions are met, there is space for averaging evaluation (more or less good entity). Therefore, we need a mixed aggregation function. The question is which one and, moreover, how to adjust its parameters. But also considering its intuitiveness for the domain expert.

Elementary conditions can be partially satisfied or satisfied to some degree. For instance, a good sell results might be expressed as less than 1 000 definitely not good, more than 1 500 definitely good, while between 1 000 and 1 500 partially good, where satisfaction increases as we approach value 1 500.

Even binary data might exhibit intensities of satisfaction. For example, a hotel room has a balcony or does not have one. But the customer might interpret the requirement as: *it is irrelevant to me*, *I prefer the balcony*, *I weakly prefer the balcony*, *I strongly prefer balcony*, *or the balcony must be*. This option is handled by the proximity relation (Hudec et al., 2024).

In this work, to cover the observations mentioned above, we focus on mixed aggregation functions formalized by a quantifier *most of*.

After this introduction the article continues with theoretical observations and examples in Section2, while Section 3 is devoted to aggregation of mandatory and optional requirements into the quantified aggregation. Section 4 is focused on discussion and business intelligence reporting. Finally, Section 5 concludes this work.

2. THEORETICAL OBSERVATIONS

Mixed aggregation functions are not always intuitive to users, or are complex. Further, they might be non-associative (like ordinal sums (Hudec et al., 2021)), non-continuous in borderline cases (like uninorms (Munar et al., 2023)), cover only specific functions (gamma operator (Dujmović, 2018)), and the like. All of them have shown benefits in a variety of tasks. But for searching the best entities, rewarding according to intensities of satisfying majority of requirements, or in business intelligence reporting is too complex.

In this work, we evaluate the aggregation by a relative quantifier *most of* and its variants to meet the basic properties of the aggregation functions (Grabisch et al., 2009) borderline cases $f(0, 0,\ldots,0) = 0$ and $f(1,1,\ldots,1) = 1$, and monotonicity.

Data evaluation and classification are mental processes of reasoning and, therefore, belong to logic (Dujmović, 2018). In these tasks, we should cover the following reasoning requirements:

- Downplay relevance of low elementary scores (conjunctive reinforcement: strict or nilpotent), the solution is lower than or equal to (equal stands for case where we observe clear non-satisfaction of at least one elementary condition) the minimal degree of all elementary conditions.
- Emphasizing the relevance of high scores (disjunctive reinforcement: strict or nilpotent), the solution is higher than or equal to (equal stands for case where we observe full satisfaction of at least one elementary condition) maximal degree of all elementary conditions.
- Neutrality, when the decrease in the value of one (or more) elementary requirements is fully compensated by the same increase in the value of another (one or more) elementary requirement.
- Pessimistic inclination, when elementary requirements should be simultaneously (at least partially) satisfied, but the result is lower than the neutrality and higher than the minimal value.
- Optimistic inclination (or the levels of substitutability) when elementary conditions should be satisfied, but the result is higher than neutrality and lower than maximal value.

Quantified aggregation has been introduced in (Kacprzyk and Ziółkowski, 1986) for relaxing relational database SQL queries managed by AND connective. Recent work (Sojka et al., 2020) has shown that the same basic structure of linguistic summaries proposed by Yager (1982) and quantified aggregation are modifications of one function, either for summarization or quantified aggregation.

The quantified linguistic summarization is formalized as follows:

$$v(\boldsymbol{x}) = \mu_Q(p) \tag{1}$$

where **x** is the vector of the values of the considered attributes of an entity and

$$p = \frac{1}{n}\sum_{i=1}^{n} \mu_{Pi}(\boldsymbol{x}) \tag{2}$$

is the proportion of entities that meet (fully or partially) the set of elementary conditions P_i, where n is the number of elementary conditions posed on a subset of attributes and μ_P and μ_q are membership functions that express the intensity of belonging to the elementary condition and quantifier, respectively.

Next, we need to formalise quantifier *most of*, or variants thereof (Kacprzyk and Zadrożny, 2009, Zadeh, 1983). Quantifier can be formalized as a monotone or strict monotone function. In the latter, we cover data evaluation, i.e., only entity which fully meets all elementary conditions gets degree of 1. In the former, we are able to classify into rejected entities, partially accepted entities and fully accepted entities, because the function has an interval of rejection and full acceptance. The monotone function is depicted in Figure 1, where $m < n \leq 1$.

DRAFT VERSION

Usually, $m = 0.5$, to express that majority means more than half. When $n < 1$, we get three zones shown in Figure 1: A - nilpotent conjunction - entities are rejected or behaviour is below threshold, B - averaging or smooth transition from rejection to acceptance, and C - nilpotent disjunction - entities are accepted or behaviour is excellent.

Quantifiers are deeply discussed in e.g., (Murinová and Novák, 2024). For aggregation, only quantifers expressed as monotone increasing or strictly incerasing are suitable to meet the main properties of aggregation functions.

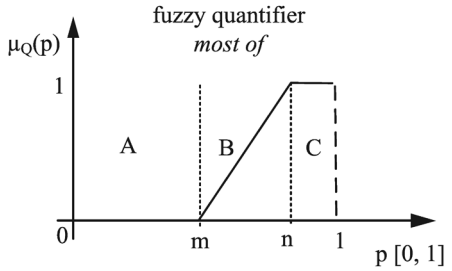


Figure 1: Quantifier *most of* as a monotone function

We propose a fuzzy aggregation mechanism based on relative quantifier *most of* as an aggregation function that produces a single holistic satisfaction score – essentially a tunable relaxation of conjunction and intensification of disjunction. This function balances between conjunctive ("all") and disjunctive ("at least one") logic in a continuous manner.

2.1 Illustrative example

Let us have ten elementary conditions P_1, P_2, …, P_{10} and a set of ten entities e_1, e_2, …, e_{10} to illustrate all the above-mentioned reasoning requirements. Parameters m and n (Figure 1) get values 0.5 and 0.85, respectively. The solution is in Table 1. For instance, entities are products, while conditions are preferred properties.

Table 1: Satisfaction degrees of elementary conditions and solutions by quantified aggregation, arithmetic mean, min and max functions

Entity	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	solution (2)	arithm. mean	min	max
e_1	1	0.8	0.9	0.2	0	0.9	0.4	0.85	0.9	0.6	**0.443**	0.655	0	1
e_2	1	0	1	1	1	0.75	0.8	0.9	0.6	0.9	**0.843**	0.795	0	1
e_3	0.75	0.85	0.9	0.85	0.95	0.7	0	1	0.9	0.85	**0.786**	0.775	0	1
e_4	1	0.9	0.2	0.3	0	0.15	0.85	0.35	0.7	1	**0.129**	0.545	0	1
e_5	0.8	0.9	0.95	0.85	0.8	0.85	0.95	0.9	0.9	0.8	**1**	0.87	0.8	0.95
e_6	0.2	0.3	0.3	0.35	0.11	0.22	0.1	0.4	0.2	0.35	**0**	0.253	0.1	0.4
e_7	0	0.8	0.95	0.75	0.9	0.45	1	0.55	0	0.65	**0.3**	0.605	0	1
e_8	1	1	1	1	1	1	1	1	1	1	**1**	1	1	1
e_9	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0
e_{10}	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	**0**	0.5	0.5	0.5

By observing entities e_8 and e_9 we see that the borderline cases are met. Entity e_{10} has idempotent behavior for all functions except quantified aggregation. It is expected because 0.5 is not majority of and, therefore, we have a nilpotent conjunction. Monotonicity is a matter of direct observation. Entities e_5 and e_6 exhibit a nilpotent disjunction (full acceptance) and a nilpotent conjunction (full rejection). When the proportion is closer to 0.5, than to 0.85 we have pessimistic inclination for entity e_1, while for entity e_2 we have optimistic inclination, as the proportion is approaching 0.85. Entities e_3 and e_4 indicate these inclinations and support the observation of monotonicity.

This function smoothly covers downplay, pessimistic inclination, through neutrality to optimistic inclination and to emphasizing entities. In this way, we covered the required aspects of mixed aggregation functions with a simple and easily applicable function.

DRAFT VERSION

2.2 Limitations

We should note that this aggregation (similarly as the arithmetic mean) does not cover the mandatory conditions or optional (full) alternatives.

3. QUANTIFIED AGGREGATION AGGREGATED WITH MANDATORY OR OPTIONAL CONDITIONS

This section focusses on aggregation when evaluation contains mandatory condition(s) and for the other holds *to more to merrier*, or optional full alternative(s) and for the other holds *to more to merrier*.

3.1 Mandatory conditions

In cost-benefit conditions, we have mandatory requirements like price. For example, when rewarding workers, they should meet the key requirement, and the other might be considered as if more of them satisfied it is better. The next example is accommodation: price should be met, while the majority of other requirements like distance to city center, breakfast quality and the like should be met. We formalize it in the following way:

$$v(\boldsymbol{x}) = \Lambda_{i=1}^{l} P_i \wedge \mu_Q(p) \tag{3}$$

where l is the number of mandatory requirements, while the proportion p is expressed as:

$$p = \frac{1}{n-l}\sum_{j=n-l}^{n} \mu_{Pi}(\boldsymbol{x}) \tag{4}$$

the other variables have the same meaning as in (1) and (2). Naturally, the question of the suitable conjunction appears. Quantified part is relaxed in comparison to conjunction. Hence, when we want to make aggregation stricter, options for conjunction are Łukasiewicz t-norm ($t(x_1, x_2) = \max(0, x_1 + x_2 - 1)$) or product t-norm ($t(x_1, x_2) = x_1 x_2$).

The initial example (Table 1) is modified that P_1 is mandatory requirement, while all the other requirements holds requirement *to more to merrier*.

Table 2: Satisfaction degrees of elementary conditions and solutions by conjunction of mandatory requirement and quantified aggregation of the other requirements

Entity	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	solution (3) by Łuk.t-norm
e_1	1	0.8	0.9	0.2	0	0.9	0.4	0.85	0.9	0.6	**0.333**
e_2	1	0	1	1	1	0.75	0.8	0.9	0.6	0.9	**0.778**
e_3	0.75	0.85	0.9	0.85	0.95	0.7	0	1	0.9	0.85	**0.544**
e_4	1	0.9	0.2	0.3	0	0.15	0.85	0.35	0.7	1	**0**
e_5	0.8	0.9	0.95	0.85	0.8	0.85	0.95	0.9	0.9	0.8	**0.800**
e_6	0.2	0.3	0.3	0.35	0.11	0.22	0.1	0.4	0.2	0.35	**0**
e_7	0	0.8	0.95	0.75	0.9	0.45	1	0.55	0	0.65	**0**
e_8	1	1	1	1	1	1	1	1	1	1	**1**
e_9	0	0	0	0	0	0	0	0	0	0	**0**
e_{10}	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	**0**

Maintaining the original requirements and modifying the P_1 condition to a mandatory one reveals that certain results remain the same. For instance, the borderline cases in the case of entities e_8 and e_9 are still satisfied. Additionally, entity e_{10} continues to hold a value of zero. When using a conjunction, such as Łukasiewicz's, in a given setting, there can be no increase in satisfaction degrees, only deterioration. This is evidenced by the remaining entities, e_1 to e_7 (except for no change in e_6|). Anyway, the monotonicity holds,

DRAFT VERSION

so this function meets properties of aggregation functions. To reach the value of 1, it is necessary to satisfy the mandatory condition (P_1) to the maximum value (1) and, concurrently, achieve nilpotent disjunction with the other conditions. This occurs when the proportion is greater than 0.85.

3.2 Optional conditions

Similarly, adding optional requirements into quantified aggregation is not a difficult task. They should be connected by disjunction with the quantified aggregation, i.e., in (3) we replace ∧ with ∨.

We formalize it in the following way:

$$v(\boldsymbol{x}) = \bigvee_{i=1}^{l} P_i \vee \mu_Q(p) \tag{5}$$

where l is the number of optional requirements, while the proportion p is expressed as:

$$p = \frac{1}{n-l} \sum_{j=n-l}^{n} \mu_{Pi}(\boldsymbol{x}) \tag{6}$$

the other variables have the same meaning as in (1) and (2). Analogously, the question of the suitable disjunction appears. Quantified part is intensified in comparison to disjunction. Hence, when we want to make aggregation weaker and therefore options for disjunction are Łukasiewicz t-conorm ($t(x_1, x_2) = \min(1, x_1 + x_2)$) or probabilistic sum t-conorm ($t(x_1, x_2) = x_1 + x_2 - x_1 x_2$).

4. DISCUSSION AND BUSINESS INTELLIGENCE REFLECTION

The proposed fuzzy-majority aggregation is scalable to any number of elementary requirements and entities; however, high intercorrelation among elementary requirements may distort the score. When many criteria are strongly dependent, a preliminary correlation (or collinearity) test is advisable before aggregating the data.

Another open issue is the assignment of weights. While averaging operators naturally incorporate weights that sum to 1, defining weights for conjunctive or disjunctive components is a more involved process. In conjunctive aggregation all conditions are very relevant, because they are mandatory. One promising avenue is to encode importance through fuzzy implications (Zadrożny et al., 2008), a topic we leave for future work.

Overall, the fuzzy-majority approach mirrors human reasoning by down-valuing alternatives that meet only a few criteria and up-valuing those that satisfy the majority. It offers a flexible yet rigorous tool that extends the querying in business intelligence.

Business intelligence queries often involve multiple criteria that managers or decision-makers consider simultaneously. The same empty-answer or over-abundance described in Sections 1–2 occurs routinely in business intelligence dashboards, where analysts filter entities by multiple key performance indicators (KPIs). For example, a sales manager might seek products that excel across majority of performance metrics (e.g., year-over-year sales growth, profit margin, customer satisfaction, and the like). Using a classical filtering approach with a conjunction could be too rigid, potentially returning no results if no product meets all criteria. By contrast, a fuzzy quantified aggregation that uses the *most of* quantifier can retrieve items that satisfy most of the conditions. For example, one could issue a query like *select products where most of (sales growth > 5%, profit margin > 10%, satisfaction > 80%, etc)* to rank products by how many criteria they fulfill, whe product reaches treshold *m* (Figure 1). In addition, sales manager is not forced to assign precise values, that is, instead of *sales growth > 5%,* manager mihgt say *sales growth around 5% and* more ensuring that 4.9% is sitll acceptable, but with a lower degree. This yields a ranked list of products satisfying the majority of criteria. Entities are ranked by intensities of satisfying quantified condition, so usual functions like HEAD showing several the best performing entities can be applied. In effect, fuzzy quantified queries retrieve significant results that conjunctive filters might overlook, while still filtering out largely irrelevant items.

5. CONCLUSION

We face the problems of diverse needs for the logic aggregation of elementary requirements for the evaluation of entities. In this work, we explored quantified aggregation “to more to merrier” that supports the natural way of human evaluation in a smooth transition from nilpotent conjunction across pessimistic and optimistic averaging evaluation to nilpotent disjunction by a simple function. However, this aggregation does not support mandatory or optional requirements. To solve this limitation, we proposed aggregation of

DRAFT VERSION

mandatory requirements by conjunctive function to quantified aggregation, as well as aggregation of optional requirements by disjunctive function to quantified aggregation. One of benefits is augmenting business intelligence dashboards with this functionality for example, to rank top entities by quantified aggregation.

The limitation is a possible correlation among elementary requirements, which might skew the result. Thus, the future research topic is recognizing correlation as well as direction of influence to exclude highly correlated and at the same time dependent attributes appearing in elementary conditions.

ACKNOWLEDGMENT

This article has been produced with the support of the the SGS project no. SP2025/081 of the Ministry of Education, Youth and Sports of the Czech Republic.

LITERATURE

Bosc, P., Hadjali, A., & Pivert, O.(2008). Empty versus overabundant answers to flexible relational queries. *Fuzzy Sets and Systems*, *159*, 1450–1467.

Dubois, D., & Prade, H. (2004). On the use of aggregation operations in information fusion processes, *Fuzzy Sets and Systems*m *142*(*1*), 143–161.

Dujmović J. (2018). Soft Computing Evaluation Logic: the LSP Decision Method and Its Applications. Hoboken: Wiley–IEEE Computer Society.

Grabisch M., Marichal J.-L., Mesiar R., & Pap E. (2009). *Aggregation functions*. Cambridge: University Press.

Hudec, M., Vučetić, M., & Barčáková, N. (2014). *An extended proximity relation and quantified aggregation for designing robust fuzzy query engine*. *Knowledge-Based Systems*, *290*, 111574.

Hudec M., Mináriková E., Mesiar R., Saranti A., & Holzinger A. (2021). Classification by ordinal sums of conjunctive and disjunctive functions for explainable AI and interpretable machine learning solutions. *Knowledge-Based Systems*, *220*, 106916.

Kacprzyk, J., & Zadrożny, S. (2009). Protoforms of Linguistic Database Summaries as a Human Consistent Tool for Using Natural Language in Data Mining. *International Journal of Software Science and Computational Intelligence*, *1*, 1-11.

