Warsaw University of Life Sciences – SGGW
Institute of Economics and Finance
Department of Econometrics and Statistics

QUANTITATIVE METHODS IN ECONOMICS

METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH

Volume XXVII, No. 1

Warsaw 2026

ISSN 2082-792X
e-ISSN 2543-8565

Department of Econometrics and Statistics, WULS – SGGW
(Katedra Ekonometrii i Statystyki SGGW w Warszawie)

Warsaw 2026, Volume XXVII, No. 1
Journal homepage: https://qme.sggw.edu.pl

Warsaw University of Life Sciences Press, Nowoursynowska 161, 02-787 Warsaw
Tel. 22 593 55 23
e-mail: wydawnictwo@sggw.edu.pl
wydawnictwo.sggw.edu.pl
Wydawnictwo SGGW
wydawnictwosggw

CONTENTS

QUANTITATIVE METHODS IN ECONOMICS
Volume XXVII, No. 1, 2026, pp. 1 – 26

Received: 02.02.2026
Accepted: 25.03.2026

BEATING THE MACHINE: EVALUATING GEMINI LLM AND SEASONAL RANDOM WALK MODELS FOR EARNINGS PER SHARE FORECASTING IN POLAND

Wojciech Kuryłek https://orcid.org/0000-0003-0692-3300
University of Warsaw, Faculty of Management, Warsaw, Poland
e-mail: wkurylek@wz.uw.edu.pl

Abstract. Emerging markets like Poland face limited analyst coverage, with only 20% of listed companies receiving professional scrutiny, necessitating automated forecasting methods. This study investigates whether Large Language Models can outperform traditional statistical approaches in predicting earnings per share (EPS). We compare Google's Gemini LLM against the seasonal random walk (SRW) model using quarterly EPS data from 267 Polish companies (2010-2019). Performance is evaluated using Mean Arctangent Absolute Percentage Error (MAAPE), with robustness checks using RMSE and MAE metrics. Results reveal a notable discrepancy: SRW achieves superior MAAPE scores, while Gemini excels on RMSE and MAE measures. This metric-dependent performance suggests model selection should align with specific error tolerance requirements. Despite Gemini's sophisticated chain-of-thought reasoning mimicking analyst approaches, simpler models prove equally effective in markets with unsophisticated EPS dynamics. These findings challenge assumptions about LLM superiority in financial forecasting and highlight the continued relevance of parsimonious statistical methods in emerging markets.

Keywords: earnings per share, seasonal random walk, Large Language Model, LLM, Gemini, Warsaw Stock Exchange

JEL codes: C01, C02, C12, C14, C58, G17

INTRODUCTION

The fundamental determination of corporate equity valuations is intrinsically linked to the product of the Price-to-Earnings (P/E) ratio and Earnings Per Share (EPS), constituting a critical component in investment evaluation methodologies. Precise

https://doi.org/10.22630/MIBE.2026.27.1.1

prognostication of these financial indicators, with particular emphasis on EPS, remains imperative as they provide essential quantitative data regarding an organization's future performance, thereby informing market valuation estimations and establishing parameters for audit expectations. The significant disparity between mature financial markets, exemplified by the United States with its comprehensive analyst coverage, and developing markets such as Poland, where merely one-fifth of listed entities receive professional analytical scrutiny, necessitates the implementation of computational and algorithmic approaches for EPS prediction in these less-covered economies.

This study addresses two primary research questions: (1) Can Large Language Models trained on extensive textual data provide more accurate EPS forecasts than traditional time series models for Polish listed companies? (2) How does the choice of error metric influence the comparative assessment of LLM versus statistical model performance? The hypothesis of Gemini superiority is partially supported: rejected when evaluated by MAAPE but supported under RMSE and MAE metrics. This work represents the inaugural application of LLMs combining textual and time series inputs for EPS forecasting. The deployment of these advanced architectures for earnings prediction constitutes uncharted territory in academic research, filling an important scholarly void. The outcomes illuminate the capabilities of AI-based techniques for financial analysis, particularly benefiting markets with sparse analyst attention.

The research herein distinctively investigates the utilization of a Large Language Model (LLM) for temporal sequence prediction of EPS values. Large Language Models demonstrate considerable capacity for processing and analyzing extensive unstructured datasets, representing sophisticated instruments for financial examination. The investigation specifically employs Google's Gemini-2.0-flash-lite model, incorporating comprehensive prompting techniques and formalized output structures. The analytical framework encompasses quarterly EPS information from 267 corporations listed on the Polish securities exchange, with temporal boundaries extending from the global financial crisis of 2008-2009 through the COVID-19 pandemic of 2020.

This investigation adopts the mean arctangent absolute percentage error (MAAPE) as introduced by Kim and Kim (2016) as an evaluation criterion, rather than exclusively utilizing the conventional mean absolute percentage error (MAPE) methodology, which exhibits susceptibility to distortion when denominators approach zero, thus enabling more accurate assessment of predictive performance.

The scholarly contribution of this work lies in its pioneering application of LLM technology to financial forecasting, specifically in the domain of EPS prediction, representing one of the earliest research endeavors in this field. The implementation of such sophisticated computational frameworks for this predictive task constitutes an unexplored area within the current academic literature, thereby addressing a significant research lacuna. The empirical findings offer meaningful perspectives regarding the efficacy of artificial intelligence-driven methodologies

for financial prognostication. Furthermore, the research enhances comprehension of EPS forecasting mechanisms within emerging economic contexts, with particular attention to the Polish market, a domain that has received comparatively limited academic attention relative to established markets such as the United States. Multiple error measurement techniques, chronological periods, and statistical verification procedures are utilized to substantiate the conclusions, while simultaneously elucidating practical ramifications for investment tactics involving Polish equities.

The purpose of this study is to evaluate whether Large Language Models, specifically Google's Gemini, can provide superior earnings per share (EPS) forecasting accuracy compared to traditional statistical methods for companies listed on the Warsaw Stock Exchange. This investigation aims to determine if the sophisticated capabilities of LLMs offer meaningful advantages over simpler approaches like the seasonal random walk model in markets characterized by limited analyst coverage and relatively unsophisticated EPS dynamics.

LITERATURE REVIEW

The trajectory of Earnings per Share (EPS) forecasting techniques commenced in the 1960s with algorithmic approaches, predominantly utilizing autoregressive integrated moving average (ARIMA) frameworks. These methodologies were investigated by numerous scholars including Ball and Watts [1972], Watts [1975], Griffin [1977], Foster [1977], and Brown and Rozeff [1979]. The efficacy of these analytical structures showed considerable variation; certain investigations indicated that basic random walk models often demonstrated comparable performance to more intricate systems, while alternative research yielded divergent conclusions. Analogous examinations for the Polish financial environment were subsequently conducted by Kuryłek (2023a).

The prominence of ARIMA frameworks persisted due to their predictive precision, as documented by Lorek [1979] and Bathke and Lorek [1984], until the latter part of the 1980s when a prevailing academic consensus emerged suggesting that forecasts generated by financial analysts exceeded the accuracy of time series model predictions [Brown et al., 1987]. The research of Conroy and Harris [1987] emphasized that analysts' projections demonstrated superior performance for immediate temporal horizons, though this advantage diminished across extended timeframes. This academic perspective remained largely unchallenged until contemporary scholarship [Lacina et al., 2011; Bradshaw et al., 2012; Pagach and Warr, 2020; Gaio et al., 2021] began to reevaluate the purported superiority of analyst forecasts compared to time series methodologies.

Concurrent with ARIMA developments, exponential smoothing techniques for EPS prediction have been explored since the late 1960s by various researchers including Elton and Gruber [1972], Ball and Watts [1972], Johnson and Schmitt [1974], Brooks and Buckmaster [1976], Ruland [1980], Brandon et al. [1987], and

Jarrett [2008], yielding inconsistent outcomes. Within the Polish market context, Kuryłek [2023b] conducted parallel investigations into these methodologies.

Multivariate cross-sectional frameworks incorporating fundamental indicators derived from financial ratios have demonstrated superior performance compared to firm-specific and common-structure ARIMA models in earnings forecasting. This superiority has been documented in research by Lorek and Willinger [1996], Lev and Thiagarajan [1993], Abarbanell and Bushee [1997], Cao and Gan [2009], Cao and Parry [2009], Ahmadpour et al. [2015], Ball and Ghysels [2017], Hou et al. [2012], and Li and Mohanram [2014]. Nevertheless, Lev and Sougiannis [2010] observed the constrained long-term effectiveness of estimate-based accounting elements. Conceptual frameworks for earnings prediction were established by Ohlson [1995, 2001], Pope-Wang [2005, 2014], and Li [2011], with Harris and Wang [2019] confirming the enhanced accuracy and reduced bias exhibited by Pope and Wang's [2005] model.