Kacprzyk, J. & Ziółkowski, A. (1986). Database queries with fuzzy linguistic quantifiers. *IEEE transactions on systems, man, and cybernetics*, *2007*, *16*(*3*), 474-479.

Munar, M., Hudec, M., Massanet, S., Mináriková, E., & Ruiz-Aguilera, D. (2023). On an Edge Detector Based on Ordinal Sums of Conjunctive and Disjunctive Aggregation Functions. In: S. Massanet, S. Montes, D. Ruiz-Aguilera, & M. González-Hidalgo (Eds). *Fuzzy Logic and Technology, and Aggregation Operators*, *EUSFLAT AGOP 2023*: V*ol. 14069. Lecture Notes in Computer Science* (pp. 271-282) Springer.

Murinová, P., & Novák, V. (2024). Logical syllogisms with “Almost all, Most, Many, A few” and “Several”. *International Journal of Approximate Reasoning*, *175*, 109284.

Sojka P., Hudec M., & Švaňa M. (2020). Linguistic summaries in evaluating elementary conditions, summarizing data and managing nested queries. *Informatica*, *31*(*4*), 841-856.

Yager, R.R. (1982). A new approach to the summarization of data. *Information Sciences*, *28*, 69-86.

Zadeh, L. A. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with Applications*, *9*(*1*), 149–184.

Zadrożny, S., De Tré, G., De Caluwe, R., & Kacprzyk, J. (2008). An overview of fuzzy approaches to flexible database querying. In J. Galindo (Ed), *Handbook of Research on Fuzzy Information Processing in Databases* (pp. 34-54). Hershey: Information Science Reference.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

LEVERAGING GENERATIVE ADVERSARIAL NETWORKS IN PROJECT MANAGEMENT FOR TASK ESTIMATION

NIKOLA VOJTEK[1], BOJAN SMUDJA[2], PAVLE MILOŠEVIĆ[3], IVANA DRAGOVIĆ[4]

[1]Daon | Smart Consulting Agency, Belgrade, nikola@smartconsulting-agency.com, 0000-0001-9380-7046
[2]Agile Frogg, Belgrade, bojan.smudja@gmail.com,
[3]University of Belgrade – Faculty of Organizational Sciences, Belgrade, pavle.milosevic@fon.bg.ac.rs, 0000-0002-5943-6023
[3]University of Belgrade – Faculty of Organizational Sciences, Belgrade, ivana.dragovic@fon.bg.ac.rs, 0000-0001-6935-7444

Abstract: *Accurate project estimation remains one of the most challenging aspects of project management, particularly when historical data is scarce, or project complexity is high. This study explores the use of Generative Adversarial Networks (GANs) to generate synthetic project data, including task duration estimates, by learning from historical data. Jira migration projects were used as an example. A new project with a distinct scope was introduced, and four GAN model variations were evaluated using Mean Absolute Error and Fréchet Distance to identify the most effective combination of parameters. The selected model was then used to generate task estimates for the new project, which were compared against traditional estimates based on the average values. Results show that the GAN-generated estimates were more conservative, highlighting potential hidden complexities and offering improved risk buffering. This approach demonstrates the potential of AI to enhance forecasting accuracy in project management and lays the foundation for future work on synthetic estimation of resources allocation, dependencies modeling, and risks identification.*
Keywords: *project management, generative adversarial networks, synthetic data.*

1. INTRODUCTION

In the last 10 years, soft computing techniques are massively applied and used in transforming various business domains. In the project management area, soft computing techniques, especially neural networks are used to allow more accurate forecasting of project outcomes. The benefits could be seen in identifying potential delays, cost overruns, and resource constraints before they occur, allowing for proactive adjustments.

AI technologies, including machine learning algorithms and predictive analytics, are transforming project management by enabling more accurate forecasting of project outcomes. These tools assist in identifying potential delays, cost overruns, and resource constraints before they occur, allowing for proactive adjustments. For instance, AI-driven models have been utilized to predict project performance metrics in urban infrastructure projects, enhancing the ability to manage timelines and budgets effectively (Sadeghi, 2024). Additionally, AI facilitates the automation of routine tasks, improves stakeholder communication through natural language processing, and supports decision-making processes by analyzing large datasets for actionable insights (Al-Arafat et al., 2024; Lan, 2023). The integration of AI into project management practices leads to increased efficiency, reduced human error, and more informed strategic planning (Davahli, 2020).

This paper deals with a specific area of the application of AI techniques in project estimations and predictive analytics, and more specifically with the data scarcity challenge. Accurate project estimations rely heavily on the availability of high-quality historical data, which can be limited in certain domains. AI techniques, including Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), offer a way to generate realistic synthetic data to overcome this limitation, enabling more accurate decision-making and risk assessment in project management. Various synthetic data generation techniques, including GANs, VAEs, and data augmentation methods, have been explored for project management applications. These techniques have been used for cross-project fault prediction (Wang et al., 2025), predictive process monitoring (Lin, Khan, & Latif, 2025), and risk identification in supply chains (Hu et al., 2025). These approaches allow project managers to generate realistic project scenarios, simulate potential risks, and optimize resource allocation when historical data is scarce or incomplete, making them valuable tools for project planning and decision-making.

DRAFT VERSION

The focus of this paper is on applying GANs for generating project task estimates, based on the resource allocations, dependencies, timelines, and other constraints, for a new project with a certain level of complexity. This approach uses examples from similar past projects to train the GAN model, allowing it to produce realistic project data. The obtained results are then compared with traditional estimation approaches to assess accuracy, variance, and overall project planning efficiency.

The rest of this paper is organized as follows. Section 2 provides a short overview of GANs and their application in project management. Section 3 describes the experiment design, including steps and specific considerations for applying GANs to project estimation. Section 4 presents the results of the experiment, followed by a summary of findings, concluding remarks, and proposals for future work in Section 5.

2. BASIC PRINCIPLES OF GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks consist of two neural networks (a Generator and a Discriminator) competing in a zero-sum game (Goodfellow et al., 2014). GANs could learn to generate realistic data without labeled training examples. So far, this approach has been applied to image synthesis, text generation, data augmentation, and now even project task simulation.GANs are a class of machine learning frameworks designed to generate realistic synthetic data. They work by training two neural networks in tandem– Generator that creates synthetic data trying to mimic the real dana, and Discriminator, with the role to evaluate the quality of the generated data by distinguishing it from real data.The two networks are set up as adversaries in a zero-sum game. The Generator tries to "fool" the Discriminator by producing data as realistic as possible.The Discriminator attempts to correctly classify data as either real (from the actual data distribution) or fake (produced by the Generator).

This approach with two networks has found diverse applications across various business domains, including project management. One study explores the application of GANs to enhance credit risk identification in supply chains (Hu et al., 2025). In another paper authors Lin, Khan and Latif (2025) proposes a novel adversarial training framework for next event prediction in business processes, and highlights the importance of using Generative Adversarial Nets in Predictive Business Process Monitoring domain. Also, in the (Wang et al., 2025), authors considered using GANs for defect prediction across different software projects.

GANs work following next steps:

1. Initialize the Networks: Both the Generator and Discriminator are initialized with random weights.
2. Training Loop: (a) The Generator creates synthetic data; (b) The Discriminator evaluates the generated data; (c) Feedback from the Discriminator is used to update both networks.
3. Convergence: The process continues until the Generator becomes good enough that the Discriminator cannot distinguish real from synthetic data.

The results generated by the GANs aim to replicate realistic patterns and relationships found in historical dana. In the context of project management, these networks are enabling the generation of plausible task breakdowns and resource plans for any new or hypothetical projects. The GANs learn to simulate complex dependencies and estimation behaviors that are otherwise difficult to recognize and model manually. The output can support project managers in scenarios such as planning, rough forecasting, and improving estimation accuracy when limited data is available.

3. EXPERIMENT OVERVIEW

3.1 Purpose and goal of the experiment

The purpose of this experiment is to evaluate if GANs can be effectively used for estimating project tasks, resource utilization, and dependencies in complex migration projects. This includes comparing synthetically generated data with traditional estimation approaches to assess accuracy, variance, and overall project planning efficiency. Specifically, it aims to determine if GANs can provide more accurate and realistic task duration estimates, resource allocations, and timelines for complex Jira migration projects.

Five project examples related to the migration of the project management tools are used for this purpose:

- Project 1: Simple Jira Core Migration (On Time)
  - Basic Jira Core migration without Confluence or plugins.
  - 6 tasks, including data backup, migration, and closure.
- Project 2: Jira and Confluence Migration (Delayed)
  - More complex migration involving multiple systems and plugins.
  - 7 tasks, including plugin compatibility testing and user training.

DRAFT VERSION

- Project 3: Comprehensive Migration with Jira Service Management (On Time)
    - Includes Jira, Confluence, and JSM with automation rules.
    - 8 tasks, including custom workflow migration and automation.
- Project 4: Complex Migration with Custom Plugins (Delayed)
    - High-security migration with custom plugins and third-party integrations.
    - 8 tasks, including security review and plugin compatibility testing.
- Project 5: Multi-Instance Consolidation (On Time)
    - Consolidates multiple Jira and Confluence instances.
    - 8 tasks, including instance analysis, consolidation, and scripting.

The projects are described using following attributes: (a) Task duration: Estimated and Actual logged times; (b) Resource Assignments: Specific team members for each task; (c) Resource Utilization: Utilization percentages based on logged time; (d) Milestones and Dependencies: Key project milestones and critical task sequences; (e) Cost Breakdown: Licensing, labor, training, and contingency costs.

New project for which data will need to be generated represents a complex Jira migration project, and includes more complex requirements, as well as slightly different scope, as provided below: (a) Migrate Jira Software from Instance A to Instance B; (b) Migrate automation rules from Instance A to Instance B; (c) Establish a new SYNC plugin connection between Instance B and Instance C; (d) Migrate specific users to Instance B; (e) Perform two migration rehearsals; (f) Conduct user training for Instance B.

This project is more complex than the previous five examples that are available, as it involves not just the migration of data but also the transfer of complex automation rules, user accounts, and plugin configurations. Additionally, the SYNC connection setup between two separate instances introduces unique technical challenges that require careful planning and execution. The need for multiple rehearsals and user training further adds to the project's complexity, making it an ideal candidate for evaluating the effectiveness of GAN-based estimation.

3.2 Experiment stages

The experiment was conducted following next steps.

Data Preparation for GANs: Key features were extracted from five existing migration projects to serve as input for the GAN models. The data underwent cleaning, structuring, and normalization to ensure consistency and quality for training. Important elements such as task durations, resource utilization, and task dependencies were included in the prepared dataset.

GAN Models Training: Four different GAN models were used to generate realistic project task data. These models were trained using the cleaned and structured data.

Models Evaluation and Accuracy: The performance of the GAN models was assessed using MAE as well as the Wasserstein distance to measure the similarity between generated and real data distributions.

Synthetic Data Generation: Using the trained GAN models, synthetic task estimates were generated for a new migration project (Project 6). The generated data included predicted task durations, required resources, and task dependencies to simulate a realistic project plan.

Traditional Estimation (Baseline): For comparison, traditional estimation methods were applied to Project 6 using historical data. Estimates for task durations, resource allocation, and project timelines were created to serve as a baseline dataset.

Evaluation and Comparison: Finally, the GAN-generated data were evaluated against the traditional estimates. The comparison focused on accuracy, variance in predictions, and the overall efficiency of the projected timelines, providing insights into the potential benefits of GAN-based estimation methods.

4. RESULTS

4.1 Models training performances

For this experiment, 4 different variances of GANs were used, as provided in Table 1. The four model variants (V1–V4) are designed and proposed by the domain expert to explore the impact of training duration, batch size, regularization techniques, and loss functions on GAN performance for project estimation tasks. Training results are provided in Figures 1-4.

DRAFT VERSION

Table 1: Models variations included in the experiment

Variant	Epochs	Batch Size	Architecture	Loss
V1	1000	32	Default (Linear + ReLU)	Binary Cross Entropy (BCE)
V2	2000	64	Same as V1	BCE
V3	1500	32	With BatchNorm + Dropout	BCE
V4	1000	32	Same as V1	Wasserstein (WGAN)

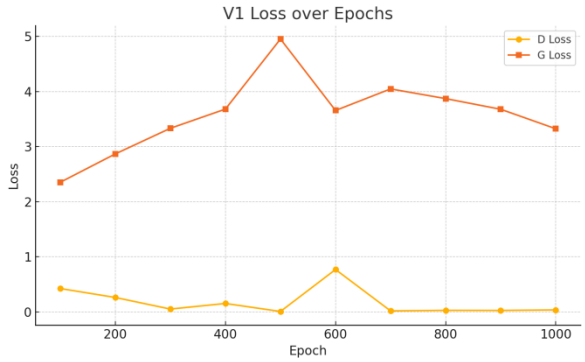


Figure 1: GAN V1, 1000 Epochs, 32 batch size

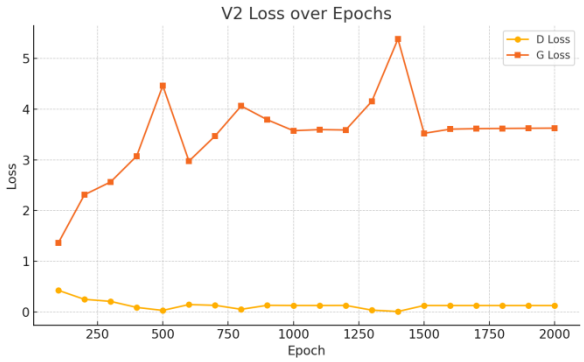


Figure 2: GAN V2, 2000 Epochs, 64 batch size

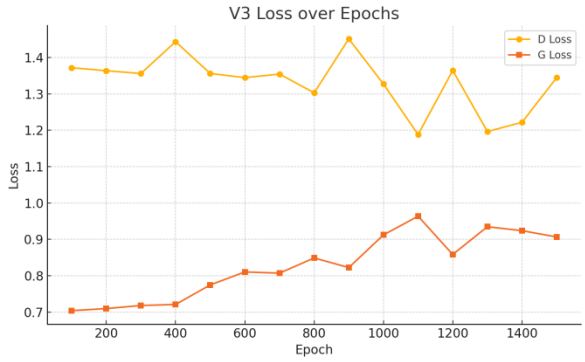


Figure 3: GAN V3, 1500 Epochs, 32 batch size

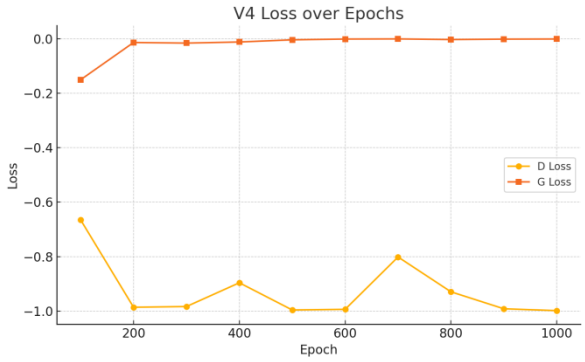


Figure 4: GAN V4, 1000 Epochs, 32 batch size

Different performance levels were observed in terms on how well GANs could learn and generate realistic project data. Model V1 showed a healthy balance between learning and accuracy, steadily improving and generating reliable results. Model V2, trained longer with more data per batch, was the most stable and showed the best overall learning behavior. Model V3, which added techniques to prevent overfitting, unfortunately ended up with one part of the system becoming too dominant, leading to weaker results. Model V4 showed very limited progress and didn't learn useful patterns effectively. All this indicates that V2 is the most stable, followed by V1. In case of the remaining two models, V3 and V4, both showed some signs of inefficiency.

Next, performance metrics were calculated and presented in Table 2. We are generating project estimations like durations, resource utilization, etc., thus MAE is the first choice. It tells how far off are synthetic estimations from historical ones, and it's easy to interpret. On the other hand, Fréchet Distance measures the difference in distribution between the real and generated data, considering both central tendency and relationships between features. GANs may generate values that look accurate individually but don't follow the same patterns as the original data. Thus, Fréchet Distance seems to be a good choice since it catches those deeper mismatches in structure. This combination of pointwise and distributional metrics is a common and recommended approach in evaluating generative model performance (Sajjadi et al., 2018).

Model V3 showed the best performance, i.e. it generates synthetic project data that is not only more accurate (lower MAE) but also more structurally realistic (lower Fréchet Distance). V1 and V2 perform almost identically despite different training setups, indicating that simply increasing epochs or batch size does not improve performance. Finally, model V4 shows potential but didn't outperform BCE-based models in this setup.