Contemporary research has increasingly concentrated on artificial neural networks for EPS forecasting, producing diverse results. Atiya et al. [1997] were pioneers in applying neural networks based on fundamental characteristics for stock price prediction, consistently outperforming alternative models. Cao et al. [2004] conducted comparative analyses of neural feedforward networks (MLP), demonstrating their superior accuracy relative to other forecasting methodologies. Conversely, Lai and Li [2006] reported that an ANN framework exhibited the least satisfactory accuracy for EPS predictions. Research by Cao and Parry [2009] illustrated that univariate neural network models surpassed linear regression frameworks and that genetic algorithms excelled in determining network weights. Cao and Gan [2009] verified the enhanced performance of neural network models optimized with genetic algorithms for predicting EPS of Chinese listed enterprises. Gupta et al. [2013] identified an optimal multiperceptron architecture for stock market price forecasting, emphasizing the significance of EPS and public sentiment. Ahmadpour et al. [2015] successfully employed standard multilayer perceptron (MLP) neural networks, achieving substantially higher accuracy with extracted rules compared to pure MLP frameworks. Chen et al. [2020] assessed various methodologies for EPS prediction, including decision trees and radial basis function networks, demonstrating the superior accuracy of ensemble approaches. Elend et al. [2020] contrasted Long Short-Term Memory (LSTM) networks with Temporal Convolution Networks (TCNs) for EPS prediction, with LSTM significantly enhancing accuracy compared to naive persistent models. Xiaoqiang [2022] furnished a comprehensive overview of machine learning techniques applicable to financial ratio forecasting, including EPS.

In the most recent academic landscape, Large Language Models (LLMs) have garnered significant scholarly attention. Cao et al. [2024] demonstrated that LLMs, which extract more comprehensive and predictive content from earnings conference calls, outperform traditional analytical benchmarks in forecasting stock price volatility. Kong et al. [2024] elucidated both the capabilities and limitations of

finance-specific large language models in addressing complex tasks, while identifying critical challenges including data quality issues, modeling complexities, and ethical considerations. Research by Abe et al. [2024] examined how LLMs can predict price movements in stock and bond portfolios utilizing economic indicators, revealing that LLM-based strategies, particularly when combined with mode ensemble techniques, outperform buy-and-hold strategies in terms of Sharpe ratio during periods of increasing consumer price index (CPI). Sarker (2024) expressed concerns regarding the trustworthiness of LLMs, characterizing these models as opaque systems. Cao and Wang [2024] emphasized the principal challenges encountered by LLMs in time series prediction contexts, observing that predictive accuracy significantly decreases when confronted with heterogeneous time series data and traditional signals containing both periodic and trend components, as well as when signals comprise complex frequency elements. Abdelsamie and Wang [2024] conducted comparative analyses of market prediction accuracy between LLM-based systems and human expertise within financial analysis domains, demonstrating that their specialized model achieved superior accuracy and efficiency in financial forecasting compared to human predictions, particularly in dynamic, information-rich contexts. Nevertheless, their research acknowledged that limitations in nuanced contextual comprehension and adaptability persist, underscoring the enduring value of human expertise. It merits emphasis that the existing literature contains no studies specifically dedicated to EPS forecasting utilizing LLMs.

The inaugural investigation of statistical forecasting methodologies for EPS among Polish listed companies was conducted by Kurylek [2023a]. This research, which employed SARIMA-type models, determined that the seasonal random walk (SRW) most effectively captures the behavior of Earnings Per Share in Poland, with no alternative model demonstrating superior performance. This finding received further corroboration for models within the exponential smoothing family [Kuryłek, 2023b]. In subsequent research, Kurylek (2024a) illustrated that even contemporary time series models developed by major technology organizations – including Facebook's Prophet, LinkedIn's SilverKite, Amazon's DeepAR, and Google's TFT – fail to exceed the performance of the SRW model in the Polish market context. Similarly, various Artificial Neural Network architectures, whether implemented in time series or multivariate configurations, demonstrated no enhancement in results [Kuryłek, 2024b; Kuryłek, 2024c]. Additionally, Kuryłek [2025] investigated the potential application of natural language processing (NLP) techniques, including FastText and FinBERT word embeddings combined with the gradient-boosting decision tree algorithm XGBoost, for EPS forecasting in Poland. This investigation concluded that the implementation of these sophisticated NLP methodologies may not be justified, as the SRW model provides a more accurate representation of market behavior. These investigations collectively indicate that neither statistical frameworks nor more advanced machine learning and deep learning methodologies have demonstrated the capacity to outperform the elementary seasonal random walk

(SRW) model in Poland. The current research examines the application of Gemini LLM (Google DeepMind, 2023) to EPS forecasting.

DATA AND METHODS

Time Series Data

Following its European Union integration in 2004, the Polish stock market demonstrated remarkable expansion, reaching a capitalization of $197 billion by late 2021 with 774 companies listed. However, analytical coverage remains deficient compared to Western European and American markets, with merely 20% of listed entities receiving analyst evaluation as of 2019. This analytical deficit highlights the necessity for implementing machine learning and statistical methodologies to forecast crucial financial metrics. The current investigation concentrates on earnings per share (EPS) information extracted from the financial analysis system EquityRT.

The investigation analyzes EPS patterns for Warsaw Stock Exchange corporations spanning from Q1 2010 through Q4 2019, positioned between two major disruptive events: the 2008-2009 financial downturn and the 2020 COVID-19 outbreak. Global market volatility has intensified substantially since 2020, initially due to the pandemic and subsequently aggravated by energy instability following the Ukrainian conflict in 2022. This context establishes 2019 as the final year of comparative stability, rendering it an appropriate foundation for model calibration and validation.

The deliberate and strategic selection of this stable timeframe was intentional. A statistical or machine learning approach, including any Large Language Model, that performs inadequately during periods of stability would likely produce unreliable outcomes in volatile conditions. Stable periods offer optimal environments where patterns are more discernible, noise is reduced, and variable relationships maintain relative consistency. Models failing under stable circumstances presumably lack capacity to identify essential dependencies, trends, or fundamental data structures. During environmental volatility, these limitations become amplified, as regime shifts, sudden disruptions, and nonlinear dynamics introduce complexities beyond simple historical extrapolation. Models exclusively based on historical trends may inadequately interpret rapid changes, generating significant errors. Additionally, volatility typically enhances uncertainty, emphasizing requirements for flexible models capable of distinguishing between temporary fluctuations and enduring structural alterations. Models demonstrating fragility in stable environments become even less reliable under unpredictable circumstances, underscoring the importance of establishing dependable baselines before application in challenging scenarios.

For forecasting purposes, data encompassing Q1 2010 through Q4 2018 (36 quarters total) serves as model training material, while Q1 2019 through Q4 2019 is reserved for out-of-sample validation. The forecasting framework includes

projections ranging from one to four quarters ahead, with additional validation utilizing 2017 and 2018 samples. This dataset provides substantial accounting information, even for publicly traded entities issuing quarterly financial statements, representing at minimum nine complete years of EPS records. Firms that discontinued publishing financial reports within the study horizon were removed from the sample. Following the application of full-time window coverage and the exclusion of stock splits and reverse splits, the final sample consisted of 267 companies. Subsequently, these EPS time series undergo transformation into textual format and incorporation into prompts for Large Language Model processing.

Textual Data

The Elektroniczny System Przekazywania Informacji (ESPI) operates as an electronic disclosure framework through which public entities on the Warsaw Stock Exchange's (WSE) primary market fulfill mandatory reporting obligations. This system facilitates transparent and immediate dissemination of critical information encompassing financial outcomes, substantial ownership modifications, organizational developments, and additional relevant occurrences that might influence securities valuation or investment determinations. The Polish Financial Supervision Authority (Komisja Nadzoru Finansowego, KNF) maintains regulatory oversight regarding ESPI reporting compliance among WSE-listed corporations, thereby enforcing market transparency standards and investor protection measures.

Concurrently, the Elektroniczna Baza Informacji (EBI) functions as a complementary digital information dissemination infrastructure predominantly utilized by corporations trading on NewConnect – an alternative exchange platform administered by the Warsaw Stock Exchange (WSE) designed for emerging enterprises and developmental ventures, particularly those in formative phases or technology sectors. The EBI framework ensures punctual and transparent communication of fundamental corporate information to the investment community and general public. While operational management of the EBI system resides with the WSE, the KNF exercises supervisory authority concerning compliance with disclosure requirements, thus maintaining market integrity and safeguarding investor interests.

HTML5 (Hypertext Markup Language version 5) serves as the technological foundation for both ESPI and EBI systems, ensuring report accessibility, legibility, and interactivity across diverse computational devices and internet browsers. The Python software package html2text performs conversion of individual corporate HTML documents into simplified, comprehensible ASCII text reports. These textual documents are chronologically arranged for each corporate entity per quarter and subsequently consolidated into unified quarterly textual compilations. The resultant textual data is thereafter incorporated into prompts that subsequently serve as input for Large Language Models (LLM).

Models

The seasonal random walk model (SRW)

The SRW can be described as:

$$EPS_t = EPS_{t-4} + \varepsilon_t \text{ where } \varepsilon_t \text{ are IID and } \varepsilon_t \sim N(0, \sigma^2) \quad (1)$$

The predictive methodology employs the earnings value from the corresponding quarter of the previous year as its prognostic output, thus circumventing any requirement for parametric estimation procedures. It implies that for most of the stocks, the best predictor of the future EPS is the EPS in the analogous quarter of the previous year. This particular forecasting technique functions as a comparative standard, whose efficacy surpassing that of alternative temporal sequence analytical models within the Polish financial marketplace has been substantiated through empirical investigations conducted by Kuryłek (2023a, 2023b, 2024). The research corpus established by Kuryłek across multiple publications demonstrates the relative advantages of this parameter-free approach when applied specifically to Polish market conditions.

Large Language Models (LLMs)

Large Language Models (LLMs) represent a significant advancement in the natural language processing (NLP) domain. The developmental trajectory of these computational systems has witnessed a transformation from rudimentary frameworks to intricate architectures. Contemporary models are engineered to interpret and generate human-like linguistic outputs through the utilization of comprehensive datasets and sophisticated neural network configurations.

The fundamental processing methodology encompasses the segmentation of textual inputs into tokens (lexical units or sub-lexical components), subsequently transformed into numerical representations – a dual procedure termed tokenization and word embedding. Sequential information is preserved via positional encoding, which integrates the ordinal arrangement of tokens, thus maintaining syntactic integrity. The structural framework consists of multiple hierarchical arrangements of self-attention mechanisms and feed-forward recurrent neural networks, with each successive layer enhancing the representational fidelity of input tokens. As Gomez (2017) elucidates, self-attention mechanisms allocate relative importance to diverse tokens within input sequences, effectively capturing extended contextual dependencies. This capability directs the model's focus toward contextually pertinent input segments, culminating in the conceptualization of transformer architecture, initially formulated as an encoder-decoder paradigm.

Transformers were conceptualized with encoder-decoder dualities for specialized linguistic tasks such as translation. However, contemporary LLMs adopt a decoder-exclusive variant optimized for autoregressive textual production. This streamlined transformer iteration processes sequential data while minimizing computational complexity. The architecture employs self-attention mechanisms alongside feed-forward recurrent neural networks for the interpretation and

generation of sequential data. This configuration demonstrates particular efficacy in generating contextually coherent output sequences predicated on antecedent tokens – LLMs undergo training to anticipate subsequent linguistic elements across extensive textual corpora. In contrast to comprehensive attention mechanisms in encoders, decoders implement masked self-attention, ensuring that positional predictions are exclusively informed by preceding tokens, a critical feature for autoregressive functionalities. The dependence on previously generated sequences for contextual comprehension may present challenges with extensive sequential inputs, potentially resulting in diminished retention of initial information as sequence processing advances.

The training methodology incorporates vast textual corpora to predict subsequent tokens in sequences, thereby acquiring linguistic patterns, grammatical structures, factual information, and rudimentary reasoning capabilities. Training datasets typically encompass diverse textual sources including literary works, academic publications, digital content, and miscellaneous textual repositories. The training protocol leverages unstructured textual data, utilizing inherent contextual information for supervisory guidance.

Despite technological advancements, LLMs exhibit certain limitations. The assimilation of extensive sequential data may result in the propagation of inherent biases present within training corpora, with bias mitigation remaining an active research domain. While LLMs demonstrate proficiency in generating innovative textual formats, occasional factual inaccuracies - termed "hallucinations" - may manifest in generated content. Furthermore, substantial computational resources and memory allocation are requisite, particularly during the training phase.

The GPT (Generative Pre-trained Transformer) series represents a preeminent exemplar within the LLM category. In 2018, OpenAI introduced the inaugural GPT model, employing a transformer-based architecture featuring unsupervised pre-training followed by task-specific optimization. The subsequent iteration, GPT-2, emerged in 2019 with expanded parametric dimensions and augmented training data, achieving superior performance across various NLP tasks without specialized fine-tuning. Released in 2020, GPT-3 further expanded to 175 billion parameters, exhibiting exceptional linguistic comprehension and generative capabilities with minimal additional training. The parametric scale of GPT-4, introduced in 2023, remains undisclosed.

In their comparative analysis, Hou and Lian (2024) examine three prominent large language models - ChatGPT, Mistral, and LLaMA - assessing their performance across dimensions of computational efficiency, linguistic precision, and ethical alignment. Research findings reveal distinctive strengths and specific constraints associated with each model: ChatGPT demonstrates superiority in linguistic accuracy, LLaMA exhibits enhanced adaptability across multilingual contexts, while Mistral introduces novel approaches to complex linguistic processing. This comparative evaluation provides substantive insights into contemporary LLM capabilities.

Current research endeavors aim to further expand these models, including investigations into multi-trillion parameter architectures. Despite persistent challenges including computational demands and inherent biases, LLMs continue their evolutionary trajectory, fostering innovations in machine-based linguistic comprehension and generation. These computational systems, exemplified by models such as GPT, have transformed the NLP landscape through the implementation of transformer architectures, facilitating advanced linguistic interpretation and production. Notwithstanding ongoing technical and ethical challenges, these models continually expand the boundaries of machine-facilitated natural language processing.

Gemini (GEMINI)

The research employs Google DeepMind's Gemini (version 2.0-flash-lite), a series of multimodal large language models (LLMs) introduced in 2023. Forecasts for earnings per share (EPS) are generated through API interfaces, necessitating 1,068 distinct API calls (representing 267 companies across 4 quarters) to produce comprehensive annual projections. These models were estimated individually using the data of each company.

Prompts

Input texts or instructions, known as prompts, are provided to the LLM to elicit specific outputs. While user prompts address particular tasks, system prompts establish the overarching operational framework across all interactions, functioning as the foundation for the AI's behavioral patterns and output generation.

```
system_prompt = f"""
You are an AI expert specializing in making financial predictions based on
historical quarterly data, and all information available until {Date:%Y-
%m-%d}.
Your task is to analyze the provided historical data and make a prediction
of
Earnings Per Share (EPS) exactly one year ahead.

Do the following steps:
0. Convert all relevant information into the Polish Zloty, if this
information
   is provided in other currency.
1. Find factors that will influence the future revenues of companies and
make
   the forecast of future revenues. -> Revenues
2. Find factors that will influence the future costs of companies and make
   the forecast of future costs. -> Costs
3. Calculate the tax factor for the last quarters by dividing Net Profit
by
   Gross Profit (Net Profit / Gross Profit) and make the forecast of tax
   factor in the future. -> Tax Factor
4. Find the information about number of Shares Outstanding in the last
quarter.
   Assume that Shares Outstanding should remain constant in the future.
   -> Shares Outstanding
5. Calculate the future EPS as (Revenues - Costs) * Tax Factor / Shares
Outstanding
"""
```

```
user_prompt = f"""
Forecast EPS for the anonymized ticker: {anonymized_ticker}
Available historical EPS data: {historical_EPS_data}
ESPI reports for the available last 5 quarters:
ESPI report 0: date of information_cut: {ending_date_0:%Y-%m-%d}
   espi reports: {espi_reports_0}
...
ESPI report 4: date of information_cut: {ending_date_4:%Y-%m-%d}
   espi reports: {espi_reports_4}
Examples of quarterly data and forecasts made on the basis of them:
Example 0: ... Example 4: ...
Make the forecast accurately and consistently. Do not hallucinate.
REMEMBER: Return ONLY valid JSON with the structure specified.
"""
```

The implemented system prompt instructs the AI to generate EPS predictions utilizing all available information up to a specified date, projecting one year forward with EPS forecasts decomposed into revenues, costs, and tax components. Tickers and names of companies are intentionally anonymized to prevent from the data leakage between the LLM's training data and the test period. The user prompt incorporates five forecast examples utilizing actual EPS figures, along with historical EPS data and ESPI reports spanning from five quarters to one year prior to the forecast date. These prompts implement Python's f-strings (string literals with an 'f' prefix containing values in curly braces {} that are interpolated into the text). Multiline text elements are denoted using triple quotes.

Structured output

The model generates structured/JSON formatted output utilizing the pydantic library in Python. Beyond EPS projections, the AI provides reasoning supporting each forecast and a certainty metric ranging from 0 to 1, indicating confidence levels. As Chong and Kao (2025) observe, in volatile financial markets where professional judgment frequently guides decision-making, the quality of reasoning can supersede predictive accuracy in importance. Consequently, explanatory model reasoning was deemed essential, implemented as a chain of thought for each prediction. This technique enables Large Language Models to decompose complex problems into sequential steps, explicitly articulating intermediate reasoning before reaching conclusions, thereby enhancing problem-solving accuracy and making logical processes transparent to users. Additionally, the model assigns a certainty score between 0 and 1 to each forecast.

Setting temperature

The temperature parameter critically balances predictability and creativity in generated text. Lower settings prioritize established patterns for more deterministic results, while higher values promote exploration, enhancing output diversity. Temperature was calibrated to minimize Mean Arctangent Absolute Percentage Error (MAAPE) for 2019, with comparative testing at settings of 0.0, 0.2, and 0.4, yielding MAAPE values of 0.743, 0.725, and 0.731 respectively. The empirical analysis identified 0.2 as the optimal temperature parameter.