DRAFT VERSION

Table 2: Models performance metrics

Variant	MAE	Fréchet Distance
V1	0.491856	2.630361
V2	**0.491856**	**2.630382**
V3	0.30594	0.457385
V4	0.488585	2.598531

4.2. Synthetic data generation and comparison with the traditional estimation

In this step, synthetic data for a new project scenario is generated using the best-performing GAN model V3. The generator, trained on the original dataset, is used to replicate the underlying distribution of task estimations and workload patterns. This synthetic dataset is intended to support further experimentation and evaluation, particularly in contexts where real project data is limited or unavailable. Synthetic data is provided in the Table 3, column 2. Total Duration of the new project is 149.52 days. In addition, comparing to the historically available data regarding the tasks that were executed in the previous 5 projects, there are three tasks that are new, and historical data about their durations are not available. Those three tasks are highlighted in the table and are: Identify and Document Ticket Content Rules, Migrate Users to Instance B and Establish SYNC Connection between Instance B and Instance C. In the next section, to compare the total duration, arbitrary value for those new tasks will be used.

Table 3: Synthetic data – Task estimates for Project 6

Task	Estimates (Synthetic)	New task	Estimates (Traditional)
Project Kickoff	12.62	No	3.2
Identify and Document Ticket Content Rules	**13.6**	**Yes**	**8.8**
Export Automation Rules from Instance A	14.42	No	6
Setup New Instance B	12.35	No	15
Migrate Users to Instance B	**18.24**	**Yes**	**8.8**
Establish SYNC Connection between Instance B and Instance C	**11.58**	**Yes**	**8.8**
Migrate Automation Rules to Instance B	5.67	No	6
Migration Rehearsal 1	2.71	No	14
Backup Instance A	5.79	No	7.66
Migration Rehearsal 2	7.14	No	14
Final Data Migration	11.95	No	14
Post-Migration Validation	11.07	No	6.25
Staff Training for New Instance B	15.97	No	7.2
Project Closure	6.41	No	3.6
Total Duration in days	***149.52***		***123.31***

Furthermore, the comparison of model V3 estimations with the traditional approach where average values are calculated for the tasks based on the historically available data (Table 3, column 4) is provided. These values are used as a baseline, and the complete comparison is provided in the next table. As noted in the previous chapter, there are three new tasks for which historical data doesn't exist. For the comparison of the total duration of the new project between the average and values that were synthetically generated, for each new task, the average duration of all other tasks was used (8.8 days).

The synthetic estimates, generated using a GAN-based approach with model V3, result in a total of 149.52 days for the new project, whereas the traditional estimation approach totals 123.31 days. This represents a 21% increase and suggests that the synthetic model included more conservative durations, potentially accounting for hidden complexities. Notably, some tasks such as "Migrate Users to Instance B," "Project Kickoff," and "Staff Training" are significantly higher in the synthetic model, highlighting areas where traditional estimation may underestimate effort. Overall, the synthetic estimates offer a broader buffer, which could reduce risk of delay but may also impact resource planning and stakeholder expectations.

Finally, more conservative estimations made by the GAN do not necessarily result in more accurate predictions, but they offer greater security in terms of scheduling, cost control, and stakeholder confidence.

5. CONCLUSION

The advances in AI technologies, such as machine learning and predictive analytics, is reshaping project management by enhancing the precision of outcome forecasting. These advanced tools enable early detection of potential risks, including schedule delays, budget overruns, and resource limitations, thereby supporting timely and informed decision-making.

This paper deals with a specific area of the application of AI techniques in project estimations and predictive analytics, and more specifically with the data scarcity challenge. The focus is on applying Generative Adversarial Networks for generating project task estimates for a new project with a certain level of complexity. The scope of the new project differed from that of the historical projects used as reference. Nevertheless, this historical data served as the training foundation for the GAN model, enabling it to generate realistic and contextually relevant project outputs. Four variations of GANs were considered as part of this experiment, each with different training duration, batch size, regularization techniques, and loss functions. For the evaluation of model performance, two accuracy metrics were used MAE and Fréchet Distance. The obtained results are compared with traditional estimation approach to assess accuracy, variance, and overall project planning efficiency. GAN-based model V3 estimated a total project duration of 149.52 days, compared to 123.31 days using traditional methods, indicating a 21% increase. This may suggest that the synthetic model accounted for potential complexities often overlooked, offering a more cautious yet potentially more realistic outlook that may help mitigate delivery risks.

This experiment demonstrated the potential of GANs to generate realistic project estimates based on historical data, even when applied to a project with a different scope. By comparing synthetic and traditional estimates, it was observed how AI-based models can reveal hidden complexities and provide more conservative planning baselines. As a next step, the approach could be extended to (1) generate additional dimensions such as resource allocations, and dependencies, and (2) validated across a broader dataset. Integrating these models into project management workflows could support more informed decision-making. Ultimately, these results may lay the foundation for implementing a Jira plugin that enhances proactive risk management.

ACKNOWLEDGMENT

This research is partially supported by the University of Belgrade – Faculty of Organizational Sciences.

LITERATURE

Sadeghi, S. (2024). Enhancing Project Performance Forecasting using Machine Learning Techniques. arXiv preprint arXiv:2411.17914.

Al-Arafat, M., Kabir, M. E., Morshed, A., & Islam, M. M. (2024). Artificial Intelligence in Project Management: Balancing Automation and Human Judgment. Frontiers in Applied Engineering and Technology, 2(01), 18-29.

Lan, C. (2023). Integrating Artificial Intelligence and Predictive Analytics for Feasibility Studies in Construction Projects: An Optimization Approach in Uncertainty Environments. SSRN.

Davahli, M. R. (2020). The Last State of Artificial Intelligence in Project Management. arXiv preprint arXiv:2012.12262.

Hu, H., Zhao, X., Huang, Y., Li, Q., & Zhang, W. (2025). Credit Risk Identification in Supply Chains Using Generative Adversarial Networks. arXiv preprint arXiv:2501.10348.

Lin, Y., Khan, R. A., & Latif, S. (2025). Predictive Business Process Monitoring via Generative Adversarial Nets. arXiv preprint arXiv:2003.11268.

Wang, J., Chen, X., Li, Z., & Yu, H. (2025). Generative Adversarial Network-based Cross-Project Fault Prediction. arXiv preprint arXiv:2105.07207.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. In Advances in Neural Information Processing Systems (NeurIPS), 27, 2672-2680.

Sajjadi, M. S. M., Bachem, O., & Lucic, M. (2018).Assessing generative models via precision and recall. In Advances in Neural Information Processing Systems (NeurIPS), 31, 5228–5237.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

APPLICATION OF COMPUTATIONAL INTELLIGENCE BASED CLUSTERING ALGORITHMS ON EDUCATIONAL DATA

MILICA ZUKANOVIĆ[1], PAVLE MILOŠEVIĆ[2], RAGNAR MÁR VILHJÁLMSSON[3], HARALDUR DADI RAGNARSSON[4], ANA POLEDICA[5]

[1] University of Belgrade – Faculty of Organizational Sciences, Belgrade, milica.zukanovic@fon.bg.ac.rs, 0000-0003-3650-8327
[2] University of Belgrade – Faculty of Organizational Sciences, Belgrade, pavle.milosevic@fon.bg.ac.rs, 0000-0002-5943-6023
[3] Bifröst University, Department of Business, Iceland, ragnarmar@bifrost.is, 0009-0007-5373-0233
[4] Bifröst University, Department of Business, Iceland, haraldurr@bifrost.is, 0009-0009-9032-014X
[5] University of Belgrade – Faculty of Organizational Sciences, Belgrade, ana.poledica@fon.bg.ac.rs, 0000-0002-5929-6446

Abstract: *Contemporary research, especially in the academic community, focuses on the increasingly frequent application of artificial intelligence (AI) in learning and teaching processes, but also on all the benefits and disadvantages that it brings with it. This study was conducted in order to examine the application of computational intelligence (CI) methods on educational data. The main goal of the paper is to group students according to similarity with the aim to analyse each group in details. Data was collected via questionnaire among students enrolled at three universities in three different countries. It represents students' opinion on the use of AI in education. Using k-medoids, self-organized maps (SOM) and fuzzy C means (FCM), two groups of students were identified according to the use of AI in learning process as the most important attribute.*

Keywords: *computational intelligence, education, clustering, FCM, SOM, AI.*

1. INTRODUCTION

Development and everyday presence of AI have revolutionized modern education. The influence of AI on the educational system is often considered two-folded (Hudec et al., 2025). On one hand, there is a strong need to include and/or enhance education in the field of AI due to market demands and trends. On the other hand, there is an ongoing work in creating a solid base for positive perceptions and acceptance of AI technologies. Governments and educational institutions have identified a clear need to rethink and redesign university curricula to better fit the growing presence of AI in every aspect of our lives (Pedro et al., 2019). Still, there is a significant difference in AI perception between regions. The Government AI Readiness Index is commonly used to assess how prepared governments are to implement AI on a global scale (Storozhenko et al., 2023). The level of AI acceptance is often correlated with government efforts and a country's development in information technology (IT). Hence, enhancing AI and IT education is a vital prerequisite for broader and more effective integration. In (Hazaimeh & Al-Ansi, 2024), the authors focus on studying AI acceptance in higher education through human interaction-based factors, including attitudes, competencies, and openness to experience. Their conclusions were based on data from more than hundred participants collected within Arab higher education institutions. The results show that attitude, competency, and openness have a positive and significant relationship with perceived benefits and AI acceptance. Further, in Malaysian and Pakistani higher education institutions, a study (Dahri et al., 2024) underlined that students are concerned about using AI tools to improve their academic performance. Also, previous interaction with AI tools were an important factor in predicting AI acceptance.

Most papers in this area of research rely on traditional methodologies, such as predefined questionnaires with Likert-scale questions, accompanied by statistical analysis and interpretation of results. However, a recent trend is to incorporate soft computing techniques into result analysis, as this more sophisticated approach may enhance insights in the subject area. Particularly, SOM stand out as a useful strategy for analysing survey data (Bedle et al, 2023). Since SOM is in fact unsupervised neural network, it is able to model overcome some of the drawbacks of standard clustering algorithms, e.g. poor modelling of complex non-convex cluster shapes, and sensitivity to initialization.

DRAFT VERSION

This research builds on the work presented in (Hudec et al., 2025) which published the initial findings on Icelandic, Slovakian, and Serbian students' perceptions of using AI in education. Herein, we are going one step further, as we aim to cluster survey respondents and analyse their differences regarding background, study area, country, etc. Namely, we aim to use three different clustering techniques to address. First, we used a standard k-means clustering technique, as a benchmark. Furthermore, we applied self-organizing maps as a more sophisticated clustering technique to overcome some of the standard limitations of k-means clustering. Finally, we employed a FCM clustering to detect possible marginal cases that should be carefully addressed and point out the level of membership to a certain cluster. Obtained results show that more sophisticated methods like SOM and FCM performed better on survey data than standard clustering methods.

This paper is organized as follows. Section 2 provides a brief literature review covering the most important aspects and challenges regarding AI in education, as well as the main directions concerning clustering techniques for grouping survey respondents. The questionnaire design and clustering dataset are presented in Section 3. Section 4 is devoted to the methodology, i.e. the clustering techniques used in this study. The results are presented in Section 5, along with a discussion and comparison with previously published work. Finally, Section 6 outlines the main conclusions and suggests directions for future research.

2. LITERATURE REVIEW

In education, technological developments and enhancements have for a long time altered, and even improved, the ways of teaching and studying. Factor in recent emergence of AI as a transformative force in higher education, with a growing body of academic literature exploring its applications in teaching and learning within universities. Integration of AI in this context covers a broad area; from the use of AI among teachers to enhance or facilitate teaching processes, to the use of AI to assist students in learning and solving assignments, to the inclusion of the AI concept within course study materials and lectures. Furthermore, AI can also be utilized for administration purposes at universities.

A review of published research over a ten year period between 2015 and 2024 studying teachers‘ use of AI technology in their teaching and professional development, a considerable gap was found between studies examining the application of AI technology in teaching (65%) versus AI‘s role in enhancing teacher professional development (35%) (Tan et al., 2024). The majority including studies on applications such as conversational AI, AI-driven learning and assessment systems, immersive technologies, visual and auditory computing, and teaching and learning analytics, against aspects of teachers' integration AI technologies into their teaching practices. This highlights the need for future research focusing more on the potential of AI in teacher professional development and how AI technologies address the development needs of teachers integrating AI technologies into their teaching practices (Tan et al., 2024).

Most research on AI in education is within computer science and STEM fields. Out of 40 research studies involving AI and reviewed by Zhang and Aslan (2021), 25 were in the fields of engineering, computer science, information technology, mathematics and science.

Among AI technologies commonly used to facilitate education are intelligent tutoring systems (ITS) where a trainee interacts with a software to acquire knowledge with a real time assessment, adjusting tutoring level accordingly (Alfaro et al., 2020; Crompton & Burke, 2023), learning management systems (LMS) used to manage learning and teaching interventions, including identifying drop-out risks to provide necessary support for students (Rosário and Dias, 2022), and content creation and curation of teaching materials to personalize students' learning experience and improving their engagement with learning resources (Zawacki-Richter et al., 2019; Chen et al., 2020).

Then there is the notion of the future role of human teachers versus AI-based teaching. Over a ten year period, Cope and al. (2020) researched how the nature of machine intelligence could enhance education. Although they determined that AI would not take over the role of a human teacher because it profoundly differs from human intelligence, they identified how AI can empower educators with better ways to teaching.

In a Swedish study of a machine learning-based teaching aid in mathematics, authors counter the notion that human actors will always be needed along with technology-enhanced learning and that teachers‘ tasks cannot be automated (Sperling et al., 2022). In his book, Neil Selwyn (2019) even argued whether robots could possibly replace teachers in the classroom and in response to increasing shortage of teachers around the world, some have proposed that AI could replace some of the functions historically performed by human power (Edwards & Cheok, 2018; Selwyn, 2019). One could argue that would not only hold true, but also, be a method of increased productivity in more complex aspects of research and teaching where AI could relieve teachers of the most automated aspects of their functions, providing them with more time to focus on expanding their knowledge of expertise.

DRAFT VERSION

3. DATA

A questionnaire was sent to enrolled students at three universities. The questions were designed to capture student's knowledge, usage and perceptions towards the use of AI in education. As data was collected in native languages of three different countries, questions were translated to best represent the same meaning across all participants.

Each questionnaire was divided into three themes. First theme was focused on perception and knowledge of AI with questions such as Do you think AI can improve education? and Which AI technology do you know? The second theme focused on benefits of using AI in education where participants were asked to rate and evaluate the benefits of AI in various areas within their university. The last theme was focused on preference towards AI in studying and teaching with participants indicating how intensively they prefer to use AI in various aspects of their learning processes and in which fields of education AI should or should not be used.

For the research presented in this paper, a dataset collected through a survey of students from Serbia, Slovakia, and Iceland was used (Hudec et al., 2025). The survey was completed by 49 students studying Information Systems and Technologies and 16 students studying Management at the University of Belgrade, Faculty of Organizational Sciences; 62 students studying Business Informatics at the University of Economics in Bratislava, Slovakia; and 54 business students from Bifröst University in Iceland. Since four students from Iceland did not fully complete the survey, their responses were excluded from further analysis. As a result, the final dataset consists of 177 instances described by 30 attributes, including responses related to the AI tools used in the learning process, their impact on various aspects of education and universities, and students' concerns or fears about AI. As the questions were answered using a Likert scale, the responses were transformed into values ranging from 1 to 5. An analysis of the dataset confirmed that there are no missing values or outliers that could affect the stability of the clustering model.

4. METHODOLOGY

Clustering is a method for grouping observations based on their similarities and it is significant for many areas, such as predicting students' performance (Mohamed et al., 2022), influence of different factors on student learning (Moubayed et al., 2020), predicting students' risk of dropout (Pecuchova & Drlik, 2024), etc. According to literature (Ferraro & Giordani, 2020), clustering can be divided into two main subgroups: hard clustering and soft clustering.

Hard clustering algorithms are partitive, i.e. divide the dataset into groups. Therefore, k-means clustering algorithm classify the data points based on features into k number of group by minimizing the sum of squares of distances between data and the corresponding cluster centroid (Madhulatha, 2011). The k-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. Hence, *k*-medoids clustering, as a variant of k-means, is more resistant to noise and extreme values (Lund & Ma, 2021). A medoid is defined as a cluster element that minimizes the total distance to all other points within the same cluster, thus ensuring greater interpretability of the results (Madhulatha, 2011).

SOM are the most popular artificial neural network for clustering because they use the same procedure for a competitive layer, determine a final neuron, and then weight vectors for the neighbours of the final neuron within a certain radius (Hagan et al., 2014). The primary goal is to represent a multidimensional input space in two dimensions. Unlike other types of neural networks, SOM preserves the topological relationships among input data, enabling meaningful grouping based on spatial proximity. One of the key advantages of SOM is that it does not require prior assumptions about the nature of the data.