Mean Arctangent Absolute Percentage Error (MAAPE)

For a specific corporate entity i, the empirical earnings per share (EPS) throughout the quarterly sequence of 2019 may be denoted as A1, ..., A4. The corresponding prognostic values for these temporal intervals facilitate the computation of forecast precision via the absolute percentage error (APE) for any quarterly period j in 2019 for any given entity i, as expressed by the mathematical formulation:

$$APE_j^i = \left|\frac{A_j^i - F_j^i}{A_j^i}\right| \tag{2}$$

Despite its prevalent application, the absolute percentage error (APE) exhibits a notable limitation: the metric becomes mathematically indeterminate or infinite when actual values approximate zero – a phenomenon not uncommon in earnings prediction contexts. Furthermore, when actual numerical outcomes are exceptionally minimal, particularly below unity, the resultant percentage discrepancies can become disproportionately amplified, creating statistical anomalies. This methodological constraint is exacerbated when actual values equal zero, producing mathematically undefined or infinite APE calculations. Addressing this methodological deficiency, Kim and Kim (2016) proposed the arctangent absolute percentage error (AAPE), offering a more methodologically sound alternative for such scenarios in predictive analytics.

$$AAPE_j^i = \begin{cases} 0 & if\ A_j^i = F_j^i = 0 \\ arctan\left(\left|\frac{A_j^i - F_j^i}{A_j^i}\right|\right) & otherwise \end{cases} \tag{3}$$

This methodological refinement utilizes the arctangent function's mathematical property of transforming values spanning from negative to positive infinity into the bounded interval [–π/2, π/2]. Thus, the Mean Arctangent Absolute Percentage Error (MAAPE) for the j-th quarterly period across the entirety of I corporate entities within the analytical sample can be mathematically expressed as:

$$MAAPE_j = \frac{1}{I}\sum_{i=1}^{I} AAPE_j^i = \frac{1}{I}\sum_{i=1}^{I} arctan\left(\left|\frac{A_j^i - F_j^i}{A_j^i}\right|\right) \tag{4}$$

The preference for MAAPE (Mean Arctangent Absolute Percentage Error) over MAPE (Mean Absolute Percentage Error) is justified by the inclusion of corporate entities with actual profitability approaching zero within the examined dataset. In scenarios where even a singular observation approximates zero, while remaining observations exhibit substantially higher values, MAPE can escalate to extraordinarily elevated magnitudes, approaching mathematical infinity. This statistical phenomenon distorts the mean calculation process, effectively diminishing the statistical significance of other observational data points.

The statistical test

To assess the statistical relevance of Mean Arctan Absolute Percentage Error disparities between models, Wilcoxon's [1945] nonparametric methodology has been implemented. This analytical approach serves as a comparative examination technique for paired samples with interdependence, eschewing distributional presuppositions beyond symmetrical characteristics and independent difference scores. The application of the Wilcoxon procedure in validation contexts, particularly for establishing statistical significance in error differentials across Ensemble Prediction System models, was comprehensively examined by Ruland [1980]. For analytical purposes, distinct probability value tables are constructed quarterly (spanning the first through fourth quarters) as well as for the aggregated quarterly information.

$$H_o: AAPEs\ of\ a\ pair\ of\ models\ are\ the\ same \quad (5)$$

The null hypothesis posits equality between Arctan Absolute Percentage Errors for model pairs. Probability values below the predetermined alpha threshold of 0.05 necessitate rejection of this null hypothesis in each analytical instance. This significance criterion maintains widespread recognition and validation within statistical literature, as substantiated by Ruland [1980] among other scholarly sources.

RESULTS

Empirical Findings

The empirical analysis presented in Table 1 demonstrates that the seasonal random walk (SRW) methodology yields superior performance, as measured by the Mean Arctangent Absolute Percentage Error (MAAPE) criterion, relative to the GEMINI framework. This pattern of superiority manifests across nearly all quarterly periods and throughout the entirety of 2019, with the exception of the second quarter where Gemini marginally outperforms SRW.

Table 1. Summary statistics on forecast errors for 2019 quarters

model	Q1 MAAPE	Q2 MAAPE	Q3 MAAPE	Q4 MAAPE	Total MAAPE
SRW	0.658	0.702	0.653	0.736	0.687
GEMINI	0.746	0.700	0.666	0.787	0.725

Source: own calculations

Statistical validation through Wilcoxon testing was implemented to determine the significance of performance differentials between the aforementioned methodologies, with corresponding p-values for each temporal segment catalogued in Table 2. The analytical evidence reveals an absence of statistically meaningful disparities in forecast accuracy between SRW and GEMINI methodologies during

three quarterly periods. However, a statistically significant divergence emerges specifically in the first quarter of 2019, as well as when considering the comprehensive annual performance. Consequently, the empirical findings present an mixed pattern.

Table 2. P-values of the Wilcoxon test of forecast errors for SRW and GEMINI in 2019

	Q1	Q2	Q3	Q4	ALL
p-value	0.008	0.972	0.801	0.224	0.025

Source: own calculations

Robustness Checks

Longitudinal examination spanning the years 2017, 2018, and 2019 reveals that the seasonal random walk (SRW) methodology consistently exhibited superior predictive accuracy compared to the GEMINI approach, as documented in Table 3. Table 4 indicates that while statistically significant differences in error metrics between these methodologies were observed in 2019, comparable statistical significance was not detected in either 2017 or 2018.

Table 3. Summary statistics on forecast errors for whole years 2017–2019

model	2017 MAAPE	2018 MAAPE	2019 MAAPE
SRW	0.686	0.711	0.687
GEMINI	0.703	0.728	0.725

Source: own calculations

Table 4. P-values of paired Wilcoxon test of forecast errors for whole years 2017–2019

	2017	2018	2019
p-value	0.064	0.338	0.025

Source: own calculations

The robustness verification protocol incorporated two additional prominent performance metrics. Table 5 presents an assessment utilizing alternative error quantification methodologies: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). This comprehensive evaluation encompasses all quarterly segments of 2019. To ensure comparative equity, Consumer Price Index (CPI) adjustments were implemented to standardize the present values of future nominal errors with contemporary error measurements.

Contrary to initial expectations, as illustrated in Table 5, the GEMINI methodology generates the most favorable results when evaluated via RMSE and MAE criteria, with the seasonal random walk model achieving marginally inferior performance. The Wilcoxon test, as presented in Table 6, confirms that the error differentials between these two methodological approaches exhibit statistically significant variations.

Table 5. Summary statistics on forecast errors for RMSE and MAE in all quarters 2019

	SRW	GEMINI
RMSE	0.937	0.861
MAE	0.705	0.655

Source: own calculations

Table 6. P-values of paired Wilcoxon test of forecast errors for RMSE and MAE in 2019

	RMSE	MAE
p-value	0.011	0.011

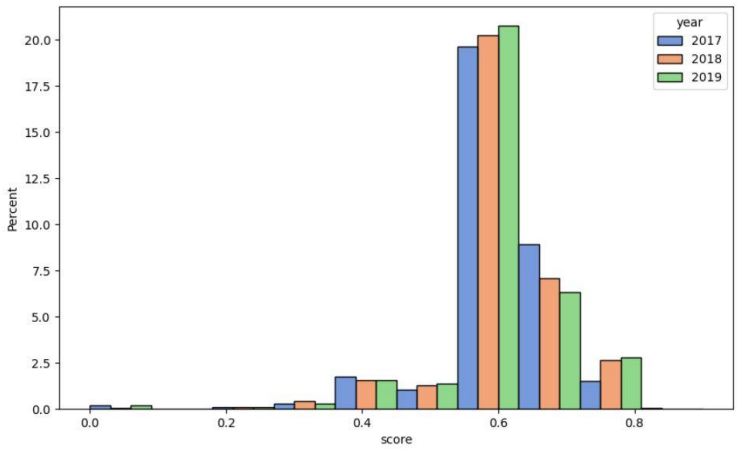
Source: own calculations

In conclusion, the empirical evidence suggests that from a MAAPE perspective, the elementary random walk model represents the optimal methodological choice for the Polish market context when compared to the artificial intelligence-based GEMINI approach. Conversely, alternative error metrics including MAE and RMSE indicate an inverse relationship regarding methodological superiority.

Certainty of forecasts

The predictive confidence metric for generative artificial intelligence systems, specifically large language models (LLMs), is quantified on a scale between zero and unity, where the extremes denote absolute uncertainty and complete conviction, respectively. This numerical assessment enables stakeholders to evaluate the dependability of model-generated content, particularly in contexts where decisions of consequence must be made.

Figure 1. Distribution of forecast certainty scores



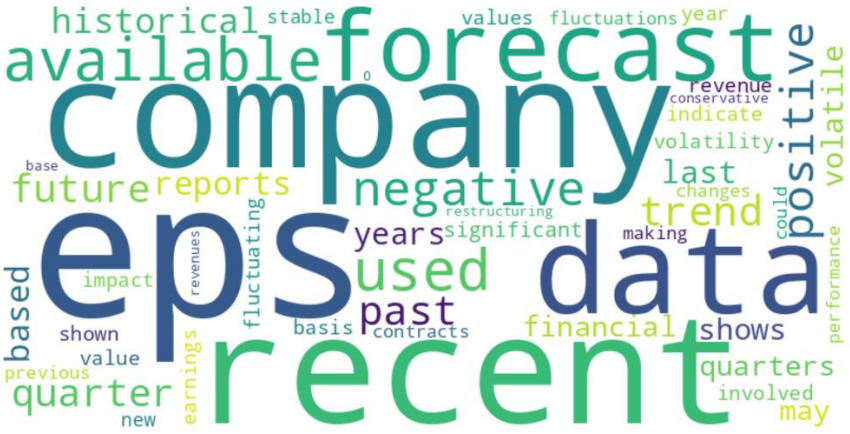
Source: own calculations

As demonstrated in Figure 1, the frequency distribution of predictive confidence metrics from the Gemini system throughout the 2017-2019 timeframe reveals that values are predominantly accumulated in the vicinity of 0.6 across all three annual periods. Substantial frequency concentration is also observed in the 0.7-0.8 range, whereas minimal distributional presence is evident in the lower confidence spectrum beneath 0.4, signifying that the model infrequently produces predictions characterized by low self-assessed reliability. The distributional characteristics suggest that the LLM consistently generates prognostications with intermediate to elevated confidence levels, with an observable progression toward more densely clustered confidence metrics over the three-year observational interval.

Chain of thought

The implementation of sequential reasoning mechanisms within sophisticated neural language frameworks facilitates the generation of intermediary analytical phases, thereby providing transparent explanatory pathways regarding predictive determinations. This methodological approach augments comprehensibility by deconstructing intricate prognostications into coherent cognitive progressions that stakeholders can scrutinize and assess. Lexical frequency visualization techniques may be employed to represent these sequential reasoning pathways across numerous forecasts by consolidating and exhibiting the most recurrent lexical units from the model's analytical stages, with dimensional prominence indicating higher occurrence rates. This methodology assists in identifying recurring patterns, conceptual frameworks, or thematic elements upon which the model relies when formulating predictions, thus providing a comprehensive overview of its analytical behavior.

Figure 2. The most frequent words in chain of thought represented by word-cloud

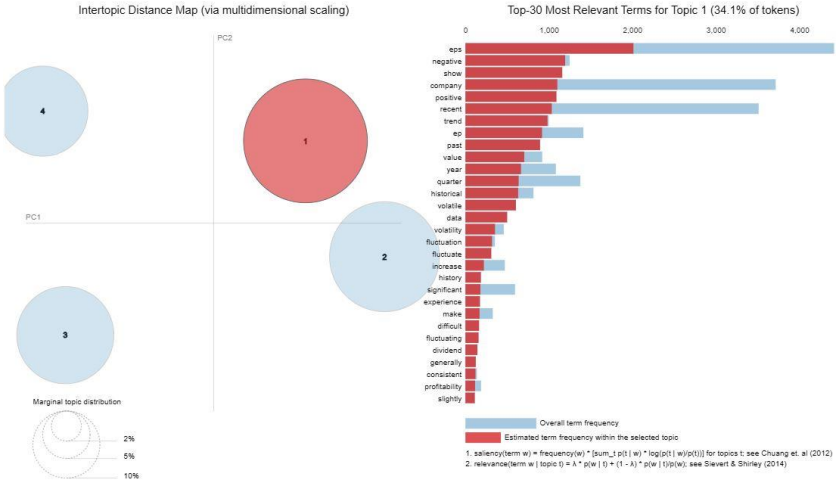


Source: own calculations

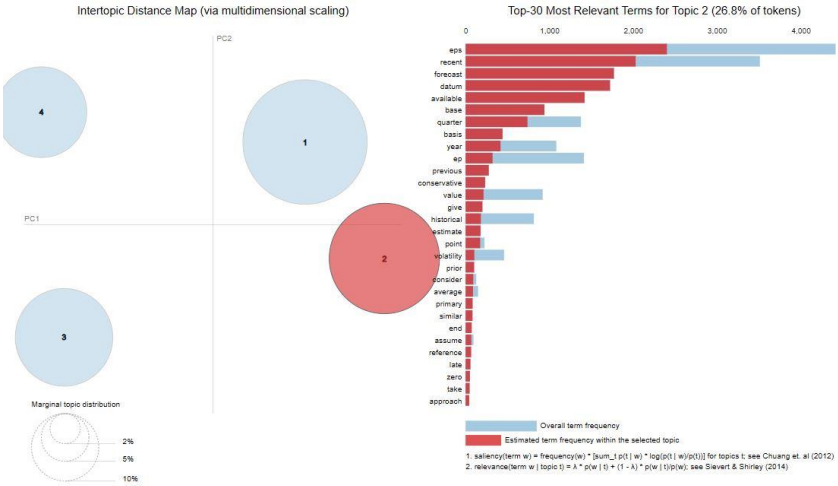
This word cloud highlights key financial forecasting terminology with “company“, “forecast“, “EPS“, “recent“, and “data“, appearing most prominently, suggesting a focus on corporate earnings predictions using recent data. The presence of terms like “volatility“, “fluctuations“, “negative“, “positive“, and “trend“ indicates analytical concerns about instability of earnings and potential swings in company performance. The inclusion of time-related words such as “historical,“ “quarter,“ “past,“ and “future“ demonstrates the temporal framework essential to financial analysis.

Figure 3. Topic model visualization in chain of thought

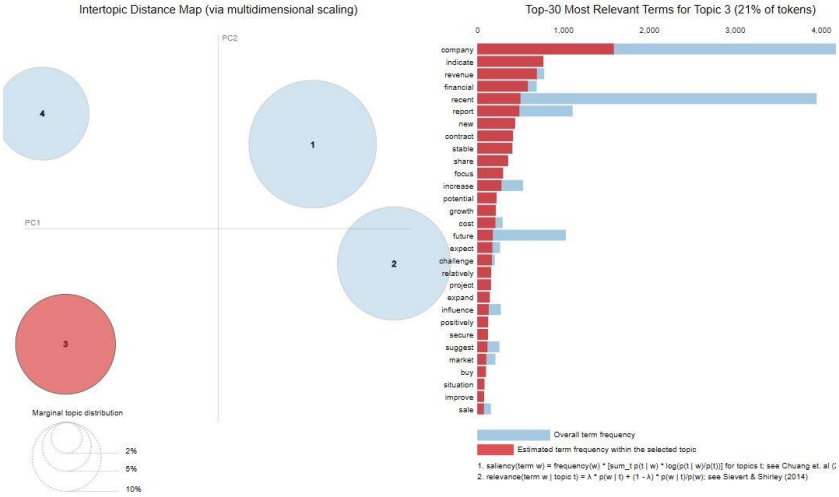
Topic 1



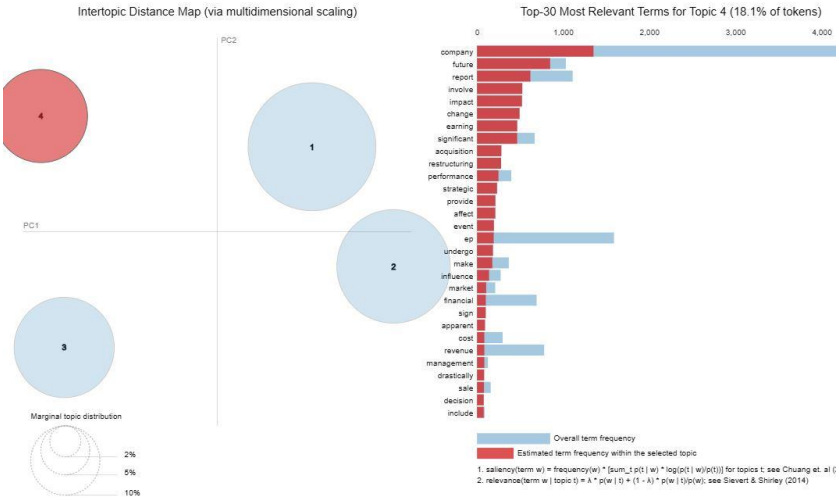
Topic 2



Topic 3



Topic 4



Source: own calculations

This graphical representation is called a Topic Model Visualization, specifically combining an Intertopic Distance Map (left side) with a Term Relevance Bar Chart (right side). The left visualization uses multidimensional scaling to display topic clusters in a two-dimensional space coming from Principal Component Analysis, showing their relative relationships, while the right side displays the most salient or relevant words for the selected topic (e.g. Topic 1) with their frequency distributions. This type of visualization is commonly used in topic modeling approaches like LDA (Latent Dirichlet Allocation) to help interpret and validate

discovered topics within a corpus. Latent Dirichlet Allocation (LDA) is a generative probabilistic model that assumes each document is a mixture of topics and each topic is a distribution over words. In topic modeling, LDA is used to uncover hidden thematic structures in large text corpora by identifying groups of words that frequently occur together, thereby revealing the underlying topics within the data.

Topic 1 appears to focus on EPS volatility and market trends. The most relevant terms include "eps", "negative", "show", "company", "positive", "recent", and "trend" suggesting this topic relates to the directional movement and fluctuations in earnings reports. This topic represents the foundational analysis of historical EPS patterns and volatility. Topic 2 centers on EPS forecast methodology and the data inputs used, containing terms like "forecast", "datum", "available", "basis", "previous", "conservative", "estimate", "reliability", and "quarter". Topic 3 relates to company fundamentals and financial outlook, mainly based on its revenues. Prominent terms include "company", "indicate", "revenue", "financial", "recent," "report," and "contract". This topic represents the contextual business factors that influence EPS expectations. Topic 4 focuses on corporate events and strategic changes. Key terms include "company," "future," "report," "involve," "impact," "change," "earnings," "acquisition," and "restructuring," suggesting this topic covers significant corporate activities that can materially affect EPS outcomes. Summarizing, the chain of thought for EPS forecasts thus progresses from: (1) analyzing historical EPS patterns and volatility, to (2) applying forecasting methodologies using available data, to (3) contextualizing within broader company financial fundamentals, to (4) adjusting for potential corporate events and strategic changes that might impact future earnings.