In soft clustering, therefore also in the FCM, each observation belongs to each cluster, although the membership functions are different. Level belonging to the cluster depends on the proximity of the observation to the centroid (Bezdek et al., 1984). The purpose of this type of clustering is the minimization of the objective function to prove the similarity between observations in a cluster and reduce the similarity between observations compared to other clusters (Bezdek et al., 1984). The advantages of fuzzy clustering are in the fact that it provides more detailed data information as well as membership function values that are suitable for analysis, and it is adaptable to numerical optimization (Hashemi et al., 2023).

Clustering was performed first using the *k*-medoids method as benchmark, and then according to literature, more sophisticated methods such as SOM and FCM were used. Since all methods require a defined number of clusters, the elbow method, as well as silhouette score, Calinski-Harabasz index and Davies-Bouldin index were applied for determining the optimal number of clusters. By applying these methods, it was decided that the instances can be optimally clustered into two clusters.

DRAFT VERSION

5. RESULTS AND DISCUSSION

Experiments were conducted with different parameters for *k*-medoids, SOM and FCM. By measuring the silhouette coefficient for each listed model, the highest coefficient was measured for the model of FCM. However, when applying the FCM, four students whose cluster membership functions were extremely close to the value 0.5 were discovered. Hence, it is difficult to assign these students to a specific cluster, and additional analysis and attention should be paid to them. Table 1 show given metrics, while Figures 1 to 3 represent the cluster distribution for each of the algorithms.

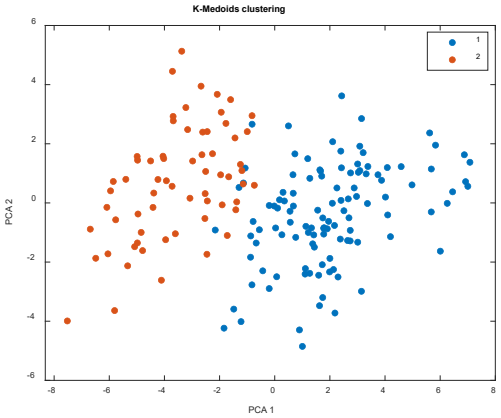


Figure 1: Result of *k*-medoids clustering

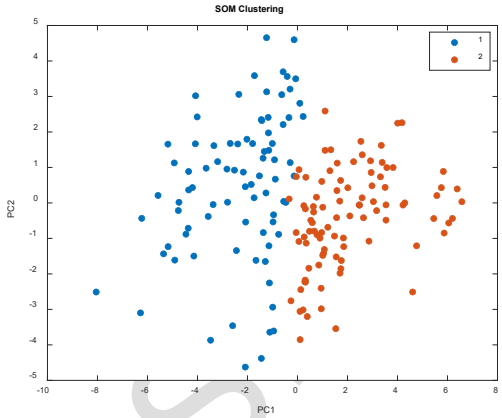


Figure 2: Result of SOM clustering

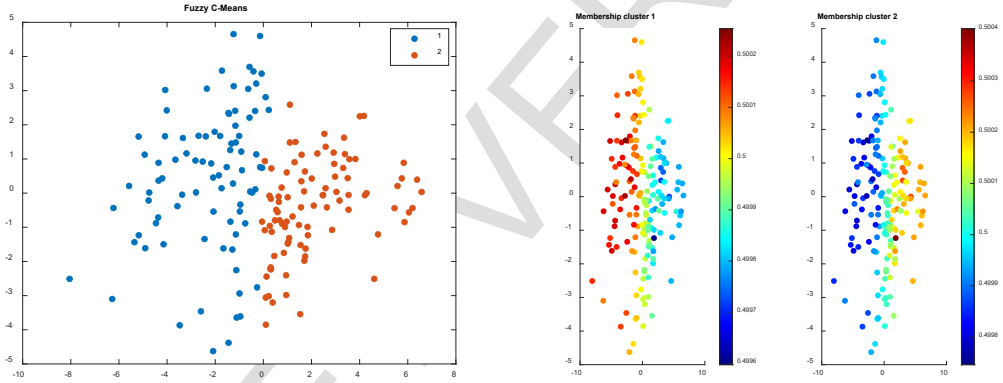


Figure 3: Result of FCM clustering

Table 1: Metrics of applied clustering algorithms

Metrics		*k*-medoids	SOM	FCM
Number of instances	Cluster 1	69	84	82
	Cluster 2	108	93	91
	'Gray area'	-	-	4
Silhouette coefficient		0.5332	0.5742	**0.5752**

Results of clustering provide significant insight into students' preferences and behavior patterns which can have a direct impact on study program curriculums and strategies for inclusion of AI into education. Based on the application of SOM and FCM algorithm with two clusters, the following groups were identified:

1. ***Analog minds*** – Students who do not use AI tools in learning process and who believe that AI negatively affects skills, especially social interactions. Also, these students are strictly against the use of AI when translating a foreign language into English and processes automation (for example applying to the university, registration for courses, dealing with financial matters, asking Erasmus mobility, etc.). The only exception where students use AI is for application for project development and assistance with literature reviews and summarization. Interestingly, this cluster is mostly made up of students studying computer science or similar study programs, but they explicitly point out that

DRAFT VERS

they are afraid of the negative impact of AI. If you are a student from Serbia, you probably belong to this cluster.

2. ***AI-powered learners*** – These students represent visionaries who believe that AI can improve processes at universities. They believe that AI has a positive impact on education, but also that it can improve it in the future. Technology-augmented students see benefits of AI in many areas at university, especially in searching for information concerning study, as well as available literature and resources for seminar essays and final thesis. They also support the use of AI learning tools, except automatic analysis of online discussion boards, in order to provide feedback to students on their participation in LMS. These students are not afraid of negative impact of artificial intelligence, and in fact, strongly support its inclusion in study programs of computer science, general business and business informatics. These are mostly Slovak students.

Students in the gray area of the FCM method have characteristics of both clusters and use AI to a moderate extent in the learning process. These students are enrolled in computer science study programs and mostly use AI when searching for information related to their studies.

6. CONCLUSION

The main goal of this study was to answer the questions of how to group students using CI methods. The paper evaluates the application of CI methods, *k*-medoids, SOM and FCM in order to cluster students from three European countries. Students were grouped according to the answers they gave in the questionnaire. Two groups were obtained according to the use of AI tools in the learning process. It was concluded that the first group includes students who do not use AI and are afraid with its negative effects, while the second group includes students who consider AI to have a positive impact on education in general. Additionally, the FCM has been shown to perform well when it comes to survey data.

The improvement of the research is reflected in the expansion of the dataset to students from other countries, which would provide a broader picture, but also to observe the problem from the teacher's point of view. Moreover, further continuation of the research can go in the direction of applying other fuzzy clustering methods.

ACKNOWLEDGMENT

This study was supported by the University of Belgrade – Faculty of Organizational Sciences, as well as the Ministry of Science, Technological Development and Innovation of Serbia under Grant No. 337-00-3/2024-05/18.

LITERATURE

Alfaro, L., Rivera, C., Castaneda, E., Zuniga-Cueva, J., Rivera-Chavez, M., & Fialho, F. (2020). A review of intelligent tutorial systems in computer and web based education. International Journal of Advanced Computer Science and Applications, 11(2), 755-763.

Bedle, H., Garneau, C. R., & Vera-Arroyo, A. (2023). Clustering energy support beliefs to reveal unique sub-populations using self-organizing maps. Heliyon, 9(7). https://doi.org/10.1016/j.heliyon.2023.e18351

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. Computers & geosciences, 10(2-3), 191-203. https://doi.org/10.1016/0098-3004(84)90020-7

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. Ieee Access, 8, 75264-75278. https://doi.org/10.1109/ACCESS.2020.2988510

Cope, B., Kalantzis, M., & Searsmith, D. (2020). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. Educational philosophy and theory, 53(12), 1229-1245. https://doi.org/10.1080/00131857.2020.1728732

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. International journal of educational technology in higher education, 20(1), 22. https://doi.org/10.1186/s41239-023-00392-8

Dahri, N. A., Yahaya, N., Al-Rahmi, W. M., Vighio, M. S., Alblehai, F., Soomro, R. B., & Shutaleva, A. (2024). Investigating AI-based academic support acceptance and its impact on students' performance in Malaysian and Pakistani higher education institutions. Education and Information Technologies, 29(14), 18695-18744. https://doi.org/10.1007/s10639-024-12599-x

DRAFT VERSION

Edwards, B. I., & Cheok, A. D. (2018). Why not robot teachers: artificial intelligence for addressing teacher shortage. Applied Artificial Intelligence, 32(4), 345-360. https://doi.org/10.1080/08839514.2018.1464286

Ferraro, M. B., & Giordani, P. (2020). Soft clustering. Wiley Interdisciplinary Reviews: Computational Statistics, 12(1), e1480. https://doi.org/10.1002/wics.1480

Hagan, M., Demuth, H., Beale, M., & DeJesus O. (2014). Neural Network Design (2nd ed.). Martin Hagan.

Hashemi, S. E., Gholian-Jouybari, F., & Hajiaghaei-Keshteli, M. (2023). A fuzzy C-means algorithm for optimizing data clustering. Expert Systems with Applications, 227, 120377. https://doi.org/10.1016/j.eswa.2023.120377

Hazaimeh, M., & Al-Ansi, A. M. (2024). Model of AI acceptance in higher education: arguing teaching staff and students perspectives. The International Journal of Information and Learning Technology, 41(4), 371-393. https://doi.org/10.1108/IJILT-01-2024-0005

Hudec, M., Skaftadóttir, H. K., Ragnarsson, H. D., Vilhjálmsson, R. M., Rakovská, E., Mokošová, D., Milošević, P., Zukanović, M., & Barčaková, N. (2025). Student Perception of Using Artificial Intelligence in Education. 11th International Conference on eDemocracy & eGovernment (ICEDEG 2025) (accepted for publication).

Lund, B., & Ma, J. (2021). A review of cluster analysis techniques and their uses in library and information science research: k-means and k-medoids clustering. Performance Measurement and Metrics, 22(3), 161-173. https://doi.org/10.1108/PMM-05-2021-0026

Madhulatha, T. S. (2011, July). Comparison between k-means and k-medoids clustering algorithms. In International Conference on Advances in Computing and Information Technology (pp. 472-481). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-22555-0_48

Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering analysis for classifying student academic performance in higher education. Applied Sciences, 12(19), 9467. https://doi.org/10.3390/app12199467

Moubayed, A., Injadat, M., Shami, A., & Lutfiyya, H. (2020). Student engagement level in an e-learning environment: Clustering using k-means. American Journal of Distance Education, 34(2), 137-156. https://doi.org/10.1080/08923647.2020.1696140

Pecuchova, J., & Drlik, M. (2024). Enhancing the Early Student Dropout Prediction Model Through Clustering Analysis of Students' Digital Traces. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3486762

Pedro, F., Subosa, M., Rivas, A., & Valverde, P. (2019). Artificial intelligence in education: Challenges and opportunities for sustainable development.

Rosário, A. T., & Dias, J. C. (2022). Learning management systems in education: Research and challenges. In Digital Active Methodologies for Educative Learning Management (pp. 47-77). IGI Global. https://doi.org/10.4018/978-1-6684-4706-2.ch003

Selwyn, N. (2019). Should robots replace teachers?: AI and the future of education. John Wiley & Sons.

Sperling, K., Stenliden, L., Nissen, J., & Heintz, F. (2022). Still w (AI) ting for the automation of teaching: An exploration of machine learning in Swedish primary education using Actor-Network Theory. European Journal of Education, 57(4), 584-600. https://doi.org/10.1111/ejed.12526

Storozhenko, L., Ignatenko, O., Yaroshovets, T., Antypenko, I., & Vlasenko, V. (2023). E-democracy in the context of the information society: prospects, challenges and opportunities. Revista Amazonia Investiga, 12(70), 63-77. https://doi.org/10.34069/AI/2023.70.10.6

Tan, X., Cheng, G., & Ling, M. H. (2024). Artificial Intelligence in Teaching and Teacher Professional Development: A Systematic Review. Computers and Education: Artificial Intelligence, 100355. https://doi.org/10.1016/j.caeai.2024.100355

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators?. International Journal of Educational Technology in Higher Education, 16(39), 1-27. https://doi.org/10.1186/s41239-019-0171-0

Zhang, K., & Aslan, A. B. (2021). AI technologies for education: Recent research & future directions. Computers and education: Artificial intelligence, 2, 100025. https://doi.org/10.1016/j.caeai.2021.100025

DRAFT VERSION

STATISTICAL MODELS

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

THEORETICAL AND EMPIRICAL RESULTS ON THE COVERAGE PROBABILITY OF THE STANDARD QUANTILE ESTIMATOR FOR THE BURR DISTRIBUTION

JELENA STANOJEVIĆ[1], MARIJA KOPRIVICA[2]

[1] University of Belgrade, Faculty of Economics and Business, Belgrade, jelena.stanojevic@ekof.bg.ac.rs, 0000-0001-5668-5297
[2] University of Belgrade, Faculty of Economics and Business, Belgrade, marija.koprivica@ekof.bg.ac.rs, 0000-0003-4239-2252

Abstract: *In this paper, we determine the coverage probability of the standard quantile estimator for the Burr Type XII distribution. The analysis is based on large deviations theory, and is conducted under the assumptions of independence and identical distribution (i.i.d.) between observations; however the results may be used in the case of negatively dependent variables as well. Numerical results are provided for various parameter settings and quantile levels, illustrating the estimator's convergence properties with increasing sample size. Our findings improve the understanding of the estimator's performance in the context of heavy-tailed distributions relevant to insurance and finance.*

Keywords: *coverage probability, standard quantile estimator, Value-at-Risk*

1. INTRODUCTION

The estimation of high quantiles plays a crucial role in many fields, particularly in risk management, finance, and insurance. High quantiles — such as the 95th or 99.5th percentilesare — are especially important in extreme value analysis and are commonly used to define risk measures, such as the widely accepted Value-at-Risk (VaR). Accurate quantile estimation is essential for developing reliable statistical models to evaluate extreme events.

Most of the literature on quantile estimation focuses on symmetric or moderately skewed distributions (Miljković, Causey, and Jovanović, 2022). However, in practice, insurance and financial loss data often exhibit extreme skewness and heavy tails. This is particularly relevant for lines of insurance prone to catastrophic claims, such as property, aviation, and marine insurance. The Burr Type XII distribution — a three-parameter, heavy-tailed distribution (Burr, 1942) — has gained attention as a suitable model for such loss data (Beirlant et al., 2001; Cummins et al., 1990; Miljković and Grün, 2016; Venter, 1983). Despite its practical relevance, theoretical results on high quantile estimation under the Burr Type XII distribution, especially in the presence of dependence structures, remain limited.

The main goal of the paper is to derive the upper bound for the probability $P\{|\widehat{x}_p(n) - x_p| \geq \varepsilon\}$, which we will denote as η. However, the probability of the true quantile value x_p belonging to the random interval $[\widehat{x}_p(n) - \varepsilon, \widehat{x}_p(n) + \varepsilon]$, i.e. the "coverage probability" is equal to $1 - \eta$. The quantity η can be understood as the convergence rate (in probability) for any fixed $\varepsilon > 0$.

We analyze the upper bound for the convergence rate, or, equivalently, the lower bound for the coverage probability, for a given ε, of the standard quantile estimator for the Burr Type XII distribution. The analysis is based on large deviations theory (Jin and Fu, 2007), and is conducted under the assumptions of independence and identical distribution (i.i.d.) between observations, however the results may be used in the case of negatively dependent variables as well. We also complement the theoretical results with numerical results to illustrate the impact of parameter settings and sample size on convergence.

2. SOME DEFINITIONS AND PRELIMINARIES

Let $(X_1, X_2, ..., X_n)$ be a sample from a random variable X with a nondegenerate distribution $F(x)$. The empirical distribution function of X is given by:

$$F_n(x) = \frac{1}{n} \sum_{k=1}^{n} I\{X_k \leq x\}, x \in R,$$

where $I\{X_k \leq x\}$ denotes the indicator function of the event $\{X_k \leq x\}$.

DRAFT VERSION

► **Definition 1.** The p-th quantile x_p of the distribution $F(x)$ is defined as:

$$x_p = \inf\{x : F(x) \geq p\},\ p \in (0,1). \tag{1}$$

In this paper, we consider the case of the direct-simulation quantile estimator, which is the standard quantile estimator in the literature, and obtain the main result for this setting.

► **Definition 2.** The direct-simulation quantile estimator of X is defined by the formula:

$$\widehat{x}_p(n) = \inf\{t : F_n(t) \geq p\},\ p \in (0,1), \tag{2}$$

where $F_n(\cdot)$ is the empirical distribution function of X.

Another important concept in this paper is negative dependence.

► **Definition 3.** Random variables X_1, X_2, …, X_n are said to be negatively dependent if the following two inequalities hold for all $x_1, x_2, ..., x_n$:

$$P\{X_1 \leq x_1, ..., X_n \leq x_n\} \leq P\{X_1 \leq x_1\} \cdot ... \cdot P\{X_n \leq x_n\},$$

and

$$P\{X_1 \geq x_1, ..., X_n \geq x_n\} \leq P\{X_1 \geq x_1\} \cdot ... \cdot P\{X_n \geq x_n\}.$$

An auxiliary result used in this paper is the following theorem, which is the main result in (Jin and Fu, 2007).