Discussion

This empirical investigation reveals that the seasonal random walk (SRW) model provides superior accuracy in representing Earnings per Share (EPS) patterns among Polish public companies. According to extensive research conducted by Kurylek [2023a, 2024a, 2024b, 2024c, 2025], this approach demonstrated exceptional performance compared to alternative models examined. The large language model (LLM) Gemini forecasts underperform relative to SRW when evaluated using the Mean Arctangent Absolute Percentage Error (MAAPE) metric. These conclusions remain consistent across multiple temporal ranges during robustness verification.

Interestingly, when alternative evaluation metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are employed, the Gemini model exhibits statistically significant superior performance compared to SRW. This discrepancy across different error metrics can be attributed to their distinct mathematical characteristics. The MAAPE metric likely imposes less severe penalties on outliers compared to RMSE and MAE, suggesting that SRW performs better in relative/percentage terms - particularly for companies with small EPS values where percentage errors can be very large - while Gemini produces lower absolute errors overall (MAE, RMSE), indicating generally accurate predictions

with occasional significant outliers that receive substantial penalties under the arctangent metric. This statistical divergence emphasizes how model selection should be determined by specific error tolerance requirements of financial applications - SRW being preferable when avoiding extreme outliers is crucial, while Gemini offers advantages when average prediction accuracy across the entire dataset is prioritized.

The examined period (2010-2019) represents a relatively stable market environment characterized by minimal disruptions and a horizontal WIG index trend. Under these conditions, SRW's fundamental assumption - that the upcoming quarter will closely mirror the corresponding quarter from the previous year - becomes increasingly reliable. Consequently, sophisticated models like GEMINI may offer minimal advantages in identifying stable patterns. Furthermore, the structural characteristics of emerging markets such as Poland may inherently favor less complex forecasting approaches. These outcomes suggest that simpler models may capture the maximum achievable predictive accuracy from historical data and effectively reflect the relatively straightforward EPS dynamics of Polish listed companies. The superior performance of less sophisticated models in the Polish context may be attributed to the simplicity of the underlying stochastic processes, consistent with the relatively unsophisticated nature of Polish public enterprises.

The observed inferior performance of GEMINI aligns with Cao and Wang [2024], who determined that LLMs' predictive accuracy significantly diminishes when confronted with diverse time series data and traditional signals containing both periodic and trend components, which EPS time series certainly exhibit. Additionally, these findings partially corroborate Abdelsamie and Wang's [2024] observation that LLMs continue to demonstrate limitations in nuanced contextual understanding and adaptability, highlighting the persistent value of human expertise.

Gemini LLM's forecast certainty score provides valuable insight into its confidence regarding EPS predictions, with scores predominantly clustering around 0.6-0.7, indicating moderate to high confidence despite variable accuracy. This metric functions as a practical instrument for investment decision-making, potentially enabling users to prioritize predictions with higher certainty scores while treating lower-confidence forecasts with appropriate caution. The chain-of-thought reasoning demonstrates that Gemini's forecasting methodology follows a logical progression from analyzing historical EPS patterns and volatility to implementing forecasting methodologies, contextualizing within broader company fundamentals, and ultimately adjusting for potential corporate events. Word cloud analysis of these reasoning chains emphasizes key financial terminology centered around volatility and trend indicators, demonstrating the model's focus on temporal frameworks essential to financial analysis. Topic modeling further illuminates four distinct dimensions of Gemini's analytical approach: historical EPS volatility assessment, forecast methodology application, company fundamental analysis, and consideration of strategic corporate events potentially impacting future earnings. This multifaceted reasoning structure indicates that despite accuracy limitations, Gemini employs a

sophisticated analytical framework mirroring the comprehensive evaluation process utilized by human financial analysts.

Given that EPS behavior follows a seasonal random walk pattern over shorter time series, and considering that stock prices derive from EPS multiplied by the price-to-earnings (P/E) multiple, one may infer that stock prices exhibit at minimum equivalent randomness to EPS. Accurately forecasting stock prices even one quarter in advance becomes exceptionally challenging when EPS behavior demonstrates random walk characteristics. During shorter timeframes where EPS remains constant, stock price forecasting essentially becomes P/E multiple prediction. Consequently, forecasting P/E multiples for periods shorter than a quarter - intervals between quarterly financial reports - could be particularly relevant for investment decisions. The seasonal random walk forecast effectively replicates values from identical quarters in previous years, suggesting that for extended forecast horizons, the P/E multiple may exert greater influence than forthcoming company earnings (EPS) when predicting future prices. This corresponds with economic theory positing that P/E multiples are influenced by anticipated future earnings growth, prospective interest rates, and market sentiment or risk premium, whereas EPS forecasts relate exclusively to imminent earnings. In both short-term and long-term investment contexts, consensus indicates that P/E multiples possess greater predictive significance than EPS forecasts.

CONCLUSIONS

The prediction of Earnings per Share (EPS) holds particular relevance within emerging financial markets such as Poland, where analyst coverage of publicly traded entities remains sparse. This investigation examines the performance capabilities of Large Language Models (LLM), specifically the Gemini architecture, in prognosticating EPS temporal sequences. An examination of quarterly EPS information from 267 Polish corporate entities spanning 2010-2019 demonstrates that the Seasonal Random Walk (SRW) methodology yielded superior results with minimal error rates as quantified by the Mean Arctangent Absolute Percentage Error. Such findings potentially reflect the relatively rudimentary organizational structures prevalent among Polish listed corporations. Notably, alternative error assessment frameworks including Mean Absolute Error and Root Mean Square Error reveal contradictory patterns, potentially attributable to these metrics' heightened sensitivity to statistical outliers.

Confidence indicators generated by the Gemini model predominantly exhibit moderate to substantial certainty values (0.6-0.7), potentially offering investment decision guidance despite inconsistent accuracy performance. A chain-of-thought examination reveals sophisticated reasoning processes that emulate human financial analytical approaches – proceeding from historical pattern recognition through methodological implementation, fundamental analysis, and strategic event

consideration – indicating that LLMs provide analytical sophistication even when failing to surpass simpler statistical frameworks in accuracy measurements.

The functional implications of these findings suggest that within abbreviated temporal sequences, employing methodologies of greater complexity than SRW for EPS forecasting within Poland may be unwarranted. However, dependency on SRW for EPS modeling suggests that projected equity valuations may exhibit considerable stochasticity, complicating accurate predictions. Consequently, forecasting price-to-earnings multiples may prove more relevant than EPS predictions for future equity valuation estimations, particularly within abbreviated investment timeframes where EPS demonstrates relative stability.

Subsequent scholarly inquiry might explore correlations between forecasting accuracy and organizational magnitude, with sectoral analysis potentially informing optimal EPS prediction model selection. Investigation of time series transformations to normalize EPS distributions could yield significant insights. Moreover, assessing the performance of diverse prediction methods and analyst projections during recessionary periods, such as the 2008-2009 financial crisis or COVID-19 pandemic, may yield meaningful findings in further research. Detecting cyclical patterns with the SRW approach may provide insights for investment strategies, potentially questioning the assumptions of the 'weak form' of the Efficient Market Hypothesis (EMH). Future studies could evaluate whether such strategies are capable of outperforming the market.

ABBREVIATIONS

In the text the following abbreviations are used:
SRW – Seasonal Random Walk model
GEMINI – the Gemini model by Google DeepMind
MAAPE – Mean Arctangent Absolute Percentage Error
MAE – Mean Absolute Error
RMSE – Root Mean Square Error

STATEMENTS AND DECLARATIONS

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.
Funding: The authors have received no funding from any source in the preparation of this work.
Data availability statement: Data will be available on request.

REFERENCES

Abarbanell J., Bushee B. (1997) Fundamental Analysis, Future EPS, and Stock Prices. Journal of Accounting Research, 35(1), 1-24. https://doi.org/10.2307/2491464

Abdelsamie M., Wang H. (2024) Comparative Analysis of LLM-based Market Prediction and Human Expertise with Sentiment Analysis and Machine Learning Integration. 2024 7th International Conference on Data Science and Information Technology (DSIT), 1-6. https://doi.org/10.1109/dsit61374.2024.10881868

Abe Y., Matsuo S., Kondo R., Hisano R. (2024) Leveraging Large Language Models for Institutional Portfolio Management: Persona-Based Ensembles. 2024 IEEE International Conference on Big Data (BigData), 4799-4808. https://doi.org/10.1109/bigdata62323.2024.10825362

Ahmadpour A., Etemadi H., Moshashaei S. (2015) Earnings Per Share Forecast using Extracted Rules from Trained Neural Network by Genetic Algorithm. Computational Economics, 46(1), 55-63. https://doi.org/10.1007/s10614-014-9455-6

Atiya A., Shaheen S., Talaat N. (1997) An Efficient Stock Market Forecasting Model using Neural Networks. IEEE International Conference on Neural Networks – Conference Proceedings. http://dx.doi.org/10.1109/ICNN.1997.614231

Ball R., Ghysels E. (2017) Automated Earnings Forecasts: Beat Analysts or Combine and Conquer? Management Science, 64(10), 4936–4952. https://doi.org/10.1287/mnsc.2017.2864

Ball R., Watts R. (1972) Some Time Series Properties of Accounting Income. The Journal of Finance, 27(3), 663-681. http://dx.doi.org/10.1111/j.1540-6261.1972.tb00991.x

Bathke Jr. A. W., Lorek K. S. (1984) The Relationship between Time-Series Models and the Security Market's Expectation of Quarterly Earnings. The Accounting Review, 59(2), 163-176.