► **Theorem 4.** *If the distribution F is strictly increasing and the sequence $\{X_n, n \geq 1\}$ consists of negatively dependent random variables, then*

$$P\{|\widehat{x}_p(n) - x_p| \geq \varepsilon\} \leq e^{-n\Delta_+(\varepsilon,n)} + e^{-n\Delta_-(\varepsilon,n)}, \quad \textit{for all } \varepsilon > 0, \tag{3}$$

where

$$\begin{aligned}
\Delta_+(\varepsilon,n) &= \sup_{-\infty<\lambda\leq 0} \left(\lambda p - \frac{\ln E\left(\exp\left(\lambda \sum_{i=1}^n I\{X_i \leq x_p + \varepsilon\}\right)\right)}{n} \right), \\
\Delta_-(\varepsilon,n) &= \sup_{0\leq\lambda<+\infty} \left(\lambda p - \frac{\ln E\left(\exp\left(\lambda \sum_{i=1}^n I\{X_i \leq x_p - \varepsilon\}\right)\right)}{n} \right).
\end{aligned}$$

Moreover, the rate is improved by negative dependence in the sense that

$$\begin{aligned}
\Delta_+(\varepsilon,n) &\geq \sup_{-\infty<\lambda\leq 0} \left(\lambda p - \ln E\left(\exp\left(\lambda I\{X \leq x_p + \varepsilon\}\right)\right)\right) > 0, \\
\Delta_-(\varepsilon,n) &\geq \sup_{0\leq\lambda<+\infty} \left(\lambda p - \ln E\left(\exp\left(\lambda I\{X \leq x_p - \varepsilon\}\right)\right)\right) > 0,
\end{aligned}$$

where the "sup" on the right-hand side denotes the appropriate supremum and represents the rate for i.i.d. samples.

3. THE MAIN RESULT

In this section, we investigate a specific case of the Burr distribution and calculate the upper bound for the convergence rate of the standard quantile estimator $\widehat{x}_p(n)$ to the true quantile value x_p, or, equivalently, the lower bound for the appropriate coverage probability.

Similar results for various forms of the Pareto distribution, including the Generalized Pareto Distribution (GPD), were presented in (Stanojević and Jović, 2016).

We consider the Burr Type XII distribution, also known as the Singh–Maddala distribution (sometimes referred to as the generalized log-logistic distribution), with the distribution function given by:

$$F(x) = 1 - \left(\frac{\theta}{\theta + x^\tau}\right)^\alpha,\ x \geq 0,$$

with appropriate parameters α, θ, $\tau > 0$. We obtain the following result.

DRAFT VERSION

► **Theorem 5.** *Let $\{X_n, n \geq 1\}$ be i.i.d. random variables with a common Burr Type XII distribution, given by $\overline{F}(x) = \left(\frac{\theta}{\theta+x^\tau}\right)^\alpha = \theta^\alpha(\theta+x^\tau)^{-\alpha}$, $x \geq 0$, where α, θ, $\tau > 0$. The upper bound for the convergence rate of the standard quantile estimator $\widehat{x}_p(n)$ to the true quantile value x_p, is given by the following inequality:*

$$P\{|\widehat{x}_p(n) - x_p| \geq \varepsilon\} \quad \leq \quad \left(\frac{p(1-p_+)}{p_+(1-p)}\right)^{-np} \cdot \left(\frac{1-p_+}{1-p}\right)^n + \left(\frac{p(1-p_-)}{p_-(1-p)}\right)^{-np} \cdot \left(\frac{1-p_-}{1-p}\right)^n, \tag{4}$$

where p_+ and p_- are: $p_+ = 1-\theta^\alpha(\theta+(x_p+\varepsilon)^\tau)^{-\alpha}$ and $p_- = 1-\theta^\alpha(\theta+(x_p-\varepsilon)^\tau)^{-\alpha}$.

Proof. The Burr Type XII distribution is strictly increasing, and in the case where $\{X_n, n \geq 1\}$ are negatively dependent, we can apply Theorem 4:

$$P\{|\widehat{x}_p(n) - x_p| \geq \varepsilon\} \leq e^{-n\Delta_+(\varepsilon,n)} + e^{-n\Delta_-(\varepsilon,n)}, \quad \text{for all } \varepsilon > 0,$$

where

$$\begin{aligned}
\Delta_+(\varepsilon,n) &= \sup_{-\infty<\lambda\leq 0}\left(\lambda p - \frac{\ln E\left(\exp\left(\lambda\sum_{i=1}^n I\{X_i \leq x_p+\varepsilon\}\right)\right)}{n}\right),\\
\Delta_-(\varepsilon,n) &= \sup_{0\leq\lambda<+\infty}\left(\lambda p - \frac{\ln E\left(\exp\left(\lambda\sum_{i=1}^n I\{X_i \leq x_p-\varepsilon\}\right)\right)}{n}\right).
\end{aligned}$$

In the case where $\{X_n, n \geq 1\}$ are i.i.d., from the same theorem we also have the following result:

$$\begin{aligned}
\Delta_+(\varepsilon,n) &\geq \sup_{-\infty<\lambda\leq 0}\left(\lambda p - \ln E\left(\exp\left(\lambda I\{X \leq x_p+\varepsilon\}\right)\right)\right),\\
\Delta_-(\varepsilon,n) &\geq \sup_{0\leq\lambda<+\infty}\left(\lambda p - \ln E\left(\exp\left(\lambda I\{X \leq x_p-\varepsilon\}\right)\right)\right).
\end{aligned}$$

Let us denote the right-hand sides of the previous inequalities by Δ_+ and Δ_-, respectively, and it is:

$$\begin{aligned}
\Delta_+ &= \sup_{-\infty<\lambda\leq 0}\left(\lambda p - \ln E\left(\exp\left(\lambda I\{X \leq x_p+\varepsilon\}\right)\right)\right),\\
\Delta_- &= \sup_{0\leq\lambda<+\infty}\left(\lambda p - \ln E\left(\exp\left(\lambda I\{X \leq x_p-\varepsilon\}\right)\right)\right).
\end{aligned}$$

If we denote:

$$p_+ = P\{X \leq x_p+\varepsilon\} = F(x_p+\varepsilon) = 1-\theta^\alpha(\theta+(x_p+\varepsilon)^\tau)^{-\alpha},$$

$$p_- = P\{X \leq x_p-\varepsilon\} = F(x_p-\varepsilon) = 1-\theta^\alpha(\theta+(x_p-\varepsilon)^\tau)^{-\alpha},$$

$$p = P\{X \leq x_p\},$$

then our goal is to derive expressions for Δ_+ and Δ_-.

The distribution of the indicator $I\{X \leq x_p+\varepsilon\}$ is given by

$$I\{X \leq x_p+\varepsilon\} : \begin{pmatrix} 0 & 1 \\ 1-p_+ & p_+ \end{pmatrix},$$

and we can compute:

$$\Delta_+ = \sup_{-\infty<\lambda\leq 0}\left(\lambda p - \ln(e^\lambda p_+ + 1 - p_+)\right).$$

The supremum of the above function is attained at $\lambda = \ln\frac{p(1-p_+)}{p_+(1-p)}$ (note, λ is always negative since $p < p_+$). Consequently, we have:

$$\Delta_+ = p\ln\frac{p(1-p_+)}{p_+(1-p)} - \ln\frac{1-p_+}{1-p}.$$

DRAFT VERSION

Similarly, the distribution of the indicator $I\{X \le x_p - \varepsilon\}$ is given by

$$I\{X \le x_p - \varepsilon\} : \begin{pmatrix} 0 & 1 \\ 1-p_- & p_- \end{pmatrix},$$

and we can compute:

$$\Delta_- = \sup_{0 \le \lambda < +\infty} \left(\lambda p - \ln(e^{\lambda} p_- + 1 - p_-) \right).$$

The supremum of the above function is attained at $\lambda = \ln \frac{p(1-p_-)}{p_-(1-p)}$ (note, λ is always positive since $p > p_-$). Consequently, we have:

$$\Delta_- = p \ln \frac{p(1-p_-)}{p_-(1-p)} - \ln \frac{1-p_-}{1-p}.$$

Finally,

$$\begin{aligned} P\{|\widehat{x}_p(n) - x_p| \ge \varepsilon\} &\le e^{-n\Delta_+} + e^{-n\Delta_-} \\ &= \left(\frac{p(1-p_+)}{p_+(1-p)} \right)^{-np} \cdot \left(\frac{1-p_+}{1-p} \right)^{n} + \left(\frac{p(1-p_-)}{p_-(1-p)} \right)^{-np} \cdot \left(\frac{1-p_-}{1-p} \right)^{n}, \end{aligned}$$

where $p_+ = 1 - \theta^{\alpha}(\theta + (x_p + \varepsilon)^{\tau})^{-\alpha}$ and $p_- = 1 - \theta^{\alpha}(\theta + (x_p - \varepsilon)^{\tau})^{-\alpha}$, and the proof is completed. ◀

Note, the proof of the above theorem is conducted using the result (Jin and Fu, 2007) of upper bond for the convergence rate for the i.i.d. case. Although the bound still holds for the negatively dependent variables, it is looser than the bounds $\Delta_+(\varepsilon, n)$ and $\Delta_-(\varepsilon, n)$, and those refinements are not utilized for the case which we consider here.

4. APPLICATION OF RESULTS

In this section, we present an application of the main theorem from the previous section along some numerical calculations. All results are obtained for two sets of parameter values — α, θ and τ — of the Burr Type XII distribution. For each set, we consider four probability levels: $p = 0.005$, $p = 0.05$, $p = 0.995$ and $p = 0.95$, representing both low and high quantiles. Additionally, for each set of values, two different ε values are considered.

Using Theorem 5 and the corresponding formula (4), for each set of parameter values we calculate the upper bound for the convergence rate of the standard quantile estimator $\widehat{x}_p(n)$ to the true quantile value x_p, or, equivalently, the lower bound for the appropriate coverage probability. Figure 1–4 summarize these results.

The following two graphs present the coverage probability depending on the sample size, for the Burr Type XII distribution with parameters $\alpha = 5.1$, $\theta = 590$, $\tau = 4.5$, focusing on one low quantile ($x_{0.005}$) and one high quantile ($x_{0.995}$), and two fixed values for ε.

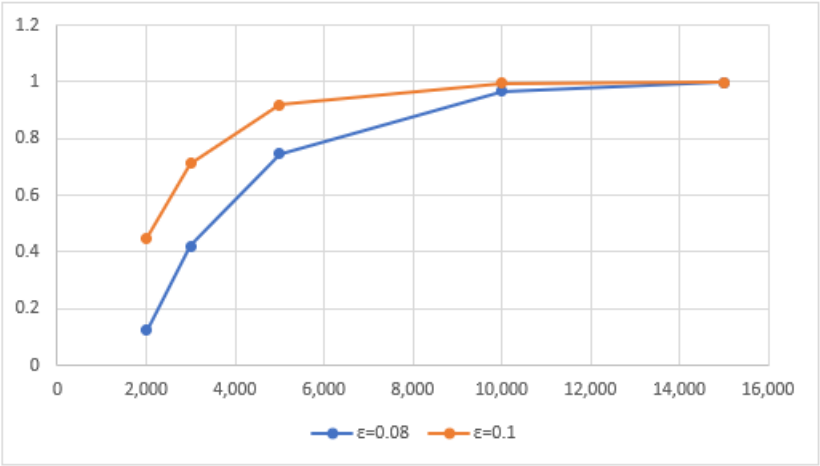


Figure 1: Coverage probability vs sample size, for the low quantile with $p = 0.005$ and $x_{0.005} = 0.886$

DRAFT VERSION

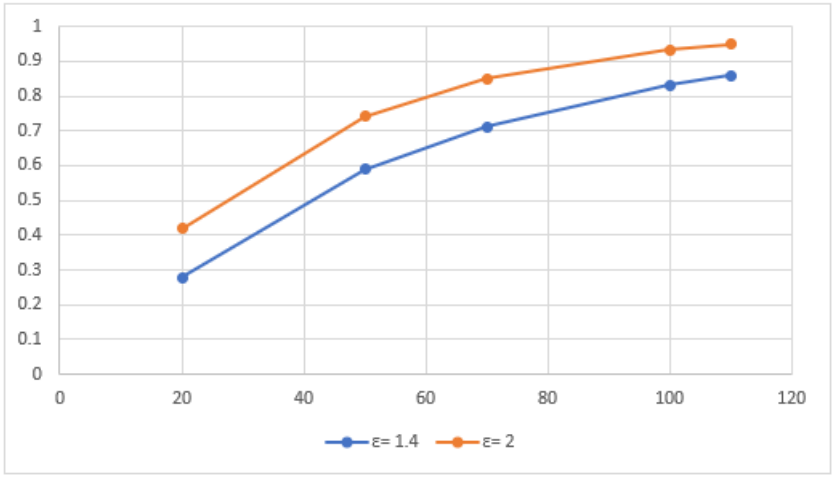


Figure 2: Coverage probability vs sample size, for the high quantile with $p = 0.995$ and $x_{0.995} = 4.719$

The Figure 1 and 2 show the coverage probability of the standard quantile estimator $\widehat{x}_p(n)$ for the true quantile value x_p, as a function of sample size and fixed other parameters. It is evident that the coverage probability increases with increasing sample size. For the low quantile $x_{0.005}$ sample sizes ranging from $2,000$ to $15,000$ provide appropriate results, while for the high quantile $x_{0.995}$, suitable results are obtained with sample sizes between 20 to 110, for considered two cases.

The following two graphs present the coverage probability depending on the sample size, for the Burr Type XII distribution with parameters $\alpha = 46$, $\theta = 470$, $\tau = 3.2$, focusing on one low quantile ($x_{0.05}$) and one high quantile ($x_{0.95}$), and two fixed values for ε.

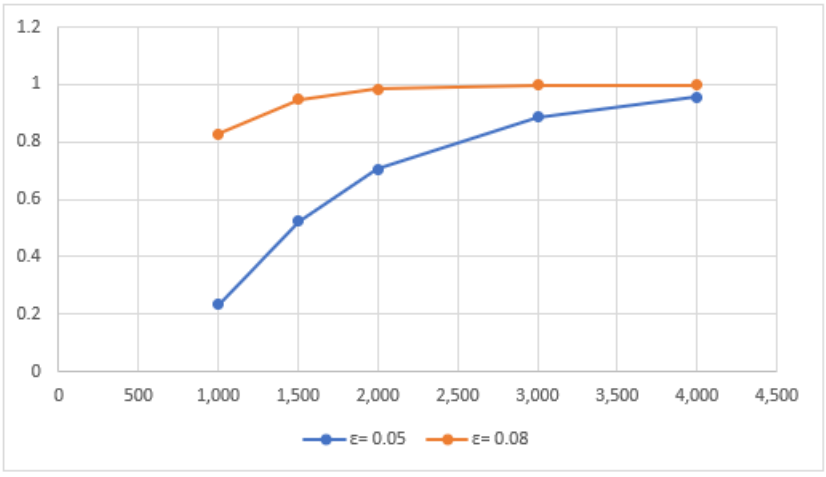


Figure 3: Coverage probability vs sample size, for the low quantile with $p = 0.05$ and $x_{0.05} = 0.817$

DRAFT VERSION

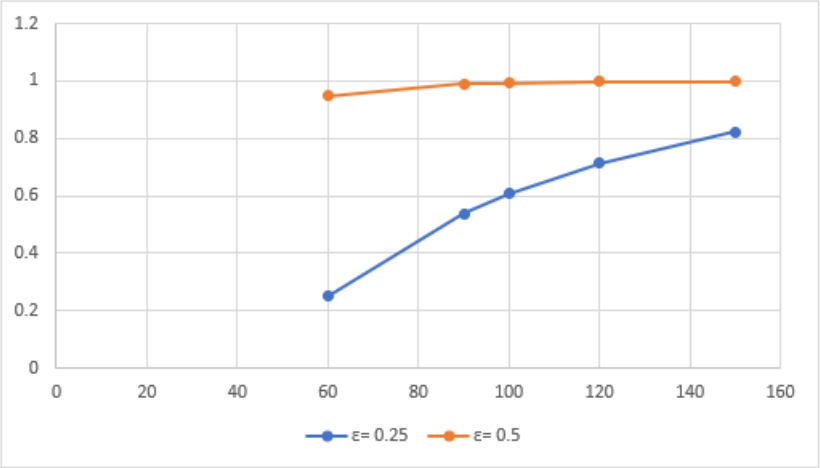


Figure 4: Coverage probability vs sample size, for the high quantile with $p = 0.95$ and $x_{0.95} = 2.943$

The Figure 3 and 4 show the coverage probability of the standard quantile estimator $\widehat{x}_p(n)$ for the true quantile value x_p, as a function of sample size and fixed other parameters. It is evident that the coverage probability increases with increasing sample size. For the low quantile $x_{0.05}$ sample sizes ranging from $1,000$ to $4,000$ provide appropriate results, while for the high quantile $x_{0.95}$, suitable results are obtained with sample sizes between 60 to 150, for considered two cases.

5. CONCLUSION

This study provides both a theoretical framework and empirical evidence for evaluating the coverage probability of the standard quantile estimator under the Burr Type XII distribution, and it is conducted under the assumptions of i.i.d. between observations; however the results may be used in the case of negatively dependent variables as well. By applying large deviations theory, we derived upper bound for the convergence rate, that describe the rate at which the estimator converges to the true quantile. The results confirmed that convergence improves with increasing sample size, in the sense that coverage probability increase with increasing sample size, and is influenced by the shape of the distribution and the quantile level. The Burr Type XII model demonstrated strong performance across parameter configurations that reflect realistic insurance loss scenarios, reinforcing its practical relevance. These results offer valuable insights for improving the reliability of quantile-based risk measures such as Value-at-Risk. Future research may explore similar bounds for alternative estimators or dependence structures, thereby broadening the applicability of this approach in actuarial and financial practice.

LITERATURE

Beirlant, J., Matthys, G., & Dierckx, G. (2001). Heavy-tailed distributions and rating. *ASTIN Bulletin: The Journal of the IAA*, *31*(1), 37–58. https://doi.org/10.2143/AST.31.1.993

Burr, I. W. (1942). Cumulative frequency functions. *The Annals of mathematical statistics*, *13*(2), 215–232. https://doi.org/doi:10.1214/aoms/1177731607

Cummins, J. D., Dionne, G., McDonald, J. B., & Pritchett, B. M. (1990). Applications of the *GB*2 family of distributions in modeling insurance loss processes. *Insurance: Mathematics and Economics*, *9*(4), 257–272. https://doi.org/10.1016/0167-6687(90)90003-V

Jin, X., & Fu, M. C. (2007). *A large deviations analysis of quantile estimation with application to value at risk*. Robert H. Smith School of Business.

Miljković, T., Causey, R., & Jovanović, M. (2022). Assessing the performance of confidence intervals for high quantiles of *B*urr *XII* and *I*nverse *B*urr mixtures. *Communications in Statistics-Simulation and Computation*, *51*(8), 4677–4699. https://doi.org/10.1080/03610918.2020.1747075

Miljković, T., & Grün, B. (2016). Modeling loss data using mixtures of distributions. *Insurance: Mathematics and Economics*, *70*, 387–396. https://doi.org/10.1016/j.insmatheco.2016.06.019

Stanojević, J., & Jovović, M. (2016). Coverage probability for one quantile estimator and its application on insurance data. *Bulletin of the international mathematical virtual institute*, *6(2016)*, 127–134.

Venter, G. (1983). Transformed beta and gamma distributions and aggregate losses. *Proceedings of the Casualty Actuarial Society*, *70*(133), 289–308.

DRAFT VERSION

TRAFFIC, TRANSPORT AND COMMUNICATIONS

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.
UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

FUEL CONSUMPTION AND CO_2 EMISSION FORECAST ON DIVERGING DIAMOND INTERCHANGE

ALEKSANDAR JOVANOVIĆ[1], KATARINA KUKIĆ[2], ANA UZELAC[3]
[1] Faculty of Engineering University of Kragujevac, Kragujevac, a.jovanovic@kg.ac.rs, 0000-0003-4208-5703
[2] University of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, k.kukic@sf.bg.ac.rs, 0000-0001-7369-2456
[3] University of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, ana.uzelac@sf.bg.ac.rs, 0000-0002-7203-9567

Abstract: *Traffic-related pollution continues to pose serious challenges to both environmental sustainability and public health. Road transportation is a major contributor, accounting for approximately 14% of global greenhouse gas emissions. This study focuses on predicting fuel consumption and emissions at Diamond Diverge Interchanges (DDIs), based on traffic demand. Accounting for the significant impact of heavy vehicle stops on overall emissions, the study introduces a novel ecological performance metric specifically designed for DDIs. Subsequently, six machine learning models are applied to predict fuel use and emissions, with the Gaussian Process Regression (GPR) model achieving the best performance for CO_2 prediction (MAE = 620.78, R^2 = 0.9989), followed closely by the Artificial Neural Network (ANN). In contrast, the Ensemble Trees model showed the weakest performance, with the highest prediction errors and the lowest R^2 value (0.9423).*

Keywords: *DDI, emissions, Bee Colony Optimization, microsimulation, machine learning.*

1. INTRODUCTION

Air pollution originating from road traffic is a major contributor to environmental harm and poses serious risks to human health. The transportation sector is responsible for nearly 14% of greenhouse gas emissions worldwide. Beyond CO_2 and fuel use, it also emits harmful substances such as nitrogen oxides (NOx), carbon monoxide (CO), and hydrocarbons (HC), which are linked to respiratory and heart-related conditions. As both vehicle numbers and travel demands grow, the environmental consequences are expected to worsen (IEA, 2019; WHO, 2019). To mitigate these issues, transportation professionals are increasingly turning to innovative intersection and interchange designs.

One such design is the Diverging Diamond Interchange (DDI), which enhances traffic operations and safety and has the potential to lower emissions and fuel use (Hughes et al., 2014). However, DDIs generally require more physical space than conventional intersections, limiting their use in densely populated urban centers. As a result, they are more often constructed in suburban areas, where there is more available land and where urban and rural traffic streams intersect. In most implementations, DDIs are controlled using fixed-time signals with two main phases. This is primarily due to the straightforward nature of such phasing, where each phase serves traffic either on the bridge or from the ramps. Although more complex phasing schemes with overlapping movements have been proposed, they have not yet been adopted in practice (Schroeder et al., 2014).

This study focuses on predicting fuel use and emissions, specifically CO_2 at DDIs with the goal to estimate environmental impact based solely on traffic intensity at different entry points of the DDI. For the purpose of this research, cycle lengths and phase splits are optimized using Bee Colony Optimization (BCO), a nature-inspired algorithm (Lučić & Teodorović, 2001). This technique has shown promise in prior applications involving alternative intersection control strategies (Jovanović & Teodorović, 2022). After optimization, simulations are conducted to record fuel consumption and emissions, forming the dataset used to train and test machine learning (ML) models. Findings highlight that the frequency of stops for heavy vehicles plays a crucial role in determining both fuel usage and emission levels. This observation offers a new angle compared to the older Robertson performance index (Robertson et al., 1980), which has been a long-standing metric in signal control evaluation. In this work, the ecological performance measure introduced by Alshayeb, Stevanovic, and Effinger (2022) is adopted as a more suitable alternative.

Machine learning has so far shown to be successful at predicting CO_2 emissions in different transport-related contexts. For example, Udoh et al. (2024) applied ML techniques to estimate emissions from light-duty vehicles using data from the UK Vehicle Certification Agency. Among six regression models tested, the Decision Tree Regressor achieved the highest accuracy, with engine power, fuel consumption, and engine

DRAFT VERSION

capacity identified as key predictors. The final model was deployed as a Streamlit-based web application, providing a practical tool for estimating emissions and supporting sustainable transport decisions. Park et al. (2023) created a model including traffic volume, speed, and wind data to forecast urban on-road CO_2 levels in Seoul. Their findings revealed significant differences in emissions across locations and times of day, demonstrating the model's utility in urban monitoring and policy planning. Using data from portable emissions measurement systems, Li et al. (2024) suggested a deep learning technique to forecast CO_2 emissions from light-duty diesel trucks. They identified key predictors such as vehicle speed, acceleration, road slope, and vehicle-specific power and created a model with good accuracy.

2. METHODOLOGY

To develop predictive models for fuel consumption and emissions at a Diverging Diamond Interchange (DDI), we first generated a comprehensive dataset using a simulation-based approach. A wide variety of traffic demand patterns were defined, covering both bridge and ramp volumes, including the proportion of heavy goods vehicles (HGVs). The methodology may be divided into subtasks, as presented in Figure 1.

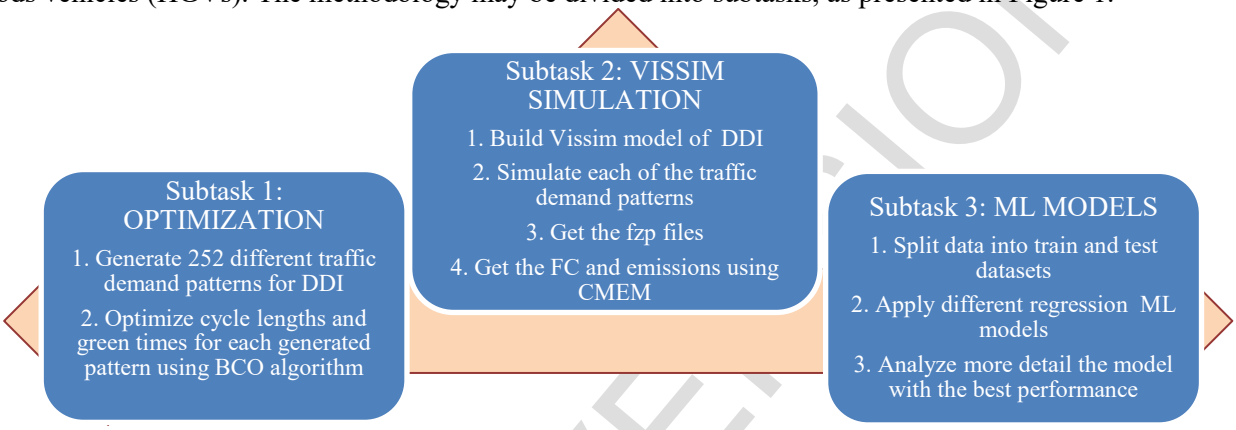


Figure 1: The methodology divided into three subtasks

Subtask 1: The goal of the fixed-time optimization at a DDI is to determine the best values for cycle length (C) and green times (g_j) that minimize the Ecological Performance Index (EPI). The optimization is subject to constraints on total cycle time ($C_{min} \leq C \leq C_{max}$), green time ranges per phase ($g_{jmin} \leq g_j \leq g_{jmax}$), and the requirement that green times and lost time together sum up to the cycle length ($\sum_{j=1}^{2} g_j = C - L$), where L is the cycle lost time, consisting of both yellow time and all-red time, each set to 3 seconds per phase. In this case, L takes the value of 12 seconds. The EPI is computed for each lane based on vehicle delay, number of stops, and a stop penalty factor, as defined by (1):

$$EPI_i = d_i + K_i \cdot S_i \quad (1)$$

where d_i is the average delay (s/veh), S_i is the average number of stops (stops/veh), and K_i is the stop penalty (Alshayeb, Stevanovic and Effinger, 2022). The total EPI for the intersection is a demand-weighted sum across all lanes (totally i traffic lanes) over the sum of traffic demands q_i at the i-th lane, and is calculated as in (2):

$$EPI = \frac{\sum_i q_i \cdot EPI_i}{\sum_i q_i} \quad (2)$$

The BCO algorithm was applied in this approach, with a population of agents (bees) iteratively searching for better signal settings, starting from feasible initial solutions and improving them through several iterations. The setup used in this study included 20 bees, 20 forward-backwards passes, one solution change per pass and a stopping condition based on CPU time of 10 seconds. At the start of the search, an initial solution is randomly generated from the feasible solution space.

Subtask 2: The modelled DDI is based on a real-world layout from Lincoln, Nebraska (USA), Figure 2 and was implemented in the VISSIM simulation environment. Simulations were run for one hour with a 150-second warm-up period, using 42 random seeds. The speed limit was set to 72 km/h, and the bridge length was 105 meters. A standard two-phase DDI signal plan was applied.

During simulations, vehicle behavior (including speed, position, acceleration, deceleration, and road slope) was recorded every second and stored in "fzp" files. These files were then processed using the Comprehensive Modal Emissions Model (CMEM) to estimate fuel use and emissions. This procedure was repeated for all traffic demand patterns, resulting in a dataset of 252 entries, where inputs are traffic demands and outputs are fuel consumption and emissions.

DRAFT VERSION

Subtask 3: The dataset was split into an 80/20 ratio to obtain 200 samples for ML models training and 50 for testing. To simplify inputs, traffic demands were grouped into bridge (A and E on Figure 2) and highway (B and F) flows. Additionally, the percentage of heavy vehicles was considered for both sources, given its strong influence on emissions, since one of the main components that influences the ecological parameters is the number of stops of heavy vehicles (Alshayeb, Stevanovic, and Effinger, 2022). As a result, four input features were defined: X_1 – traffic demands from the highway, X_2 – the percentage of heavy vehicles from the highway, X_3 – traffic demands on the bridge and X_4 – the percentage of heavy vehicles on the bridge. The Vissim & CMEM model then provided five outputs: fuel consumption (FC), carbon dioxide (CO_2), hydrocarbons (HC), carbon monoxide (CO), and nitrogen oxides (No_x). This approach enabled the creation of dataset that will be used within predictive ML models capable of learning from traffic patterns and estimating environmental impacts.

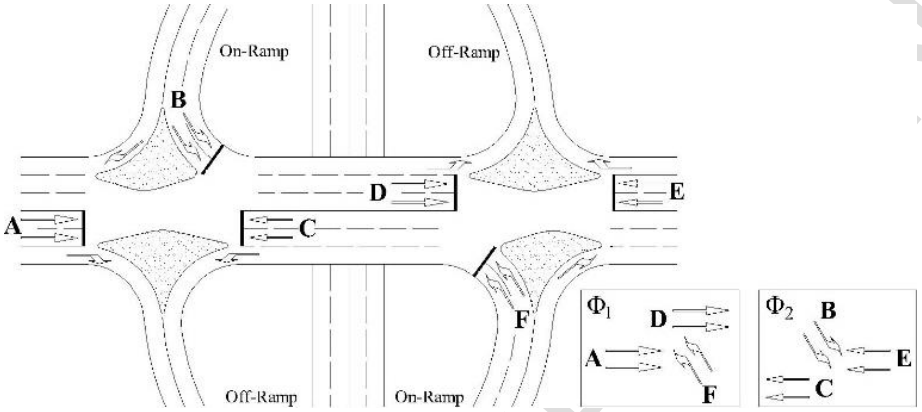


Figure 2: DDI layout

To predict fuel consumption (FC) and carbon dioxide (CO_2) emissions, this study applies a set of regression models, including both traditional and advanced approaches, chosen for their strong performance in multivariable prediction tasks. Specifically, the following regression models were considered: Multiple Linear Regression (MLR), Support Vector Regression (SVR), Random Forest Regression (RFR), Gaussian Process Regression (GPR), Ensemble Trees (Boosted Regression) and Artificial Neural Network (ANN) with a single hidden layer. Two sets of models were developed with input variables X_1, X_2, X_3 and X_4, and the output in one set of models is FC, and in the other set is CO_2.

In order to evaluate regression models, we used four commonly employed metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Mean Absolute Percentage Error (MAPE). MAE represents the average of the absolute differences between the predicted and actual values. It provides a straightforward interpretation of the average magnitude of prediction errors, without considering their direction. MAE is less sensitive to outliers, and it is calculated by formula (3):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

where n is the number of instances in the test dataset, y_i is the actual value of output for the ith data point, and $\hat{y}_i$ is the predicted value of output by the model for the ith data point.

RMSE is the square root of the average of squared differences between predicted and observed values. Because it squares the errors before averaging, RMSE penalizes larger errors more strongly than MAE, making it useful when large deviations are particularly undesirable. It is calculated by formula (4):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

R^2, or the coefficient of determination, indicates the proportion of variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, where with higher value the model explains more of the variability in the response variable. It is obtained as in (5), where $\bar{y}_i$ is mean of all observed values:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{5}$$

Finally, MAPE expresses the average absolute error as a percentage of actual values, making it useful for understanding model accuracy in relative terms. However, it should be used with caution when actual values are close to zero, as this can result in very high or undefined percentage errors. It is calculated by formula (6):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|} \tag{6}$$

DRAFT VERSION

Before presenting the obtained results, we will briefly explain several regression models applied in this study. The MLR model served as a baseline due to its simplicity and interpretability. It assumes a linear relationship between inputs and the target variable, making it easy to implement but limited in capturing complex, nonlinear patterns. The SVR model uses a kernel function to project data into a higher-dimensional space where a linear fit can approximate nonlinear relationships (Support Vector Machines algorithm for regression problems). Random Forest and Ensemble Trees (Boosting) are representatives of ensemble learning techniques. RFR builds multiple decision trees using bootstrap samples and averages their predictions, reducing variance and improving generalization. In contrast, Boosted Regression builds trees sequentially, where each new tree attempts to correct the errors of the previous one. This often leads to improved accuracy but requires careful tuning to avoid overfitting. GPR is a probabilistic, non-parametric model that assumes a distribution over functions and provides both predictions and uncertainty estimates. It is a Bayesian approach that can model certainty in predictions, especially effective in small to medium datasets and is valued for its flexibility and ability to model uncertainty explicitly, though computational cost can grow significantly with larger datasets. Lastly, ANNs are a class of ML models inspired by the structure and function of the human brain. ANNs consist of interconnected nodes (neurons) arranged in layers that process input data and make predictions. It is highly effective at capturing complex nonlinear relationships in the data. While neural networks require more data and computational resources, they offer strong predictive performance when properly trained and regularized.