Bradshaw M., Drake M., Myers J., Myers L. (2012) A Re-Examination of Analysts' Superiority over Time-Series Forecasts of Annual Earnings. Review of Accounting Studies, 17(4), 944-968. http://dx.doi.org/10.1007/s11142-012-9185-8

Brandon Ch., Jarrett J. E., Khumawala S. B. (1987) A Comparative Study of the Forecasting Accuracy of Holt-Winters and Economic Indicator Models of Earnings Per Share for Financial Decision Making. Managerial Finance, 13(2), 10-15. http://dx.doi.org/10.1108/eb013581

Brooks L. D., Buckmaster D. A. (1976) Further Evidence of the Time Series Properties of Accounting Income. The Journal of Finance, 31(5), 1359-1373. http://dx.doi.org/10.1111/j.1540-6261.1976.tb03218.x

Brown L. D., Griffin P. A., Hagerman R. L., Zmijewski M. E. (1987) Security Analyst Superiority Relative to Univariate Time-Series Models in Forecasting Quarterly Earnings. Journal of Accounting and Economics, 9(1), 61-87. http://dx.doi.org/10.1016/0165-4101(87)90017-6

Brown L. D., Rozeff M. S. (1979) Univariate Time-Series Models of Quarterly Accounting Earnings Per Share: A Proposed Model. Journal of Accounting Research, 17(1), 179-189. http://dx.doi.org/10.2307/2490312

Cao Q., Gan Q. (2009) Forecasting EPS of Chinese Listed Companies using a Neural Network with Genetic Algorithm. 15th Americas Conference on Information Systems 2009, AMCIS 2009, 2791-2981.

Cao Q., Parry M. (2009) Neural Network Earnings Per Share Forecasting Models: A Comparison of Backward Propagation and the Genetic Algorithm. Decision Support Systems, 47(1), 32-41. https://doi.org/10.1016/j.dss.2008.12.011

Cao Q., Schniederjans M. J., Zhang W. (2004) Neural Network Earnings Per Share Forecasting Models: A Comparative Analysis of Alternative Methods. Decision Sciences, 35(2), 205-237. https://doi.org/10.1111/j.00117315.2004.02674.x

Cao R., Wang Q. (2024) An Evaluation of Standard Statistical Models and LLMs on Time Series Forecasting. 2024 IEEE International Conference on Future Machine Learning and Data Science (FMLDS), 533-538. https://doi.org/10.1109/fmlds63805.2024.00098

Cao Y., Chen Z., Pei Q., Lee N., Subbalakshmi K. P., Ndiaye P. M. (2024) ECC Analyzer: Extracting Trading Signal from Earnings Conference Calls using Large Language Model for Stock Volatility Prediction. Proceedings of the 5th ACM International Conference on AI in Finance, 257-265. https://doi.org/10.1145/3677052.3698689

Chen Y., Chen S., Huang H., Sangaiah A. (2020) Applied Identification of Industry Data Science using an Advanced Multi-Componential Discretization Model. Symmetry, 12(10), 1-28. https://doi.org/10.3390/sym12101620

Chong C. Y., Kao H.-Y. (2025) Rationale-Driven Predictions for Stock Movements: A Multi-model Integration and Stack Generalization Approach. Technologies and Applications of Artificial Intelligence, 124-138. https://doi.org/10.1007/978-981-96-4589-3_9

Conroy R., Harris R. (1987) Consensus Forecasts of Corporate Earnings: Analysts' Forecasts and Time Series Methods. Management Science, 33(6), 725-738. http://dx.doi.org/10.1287/mnsc.33.6.725

Elend L., Kramer O., Lopatta K., Tideman S. (2020) Earnings Prediction with Deep Learning. German Conference on Artificial Intelligence (Künstliche Intelligenz) KI 2020: Advances in Artificial Intelligence, 267-274. http://dx.doi.org/10.1007/978-3-030-58285-2_22

Elton E. J., Gruber M. J. (1972) Earnings Estimates and the Accuracy of Expectational Data. Management Science, 18(8), B409-B424. http://dx.doi.org/10.1287/mnsc.18.8.B409

Foster G. (1977) Quarterly Accounting Data: Time-Series Properties and Predictive-Ability Results. The Accounting Review, 52(1), 1-21.

Gaio L., Gatsios R., Lima F., Piamenta Jr. T. (2021) Re-Examining Analyst Superiority in Forecasting Results of Publicly-Traded Brazilian Companies. Revista de Administracao Mackenzie, 22(1), eRAMF210164. https://doi.org/10.1590/1678-6971/eramf210164

Gomez A. N., Kaiser L., Jones L., Parmar N., Polosukhin I., Vaswani A., Shazeer N., Uszkoreit J. (2017) Attention Is All You Need. arXiv preprint arXiv:1706.03762.

Google DeepMind. (2023) Gemini: A Family of Highly Capable Multimodal Models. arXiv. https://arxiv.org/abs/2312.1180

Griffin P. (1977) The Time-Series Behavior of Quarterly Earnings: Preliminary Evidence. Journal of Accounting Research, 15(1), 71-83. http://dx.doi.org/10.2307/2490556

Gupta R., Khirbat G., Singh S. (2013) Optimal Neural Network Architecture for Stock Market Forecasting. Proceedings – 2013 International Conference on Communication Systems and Network Technologies, CSNT 2013, 557-561. https://doi.org/10.1109/csnt.2013.120

Harris R. D. F., Wang P. (2019) Model-Based Earnings Forecasts vs. Financial Analysts' Earnings Forecasts. British Accounting Review, 51(4), 424-437. https://doi.org/10.1016/j.bar.2018.10.002

Hou K., van Dijk M., Zhang Y. (2012) The Implied Cost of Capital: A New Approach. Journal of Accounting and Economics, 53(3), 504-526. https://doi.org/10.1016/j.jacceco.2011.12.001

Hou G., Lian Q. (2024) Benchmarking of Commercial Large Language Models: ChatGPT, Mistral, and Llama. ResearchSquare. https://doi.org/10.21203/rs.3.rs-4376810/v1

Jarrett J. E. (2008) Evaluating Methods for Forecasting Earnings Per Share. Managerial Finance, 16, 30-35. http://dx.doi.org/10.1108/eb013647

Johnson T. E., Schmitt T. G. (1974) Effectiveness of Earnings Per Share Forecasts. Financial Management, 3(2), 64-72. http://dx.doi.org/10.2307/3665292

Kim S., Kim H. (2016) A New Metric of Absolute Percentage Error for Intermittent Demand Forecasts. International Journal of Forecasting, 32(3), 669-679. http://dx.doi.org/10.1016/j.ijforecast.2015.12.003

Kong Y., Nie Y., Dong X., Mulvey J. M., Poor H. V., Wen Q., Zohren S. (2024) Large Language Models for Financial and Investment Management: Models, Opportunities, and Challenges. The Journal of Portfolio Management, 51(2), 211-231. https://doi.org/10.3905/jpm.2024.1.646

Kuryłek W. (2023a) The Modeling of Earnings Per Share of Polish Companies for the Post-Financial Crisis Period using Random Walk and ARIMA Models. Journal of Banking and Financial Economics, 1(19), 26-43. http://dx.doi.org/10.7172/2353-6845.jbfe.2023.1.2

Kuryłek W. (2023b) Can Exponential Smoothing Do Better than Seasonal Random Walk for Earnings Per Share Forecasting in Poland? Bank Credit, 54(6), 651-672.

Kuryłek W. (2024a) Can We Profit from BigTechs' Time Series Models in Predicting Earnings Per Share? Evidence from Poland. Data Science in Finance and Economics, 4(2), 218-235. http://dx.doi.org/10.3934/DSFE.2024008

Kuryłek W. (2024b) Artificial Neural Networks and Gradient-Boosting Decision Trees in Time Series Forecasting of Earnings Per Share in Poland. Eastern European Economics, 1-22. https://doi.org/10.1080/00128775.2024.2429137

Kuryłek W. (2024c) If Multilayer Perceptron Network May Help in Multivariate EPS Forecasting. Evidence from Poland. Quantitative Methods in Economics, 25(3), 107-123. https://doi.org/10.22630/mibe.2024.25.3.10

Kuryłek W. (2025) Are Natural Language Processing Methods Applicable to EPS Forecasting in Poland? Data Science in Finance and Economics, 5(1), 35-52. https://doi.org/10.3934/dsfe.2025003

Lacina M., Lee B., Xu R. (2011) An Evaluation of Financial Analysts and Naïve Methods in Forecasting Long-Term Earnings. [in:] K. D. Lawrence R. K. Klimberg (Eds.), Advances in Business and Management Forecasting (pp. 77-101) Bingley, UK: Emerald. http://dx.doi.org/10.1108/S1477-4070(2011)0000008009

Lai S., Li H. (2006) The Predictive Power of Quarterly Earnings Per Share based on Time Series and Artificial Intelligence model. Applied Financial Economics, 16(18), 1375-1388. http://dx.doi.org/10.1080/09603100600592752

Lev B., Thiagarajan S. (1993) Fundamental Information Analysis. Journal of Accounting Research, 31(2), 190–215. http://doi.org/10.2307/2491270

Lev B., Li S., Sougiannis T. (2010) The Usefulness of Accounting Estimates for Predicting Cash Flows and Earnings. Review of Accounting Studies, 15(4), 779-807. https://doi.org/10.1007/s11142-009-9107-6

Li K. K. (2011) How Well Do Investors Understand Loss Persistence? Review of Accounting Studies, 16(3), 630-667. https://doi.org/10.1007/s11142-011-9157-4

Li K. K., Mohanram P. (2014) Evaluating Cross-Sectional Forecasting Models for the Implied Cost of Capital. Review of Accounting Studies, 19(3), 1152-1185. https://doi.org/10.1007/s11142-014-9282-y

Lorek K. S. (1979) Predicting Annual Net Earnings with Quarterly Earnings Time-Series Models. Journal of Accounting Research, 17(1), 190-204. http://dx.doi.org/10.2307/2490313

Lorek K. S., Willinger G. L. (1996) A Multivariate Time-Series Model for Cash-Flow Data. Accounting Review, 71, 81-101.