3. RESULTS

All applied models were trained using default parameters provided by MATLAB's Regression Learner and Statistics and Machine Learning Toolbox. All models were trained and tested using an 80/20 split of the dataset. To ensure fair comparison, 10-fold cross-validation was performed, meaning that the given data sample is to be split into 10 groups. Recall briefly that k-fold cross-validation estimates model performance on unseen data using a k-subset of the data. It provides a more realistic evaluation than methods like a basic train/test split. This process evaluates the model design by retraining it with different data subsets. The steps are: 1. shuffle the dataset and split it into k parts (10 in this study); 2. for each part, use it as a test set and the remaining parts as a training set; 3. train the model, evaluate it, and save the score; 4. the final results represent average performance across folds, providing a more robust insight into generalization performance.

We developed six ML regression models with the following parameters:

- LMR (with `fitlm` MATLAB function): a simple, interpretable model that serves as a baseline model
- SVR (with `fitrsvm` MATLAB function): Kernel: 'linear' showed better results than 'gaussian', other parameters set to default values (BoxConstraint (C): 1, Epsilon: 0.1, Standardization: Enabled)
- RFR (with `fitrensemble` MATLAB function): Method: Bagging, 100 learning trees, Learner type: Regression trees with default depth
- Boosted Trees (with `fitrensemble` MATLAB function): Method: Least-squares boosting, 100 learning trees, Learner rate: 0.1, Trees are grown sequentially with focus on correcting previous errors
- GPR (with `fitrgp` MATLAB function): Kernel function: Squared Exponential, Basis function: Constant, Noise standard deviation and kernel parameters are estimated from the data
- ANN (with `fitnet` MATLAB function): Architecture: Feedforward NN, Hidden layer: 1 layer with 10 neurons, Training function: Levenberg–Marquardt (default function for small-to-medium datasets).

In Table 1, we present an evaluation of applied regression models for the prediction of CO_2 emissions, with the following performance metrics: MAE, RMSE, MAPE and R^2.

Table 1: Evaluation of performed regression models for the prediction of CO_2 emissions

Model	MAE	RMSE	MAPE	R^2
MLR	1618.3	2280.6	1.0736	0.99774
SVR	3542.9	4382.2	2.2076	0.99104
RFR	6110.1	9344.5	4.1557	0.95825
Ensemble Trees	8417.4	11109	5.2746	0.94233
GPR	620.78	1339.9	0.36664	0.9989
ANN	750.2	1432.7	0.45287	0.99886

Table 1 summarizes the predictive performance of six regression models. Among them, Gaussian Process Regression (GPR) achieved the best overall results, indicating excellent accuracy and model fit. Note that Multiple Linear Regression (MLR) also demonstrated strong performance, with the additional benefit of model interpretability through explicit coefficients, confidence intervals, and significance testing. So, we will briefly

DRAFT VERSION

present MLR in more detail and interpret the obtained results. Table 2 presents estimated coefficients and interpretation of the MLR Model described by the equation:

$$CO_2 emission = -53735 + 197.2 \cdot X_1 + 13264 \cdot X_2 + 97.3 \cdot X_3 + 12685 \cdot X_4 + \varepsilon$$

Table 2: Estimated coefficients and interpretation of the MLR model for prediction of CO2 emissions

Input	Coefficient	t-stat	p-value	Interpretation
Intercept	-53735	-60.8	≈ 0	Baseline emission level (has no practical meaning by itself — as it is unrealistic for all inputs to be zero).
X_1	197.2	25.2	<0.001	An increase in highway traffic demand by 1 unit increases CO_2 emissions by 197.2 g, assuming all other factors remain constant.
X_2	13264	17.6	<0.001	Each additional percentage of heavy vehicles from the highway increase leads to a 13.26 kg rise in CO_2 emissions.
X_3	97.3	11.9	<0.001	An increase in traffic demand on the bridge by 1 unit results in a 97.3 g increase in CO_2 emissions.
X_4	12685	15.9	<0.001	Each additional percentage of heavy vehicles from the bridge increases emissions by approximately 12.7 kg.

From the presented results, we conclude that the MLR model shows excellent performance and interpretability, with an R^2 of 0.998. Among the predictors, the percentage of heavy vehicles from the highway and from the bridge has the largest impact on CO_2 emissions. This result is expected, as a higher share of freight vehicles significantly increases fuel consumption and emissions.

In the same way, we developed another six ML regression models for the prediction of fuel consumption (FC) in dependence of the same four input variables. The obtained performances are presented in Table 3.

Table 3: Evaluation of performed regression models for the prediction of fuel consumption (FC)

Model	MAE	RMSE	MAPE	R^2
MLR	2858.4	3507.5	5.0792	0.95264
SVR	2824.9	3642.2	5.1302	0.93983
RFR	1998.1	3069.9	3.2852	0.95806
Ensemble Trees	2596.4	3621.6	4.2552	0.94282
GPR	309.6	477.92	0.55646	0.999
ANN	242.61	382.91	0.42099	0.99936

Among the tested models, the ANN and GPR achieved the best performance, with extremely low MAE, RMSE, and MAPE values, and R^2 values close to 1. This indicates high predictive accuracy. In contrast, linear regression and SVR performed reasonably well but showed noticeably higher error rates. Overall, the non-linear models captured the complexity of fuel consumption data more effectively than linear approaches.

Both the Gaussian Process Regression (GPR) and Artificial Neural Network (ANN) models demonstrated outstanding accuracy in predicting fuel consumption. The ANN model achieved the highest R^2 value (0.99936) and the lowest error metrics overall, indicating its strong ability to capture complex nonlinear relationships. However, due to its black-box nature, it lacks interpretability, which can be a limitation in practical transportation engineering contexts.

On the other hand, the GPR model also delivered highly accurate predictions (R^2 = 0.999), with the added benefit of providing confidence intervals around its predictions, which enhances model transparency and reliability. This feature makes GPR particularly useful in scenarios where understanding the uncertainty in predictions is crucial.

Figure 3 presents the 95% confidence intervals of the GPR model's predictions concerning input features X1 (traffic demands on the highway) and X_3 (traffic demands from the bridge), offering visual insight into how the model's confidence varies across different traffic conditions. It illustrates the effect of individual input features (X_1 and X_3) on the predicted fuel consumption using the GPR model. For each subplot, one feature is varied across its observed range while the others are fixed at their mean values. The blue line shows the model prediction, while the shaded area represents the 95% confidence interval. These plots reveal how sensitive the fuel consumption is to each input, with X_1 (traffic demand on the highway) and X_3 (traffic demand on the bridge) both showing strong and nearly linear relationships with fuel use. The narrow confidence bands indicate high model certainty in these regions.

DRAFT VERSION

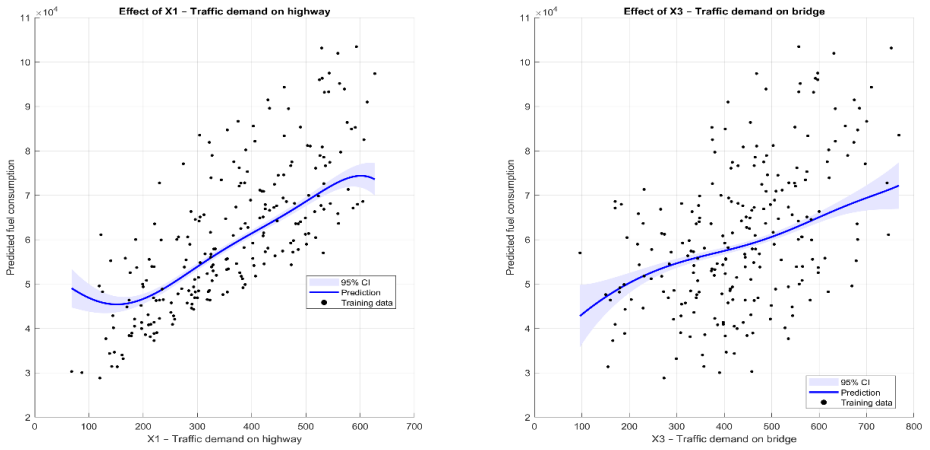


Figure 3: Confidence intervals of GPR predictions

4. CONCLUSION

This study advances the application of predictive modelling and optimization to tackle pressing issues related to traffic-related emissions and fuel consumption at Diamond Diverge Interchanges (DDIs). By relying solely on traffic demand patterns as input, we have developed models capable of accurately predicting CO_2 and fuel consumption. BCO is used to optimize signal control, demonstrating the promise of bio-inspired algorithms in enhancing ecological performance.

In addition, six machine learning algorithms were developed and applied to predict fuel consumption and CO_2 emissions. ANN and GPR models showed the best performances. Moreover, the MLR model demonstrated strong performance in predicting CO_2 emissions (R^2 of 0.998), which is why it was used for interpreting the obtained results. The model shows that all analyzed variables significantly affect CO_2 emissions ($p < 0.001$), with the percentage of heavy vehicles from the highway (X_2) being the most important variable, as it has the highest coefficient (13.26 kg per percent) and thus the strongest impact on emission increase.

Future research will be directed to applying this methodology over the whole corridor or network at classical intersections.

LITERATURE

International Energy Agency (IEA), (2019). Global EV Outlook 2019; IEA: Paris, France.

World Health Organization, (2019). Ambient air pollution: training for health care providers (No. WHO/CED/PHE/EPE/19.12. 14). World Health Organization.

Hughes, W., Jagannathan, R., Sengupta, D. and Hummer, J. (2010). Alternative intersections/interchanges: informational report (AIIR) (No. FHWA-HRT-09-060). *Federal Highway Administration. Office of Research, Development, and Technology.* United States.

Lucic, P. & Teodorovic, D. (2001). Bee system: modeling combinatorial optimization transportation engineering problems by swarm intelligence. In *Preprints of the TRISTAN IV triennial symposium on transportation analysis* (pp. 441-445).

Jovanović, A. & Teodorović, D. (2022). Fixed-time traffic control at superstreet intersections by bee colony optimization. *Transportation research record*, *2676*(4), (pp. 228-241).

Alshayeb, S., Stevanovic, A. & Effinger, J.R. (2022). Investigating impacts of various operational conditions on fuel consumption and stop penalty at signalized intersections. *International Journal of Transportation Science and Technology*, *11*(4), (pp. 690-710).

Teodorović, D., Davidović, T., Šelmić, M. and Nikolić, M. (2021). Bee colony optimization and its applications. *Handbook of AI-based Metaheuristics*, (pp. 301-322).

Udoh, J., Lu, J., & Xu, Q. (2024). Application of machine learning to predict CO_2 emissions in light-duty vehicles. *Sensors, 24*(24), 8219.

Park, C., Jeong, S., Kim, C., Shin, J., & Joo, J. (2023). Machine learning-based estimation of urban on-road CO_2 concentration in Seoul. *Environmental Research, 231*, 116256.

Li, S., Tong, Z., & Haroon, M. (2024). Estimation of transport CO_2 emissions using machine learning algorithm. *Transportation Research Part D: Transport and Environment, 133*, 104276.

DRAFT VERSION

SYM-OP-IS 2025
52nd Symposium on Operational Research
Palić, 7 - 10.9.2025.

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

Original article

SELECTION OF CITY LOGISTICS DEVELOPMENT SCENARIOS BY APPLYING A HYBRID MCDM MODEL

SNEŽANA TADIĆ[1], MLADEN KRSTIĆ[2], ALEKSA MARAVIĆ[3], MILOŠ VELJOVIĆ[4]

[1] University of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, s.tadic@sf.bg.ac.rs, 0000-0003-4651-3699
[2] University of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, m.krstic@sf.bg.ac.rs, 0000-0002-7937-0543
[3] University of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, aleksamaravic16012002@gmail.com, 0009-0004-7188-4154
[4] University of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, m.veljovic@sf.bg.ac.rs, 0000-0001-8024-3296

Abstract: *The paper presents the selection process for the development of city logistics (CL) scenarios. Respecting current trends, four scenarios are defined. Each of the scenarios has certain characteristics that can be advantages or disadvantages in a comparative analysis. Given that the choice of development scenarios is influenced by several factors, their ranking in terms of different criteria requires the application of Multi-Criteria Decision-Making (MCDM) methods. To solve the problem, a hybrid model was applied, which is based on a combination of fuzzy FActor RElationship (FARE) and Axial Distance-based Aggregated Measurement (ADAM) methods.*

Keywords: *city logistics, development scenario, MCDM, fuzzy FARE, ADAM.*

1. INTRODUCTION

As urban areas continue to expand and the rate of urbanization increases, logistics needs have become more diverse and random, and logistics operations have become more complicated. Therefore, an efficient and reliable system of city logistics (CL) has become an important part for promoting the economic development of cities and improving living standards (Tadić & Zečević, 2016; Zhang, 2022). The growth and development of traffic in goods, materials and freight follows the growth and development of CL.

CL research covers a wide range of issues, including storage, handling, distribution and other aspects, related to industry, trade, environment, etc. The planning of the CL system should take into account the requirements of companies, governments and residents and unite the interests of all parties to achieve a general system optimization (Tadić & Zečević, 2015). Numerous papers have dealt with the perspectives of the CL development (e.g. Taniguchi & Thompson, 2014). Also, many papers have evaluated the initiatives, concepts, scenarios and solutions of CL (Büyüközkan & Mukul, 2019; Tadić et al., 2014a; 2014b; 2018).

In this paper, the scenarios of the development of CL are presented and the process of their evaluation is carried out. Each scenario takes into account the interdependencies of the CL system and current trends (Tadić & Zečević, 2016a). Since the choice of scenarios depends on several factors, their evaluation requires the application of Multi-Criteria Decision-Making (MCDM) methods. To solve the problem, a MCDM model was applied, based on a combination of fuzzy FActor RElationship (FARE) and Axial Distance-based Aggregated Measurement (ADAM) methods, which achieved the main goal and contribution of the work.

After the introduction, in the second chapter, the scenarios of the development of CL are described, and in the third, the criteria for their evaluation. In the fourth chapter, the procedure for choosing the most favorable scenario using the MCDM model is presented. Finally, concluding remarks are given.

2. CITY LOGISTICS DEVELOPMENT SCENARIOS

There are different prospects and possibilities for the future development of CL, and accordingly also potential scenarios. ***Sustainable CL (A_1).*** The main cause of the negative effects of logistics on sustainability is the lack of planning activities and comprehensive, long-term CL policies (Tadić & Zečević, 2015; Kovač et al., 2022). Plans should cover the entire city, all modes of transport, all flows, respect all characteristics of the urban entity and all ecological, social and economic goals of all stakeholders (Tadić, 2014, Tadić & Zečević, 2016b). Alternative forms of transport, as well as cycle logistics, which can reduce carbon dioxide (CO_2) emissions (Allen et al., 2012), as well as greenhouse gases, particles and nitrogen oxides (Assmann et

DRAFT VERSION

al., 2020), are gaining importance. ***Autonomous CL (A_2).*** Efforts to overcome economic, sociopolitical and environmental challenges and improve the quality of life in the city led to the concept of autonomy. Innovations are of particular importance for solving the problem of home delivery, and can be seen through technologies or business models (automated and autonomous delivery systems; underground delivery; lockers, etc.). New solutions, such as autonomous vehicles, robots and drones, present the potential to improve efficiency. The main barriers and concerns of the authorities remain issues of safety and security. In addition, it is important to respect both the generators of the flow of goods (senders and receivers) and their implementers (logistics providers) when automating CL activities (Paddeu, 2023). Of particular importance in this scenario is the concept of hyper-connected CL, which is based on a greater degree of cooperation and consolidation, within and between cities. Research indicates the potential advantages of the concept, but also the need to make it more precise and comprehensive (Crainic et al., 2023). ***Current state of CL (A_3).*** City deliveries are short, frequent, smaller in volume and contain more items (Tadić & Zečević, 2016b). More and more home deliveries are being made with little or no optimization, in a fragmented way (Attard, 2023). Many local governments still ignore CL in urban planning (Cui et al., 2015). Some of the main problems of realizing flows in the city, such as (Giuliano et al., 2013): air pollution, traffic congestion, reduced mobility and noise, are the result of inadequate logistics planning (Tadić & Zečević, 2015). Deliveries are made by operators and other organizations, as well as private individuals. However, policy makers tend to ignore such developments. A large number of CL initiatives and solutions have been defined and tested, but decisions are made ad hoc, partially, so the effects are limited (Tadić & Zečević, 2016b). One of the more recent initiatives is the integration of freight and passenger transport (Nocera et al., 2023; Tadić et al., 2022), but the most common of the initiatives are regulations, and those based on bans and fines. ***Human-centered CL (A_4).*** Considering the development tendencies of various natural and anthropogenic crises, logistics "tailored to man", with a humane, resilient and even humanitarian character, may have greater importance and representation in the future, among other things in CL. It is especially important to improve the coordination of information exchange during disasters/crises (Mohammed Zain et al., 2023; Tadić et al., 2025). A more intensive development of healthcare logistics, which is very complex, is also needed (Tadić et al., 2015). Solutions such as consolidation centers can contribute to the reduction of congestion in hospitals, the optimal use of space, better service, etc. (Breen et al., 2023). Cities, especially megacities, also face other problems of existence, including food supply and poor air quality online channels play a role ("click and collect", digital platforms, mobile applications, as well as "ghost kitchens and dark stores").

3. CRITERIA FOR THE EVALUATION OF CITY LOGISTICS DEVELOPMENT SCENARIOS

There are different indicators of urban sustainability (Fehr et al., 2004), so it is possible to evaluate CL scenarios from the point of view of a large number of criteria. In this work, it was done on the basis of nine criteria that depict all aspects of sustainability (social, economic and ecological), which were chosen based on previous research (Tadić et al., 2014a; 2014b; 2018).

Mobility (C_1) refers to the movement of passenger and freight vehicles in the city. With increasing mobility, the quality of life in the city also improves. ***Accessibility (C_2)*** refers to the ability to access shipping/delivery locations. ***Occupation/release of public space (C_3)*** refers to the occupation of public areas or the possibility of their release in favor of some more attractive content. ***Safety (C_4)*** implies the improvement of traffic conditions, which are a consequence of the reduction in the number of delivery vehicles. ***Investments (C_5)*** refer to financial investments necessary for the implementation and implementation of the CL scenario. ***The possibility of implementation (C_6)*** refers to the complexity of administrative, regulatory, organizational, socio-cultural and other issues that accompany the implementation of a certain scenario. ***Modal redistribution of transport work (C_7)*** assesses the impact of scenarios on the use of alternative modes of transport. ***The complexity of the implementation of the logistics chain (C_8)*** depends on the degree of transformation of commodity flows. Certain scenarios transform distribution chains into multi-echelon systems, making them more complex to implement. ***The environmental aspect (C_9)*** is focused on reducing the negative impact of logistics activities on the environment. This refers to the reduction of fuel consumption, traffic congestion, harmful emissions into the air, emissions of noise and vibrations.

4. SELECTION OF CITY LOGISTICS DEVELOPMENT SCENARIOS

To solve the problem, a model consisting of two methods will be used, fuzzy FARE for determining the weight of criteria (Ginevičius, 2011) and ADAM for ranking of alternatives (Krstić et al., 2023).

Step 1: Defining the structure of the problem. Sets of alternatives and criteria for their evaluation are established, which was done in previous chapters.

DRAFT VERSION

Step 2: Defining scales for evaluating criteria and alternatives. The importance of criteria and alternatives is described by linguistic evaluations that are transformed into triangular fuzzy numbers (TFN) (in the case of the fuzzy FARE), i.e. crisp numerical values (in the case of the ADAM): None: N, (1, 1, 2), 1; Very low: VL, (1, 2, 3), 2; Low: L, (2, 3, 4), 3; Fairly low: FL, (3, 4, 5), 4; Medium: M, (4, 5, 6), 5; Fairly high: FH, (5, 6, 7), 6; High: H, (6, 7, 8), 7; Very high: VH, (7, 8, 9), 8; Extremely high: EH, (8, 9, 10), 9.

Step 3: Obtaining criterion weights.

Step 3.1: First, on the basis of the author's expert knowledge, experience and insight into the relevant literature (e.g. Tadić, 2014; Taniguchi, 2014; Zhang, 2022), linguistic evaluations of the criteria comparison (Table 1) were assigned, which are converted into TFN. This way of describing the relationship between the importance of the criteria is the most adequate because the fuzzy environment can adequately deal with ambiguity, inaccuracies and ambiguities in assessments. A matrix is created for evaluating the criteria $\tilde{A}$.

$$\tilde{A} = \left[\tilde{a}_{ij}\right]_{n\times n} \tag{1}$$

where $\tilde{a}_{ij}$ *= (l, m, u) = (l, m, n)* is the evaluation of the importance of criterion *i* in relation to criterion *j*. Items *l*, *m* i *u* are the lower, middle and upper value of TFNs; *n* is the number of criteria; and the following applies:

$$\tilde{a}_{ji} = -\,\tilde{a}_{ij} \tag{2}$$

and evaluation is considered consistent if:

$$\sum_{j=1}^{n} u = -\sum_{j=1}^{n} l \tag{3}$$

Table 1: Input and output data for the fuzzy FARE method

	C₁	**C₂**	**C₃**	**C₄**	**C₅**	**C₆**	**C₇**	**C₈**	**C₉**	**Crisp weight**
C₁		H								0,881
C₂										0,806
C₃	FH	M		FH			FH	M		1,108
C₄	M	FH					M			0,979
C₅	EH	VH	H	H		L	FL	VL	L	1,269
C₆	VH	H	H	H			FH	N	L	1,230
C₇	H	FH						FL		0,993
C₈	VH	H		FH					VL	1,069
C₉	FH	FH	M	M			L			1,084

Step 3.2: Obtaining the potential impact of criteria $\tilde{P}$:

$$\tilde{P} = H(n-1) \tag{4}$$

where *H* s the highest value of the scale used for estimations.

Step 3.3: Obtaining the overall impact (importance) of the criteria $\tilde{P}_j$:

$$\tilde{P}_j = \sum_{j=1}^{n} \tilde{a}_{ij}\,, \forall j = 1, \ldots, n; j \neq i \tag{5}$$

Step 3.4: Obtaining the final weights of fuzzy criteria (Table 1):

$$W = \frac{\tilde{P}_j}{\tilde{P}_H}, \forall j = 1, \ldots, n \tag{6}$$

where $\tilde{P}_H$ is the total potential importance of the obtained criteria:

$$\tilde{P}_H = \left(\min_j l^{\tilde{P}_j r}\,, \operatorname{mean}_j m^{\tilde{P}_j r}\,, \max_j u^{\tilde{P}_j r}\right) \tag{7}$$

where P_j s the actual total impact of the criteria *j*:

DRAFT VERSION

$$\tilde{P}_j^r = \tilde{P}_J + \tilde{P}, \forall j = 1, \dots, n \tag{8}$$

Step 4: Ranking of alternatives using the ADAM method.

Step 4.1: Define the matrix E. Elements are the evaluation e_{qj} of the alternative q according to criterion j:

$$E = \left[e_{qj}\right]_{m \times n} \tag{9}$$

where m is the number of alternatives, and n is the number of criteria (Table 2). The grades, as in the case of the weights of the criteria, are defined on the basis of the author's knowledge and research experience.

Step 4.2: Define a sorted decision matrix S whose elements s_{qj} are sorted ratings e_{qj} decreasing according to the importance (weight) of the criteria:

$$S = \left[s_{qj}\right]_{m \times n} \tag{10}$$

Step 4.3: Define a normalized sorted matrix N whose elements are n_{qj} normalized scores:

$$n_{qj} = \begin{cases} \dfrac{s_{qj}}{\max\limits_{q} s_{qj}}, & \text{for } j \in B \\ \dfrac{\min\limits_{q} s_{qj}}{s_{qj}}, & \text{for } j \in C \end{cases} \tag{11}$$

where B is a set of benefit criteria, and C is a set of cost criteria.

Step 4.4: Find the coordinates (x, y, z) of the reference (R_{qj}) and difficult reference (P_{qj}) points that define the complex polyhedron:

$$x_{qj} = n_{qj} \times \sin \alpha_j, \forall j = 1, \dots, n; \ \forall q = 1, \dots, m \tag{12}$$

$$y_{qj} = n_{qj} \times \cos \alpha_j, \forall j = 1, \dots, n; \ \forall q = 1, \dots, m \tag{13}$$

$$z_{qj} = \begin{cases} 0, \text{za } R_{qj} \\ w_j, \text{za } O_{qj} \end{cases}, \ \forall j = 1, \dots, n; \ \forall q = 1, \dots, m \tag{14}$$

where α_j is the angle that determines the direction of the vector that defines the value of the alternative:

$$\alpha_j = (j-1)\frac{90°}{n-1}, \forall j = 1, \dots, n \tag{15}$$

Step 4.5: Find the volumes of complex polyhedra V_q^C:

$$V_q^C = \sum_{k=1}^{n-1} V_k, \forall q = 1, \dots, m \tag{16}$$

where V_k is the volumes of the pyramids of which it is composed:

$$V_k = \frac{1}{3} B_k \times h_k, \forall k = 1, \dots, n-1 \tag{17}$$

where B_k is the area of the base of the pyramid defined by the reference and difficult reference points of two consecutive criteria:

$$B_k = c_k \times a_k + \frac{a_k \times (b_k - c_k)}{2} \tag{18}$$

a_k is the Euclidean distance between the reference points of two consecutive criteria that is obtained:

$$a_k = \sqrt{\left(x_{j+1} - x_j\right)^2 + \left(y_{j+1} - y_j\right)^2} \tag{19}$$

b_k and c_k are vector intensities corresponding to the weights of two consecutive criteria:

$$b_k = z_j \tag{20}$$

DRAFT VERSION

$$c_k = z_{j+1} \tag{21}$$

h is the height of the pyramid from the base to the top of the pyramid located in the coordinate origin (O):

$$h_k = \frac{2\sqrt{s_k(s_k - a_k)(s_k - d_k)(s_k - e_k)}}{a_k} \tag{22}$$

where s_k is the semicircumference of the triangle defined by the x coordinates of two consecutive criteria and the coordinate origin:

$$s_k = \frac{a_k + d_k + e_k}{2} \tag{23}$$

where d_k and e_k are the Euclidean distances of reference points of two consecutive criteria from the coordinate origin:

$$d_k = \sqrt{x_j^2 + y_j^2} \tag{24}$$

$$e_k = \sqrt{x_{j+1}^2 + y_{j+1}^2} \tag{25}$$

Step 5: Rank the alternatives according to the decreasing values of the volumes V_q^C $(q = 1, \dots, m)$. The ranking of alternatives was performed using ADAM 1.2-beta software (http://adam-mcdm.com).

Table 2: Input and output data for the ADAM method

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	Volume	Rank
A_1	8	2	4	7	3	2	9	3	9	0,3575	1
A_2	2	4	7	2	2	3	1	4	6	0,1804	2
A_3	5	8	2	6	7	8	5	6	3	0,1339	4
A_4	3	6	8	4	5	5	3	2	4	0,1691	3

The best-rated alternative is sustainable CL. The implementation of this scenario, although demanding in terms of investment and accompanying regulations, is expected to reconcile the often conflicting interests of stakeholders. The implementation of logistics services would be more efficient, the traffic network would be relieved, environmentally friendly modes of transport would be promoted, urban residents would be guaranteed to achieve environmental quality, etc.

5. CONCLUSION

Cities develop and change, and with that, CL. This development may lead to different scenarios in the future, and which one will prevail is not easy to assess. In this paper, the answer to this question was given using two methods, fuzzy FARE and ADAM. Development of CL in the direction of sustainability, was chosen as the most favorable solution. Bearing in mind the world's social and natural challenges, it is highly expected and desirable that CL develops in this direction. This scenario was chosen as a compromise solution because it provides an opportunity for the development of such a CL system which, in addition to increasing the efficiency of the implementation of logistics activities, also respects the needs of the community, which in most cases are related to improving the quality of life. In the context of future research, papers dealing with this topic could analyze the defined scenarios in more detail. In addition, the MCDM model could be extended to include a larger number of criteria or perspectives from different stakeholders.

LITERATURE

Allen, J., Browne, M., Woodburn, A., & Leonardi, J. (2012). The Role of Urban Consolidation Centres in Sustainable Freight Transport, *Transport Reviews*, 32(4), 473–490.

Assmann, T., Lang, S., Müller, F., & Schenk, M. (2020). Impact assessment model for the imple mentation of cargo bike transshipment points in urban districts, *Sustainability (Switzerland)*, 12(10), 4082–4101.

Attard, M. (2023). Climate Change and Urban Logistics. In Monios, J., Budd, L., & Ison, S. (Eds.), *The Routledge Handbook of Urban Logistics* (323–332), Routledge.

Breen, L., Schiffling, S., & Xie, Y. (2023). Healthcare and Urban Logistics. In Monios, J., Budd, L., & Ison, S. (Eds.), *The Routledge Handbook of Urban Logistics* (159–174), Routledge.

DRAFT VERSION

Büyüközkan, G., & Mukul, E. (2019). Evaluation of smart city logistics solutions with fuzzy MCDM methods, *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 25(9), 1033–1040.

Crainic, T. G., Klibi, W., & Montreuil, B. (2023). Hyperconnected city logistics: a conceptual framework. In Marcucci, E., Gatta, V., & Le Pira, M. (Eds.), *Handbook on City Logistics and Urban Freight* (398–421), Edward Elgar Publishing.

Cui, J., Dodson, J., & Hall, P.V. (2015). Planning for urban freight transport: An overview, *Transport Reviews*, 35(5), 583–598.

Fehr, M., Alves de Sousa, K., Pereira, A., & Pelizer, L. (2004). Proposal of Indicators to assess urban sustainability in Brazil, *Environment, Development and Sustainability*, 6(3), 355–366.

Ginevičius, R. (2011). A new determining method for the criteria weights in multicriteria evaluation. *International Journal of Information Technology & Decision Making*, 10(6), 1067–1095.

Giuliano, G., O'Brien, T., Dablanc, L., & Holliday, K. (2013). *Synthesis Of Freight Research In Urban Transportation Planning*, NCFRP Report 23, Washington DC.

Kovač, M., Tadić, S., & Krstić, M. (2022). Planning and development of sustanable logistics systems at a micro-level, *Tehnika*, 69(2), 225–231.

Krstić, M., Agnusdei, G.P., Tadić, S., Kovač, M., & Miglietta, P.P. (2023). A Novel Axial-Distance-Based Aggregated Measurement (ADAM) Method for the Evaluation of Agri-Food Circular-Economy-Based Business Models. *Mathematics*, 11(6).

Mohammed Zain, R., Mohd Zahari, H., & Mohd Zainol N.A. (2023). Inter-agency information sharing coordination on humanitarian logistics support for urban disaster management in Kuala Lumpur, *Frontiers Sustainable Cities*, vol. 5.

Paddeu, D. (2023). New technologies and autonomous vehicles for urban goods distribution. In Marcucci, E., Gatta, V., & Le Pira, M. (Eds.), *Handbook on City Logistics and Urban Freight* (444–462), Edward Elgar Publishing.

Tadić, S. (2014). *Integrated city logistics solutions performance modelling*. University of Belgrade, Faculty of Transport and Traffic Engineering, PhD thesis [in Serbian].

Tadić, S., & Zečević, S. (2015). Integrated planning aimed at sustainability city logistics solutions, *Tehnika*, 65(1), 164–173.

Tadić, S., & Zečević, S. (2016a). Global trends and their impact on city logistics management, *Tehnika*, 71(3), 459–464.

Tadić, S., & Zečević, S. (2016b). *Modelling city logistics concepts*. Faculty of Transport and Traffic Engineering, Belgrade [in Serbian].

Tadić, S., Veljović, M., & Zečević, S. (2022). Crowd logistics: household as a logistics service provider, *International Journal for Traffic and Transport Engineering*, 12(1), 111–122.

Tadić, S., Veljović, M., Krstić, M., & Zečević, S. (2025). Humanitarian logistics: a framework for structuring and research, *Transportation Research Procedia, 21th international conference on transport science*, ICTS 2024, Portoroz, Slovenia, 20-21.05. 2024, 83, 72–79 [in Serbian].

Tadić, S., Zečević, S., & Krstić, M. (2014a). A novel hybrid MCDM model based on fuzzy DEMATEL, fuzzy ANP and fuzzy VIKOR for city logistics concept selection, *Expert Systems with Application*, 18, 8112–8128.

Tadić, S., Zečević, S., & Krstić, M. (2014b). Ranking of logistics system scenarios for central business district, *Promet-Traffic & Transportation*, 26(2), 159–167.

Tadić, S., Zečević, S., & Krstić, M. (2018). Assessment of the political city logistics initiatives sustainability, *Transportation Research Procedia*, 30, 285–294.

Taniguchi, E. (2014). Concepts of city logistics for sustainable and liveable cities. *Procedia-social and behavioral sciences*, 151, 310-317.

Taniguchi, E., & Thompson, R. G. (Eds.). (2014). *City logistics: Mapping the future*. CRC Press.

Zhang, H. (2022). *Reliability Optimization of Urban Logistics Systems*. Springer.

DRAFT VERSION