Ohlson J. A. (1995) Earnings, Book Values, and Dividends in Equity Valuation. Contemporary Accounting Research, 11(2), 661-687. https://doi.org/10.1092/7tpj-rxqn-tqc7-ffae

Ohlson J. A. (2001) Earnings, Book Values, and Dividends in Equity Valuation: An Empirical Perspective. Contemporary Accounting Research, 18(1), 107-120. https://doi.org/10.1092/7tpj-rxqn-tqc7-ffae

Pagach D. P., Warr R. S. (2020) Analysts Versus Time-Series Forecasts of Quarterly Earnings: A Maintained Hypothesis Revisited. Advances in Accounting, 51, 1-15. http://dx.doi.org/10.1016/j.adiac.2020.100497

Pope P. F., Wang P. (2005) Earnings Components, Accounting Bias and Equity Valuation. Review of Accounting Studies, 10(4), 387-407. https://doi.org/10.1007/s11142-005-4207-4

Pope P., Wang P. (2014) On the Relevance of Earnings Components: Valuation and Forecasting Links. Review of Quantitative Finance and Accounting, 42, 399-413. https://doi.org/10.1007/s11156-013-0347-y

Ruland W. (1980) On the Choice of Simple Extrapolative Model Forecasts of Annual Earnings. Financial Management, 9(2), 30-37. http://dx.doi.org/10.2307/3665165

Sarker I. H. (2024) LLM Potentiality and Awareness: A Position Paper from the Perspective of Trustworthy and Responsible AI Modeling. https://doi.org/10.36227/techrxiv.170905626.67078570/v1

Watts R. L. (1975) The Time Series Behavior of Quarterly Earnings. Working paper, Department of Commerce, University of New Castle.

Wilcoxon F. (1945) Individual Comparisons by Ranking Methods. Biometrics, 1, 80-83. http://dx.doi.org/10.2307/3001968

Xiaoqiang W. (2022) Research on Enterprise Financial Performance Evaluation Method based on Data Mining. 2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI) https://doi.org/10.1109/icetci55101.2022.9832404

QUANTITATIVE METHODS IN ECONOMICS
Volume XXVII, No. 1, 2026, pp. 27 – 40

Received: 10.01.2026
Accepted: 17.02.2026

PROXIMAL POLICY OPTIMISATION VERSUS ANT COLONY OPTIMISATION FOR THE THREE-DIMENSIONAL BIN PACKING PROBLEM: A COMPARATIVE STUDY

Michał Sawicki
Łazarski University, Warsaw, Poland
Tomasz Woźniakowski https://orcid.org/0000-0002-0779-4769
Department of Econometrics and Statistics
Warsaw University of Life Sciences – SGGW, Poland
e-mail: tomasz_wozniakowski@sggw.edu.pl

Abstract: This paper compares a Proximal Policy Optimisation (PPO) deep reinforcement-learning agent with an Ant Colony Optimisation (ACO) solver on the offline, heterogeneous-bin three-dimensional bin packing problem (3D-BPP). Both algorithms were evaluated on fifty synthetic instances using a unified composite scoring function covering placement ratio, volume utilisation, bin-count penalty and mean per-bin waste. PPO achieves a higher mean composite score (0.346 vs. 0.283), wins on 38 of 50 instances with an average winning margin of 0.101, and resolves each instance in under 60 seconds on a commodity CPU. ACO exhibits greater score variance and resolves instances in up to 1,706 seconds, but its training-free character makes it relevant when the instance distribution changes too rapidly for policy retraining. The PPO training cost of approximately 5.5 hours is recovered after 58 instances compared with ACO at mean inference times. A paired Wilcoxon signed-rank test is identified as the appropriate significance test once per-instance data are made available.

Keywords: 3D bin packing, Proximal Policy Optimisation, Ant Colony Optimisation, deep reinforcement learning, swarm intelligence, logistics optimisation

JEL classification: C61, C63, M11

https://doi.org/10.22630/MIBE.2026.27.1.2

INTRODUCTION

The three-dimensional bin packing problem (3D-BPP) requires placing a set of rectangular boxes into a finite collection of bins such that boxes do not overlap, do not exceed bin boundaries, and the number of bins used is minimised. The problem arises directly in vehicle loading, pallet building and warehouse slot allocation. In high-volume logistics settings, a one-percent improvement in average bin fill translates into fewer truck movements and lower fuel cost per unit shipped [Wäscher et al. 2007; Gzara et al. 2020].

Because 3D-BPP is strongly NP-hard [Garey, Johnson 1979; Christensen et al. 2016], optimal solvers are tractable only for small instances. Two algorithmic families have attracted sustained research attention. Ant Colony Optimisation (ACO) solves each instance from scratch: simulated ants build candidate packings guided by pheromone trails that reinforce good solutions and decay over time [Dorigo, Gambardella 1997; Dorigo, Stützle 2018]. Proximal Policy Optimisation (PPO), a deep reinforcement-learning (DRL) method, invests computation in an off-line training phase and then resolves new instances through fast policy inference [Schulman et al. 2017].

The two approaches imply different cost structures. ACO requires no training but spends computation at inference time on every instance. PPO requires substantial training but resolves instances in seconds once trained. Whether the PPO training overhead is economically justified depends on how many instances an organisation solves and at what rate the instance distribution changes. This trade-off is well-known in principle but rarely quantified empirically under identical experimental conditions for 3D-BPP specifically.

This paper places both algorithms on the same fifty synthetic offline 3D-BPP instances, evaluates them with the same composite scoring function, and reports (i) packing quality statistics, (ii) instance-level win/loss counts, (iii) runtime distributions, and (iv) a break-even analysis quantifying how many instances PPO must solve to recover its training cost relative to ACO. The paper is structured as follows: Section 2 reviews the relevant literature; Section 3 describes algorithm implementations and the benchmark protocol; Section 4 reports results; Section 5 discusses findings and limitations; Section 6 concludes.

LITERATURE REVIEW

Problem Definition and Typology

Wäscher, Haußner and Schumann [2007] provide the canonical taxonomy of cutting-and-packing problems, classifying them by objective function, item–container relationship and temporal availability of item data. Under that taxonomy, 3D-BPP is an offline V-type problem: all box dimensions w_i, h_i, d_i are known at the start, the algorithm may open as many bins of given dimensions as needed, and the

goal is to minimise the bin count. Earlier typologies by Dyckhoff [1990] and Dyckhoff and Finke [1992] established the terminological foundation. NP-hardness of 3D-BPP follows from the one-dimensional case [Garey, Johnson 1979]; Christensen et al. [2016] survey approximation results and complexity boundaries. In practice, instances with more than a few dozen items require heuristic or learning-based solvers.

Swarm Intelligence Approaches

ACO was introduced by Dorigo and Gambardella [1997] for the Travelling Salesman Problem and adapted to bin packing by Levine and Ducatelle [2004] for the one-dimensional case. For 3D-BPP the dominant hybrid couples ACO with the three-dimensional heuristic algorithm (3DHA) of Silveira et al. [2013]: ants determine the box sequence and 3DHA resolves placement via an extreme-point first-fit rule. Viegas et al. [2014, 2015] applied this hybrid to the steel-cutting variant of 3D-BPP, finding ACO competitive with tabu search and simulated annealing within practical time budgets. Li and Zhang [2015] extended the approach to heterogeneous containers. Maia and Borenstein [2024] combined ACO with column generation, obtaining improved lower bounds on large instances. Dorigo and Stützle [2018] survey pheromone update variants including the MAX–MIN Ant System (MMAS) used in the present study.

Deep Reinforcement Learning Approaches

Hu et al. [2017] first applied DRL to a 3D packing variant, training a policy to sequence boxes so as to minimise bin surface area. Schulman et al. [2017] introduced PPO as a policy-gradient method that clips the surrogate objective to keep updates within a trust region, achieving stable convergence on high-dimensional action spaces. For 3D-BPP, the standard design factorises the action space: a deterministic geometric routine identifies feasible placement anchors, and the neural network selects among (item, orientation) pairs. This keeps the learned component compact while offloading geometric feasibility checking to deterministic code [Murdivien, Um 2023]. Que et al. [2023] coupled a transformer encoder with PPO (TransPack), achieving state-of-the-art volume utilisation on offline instances by using multi-head attention to represent item sets of variable size. Xiong et al. [2024] extended this to the online setting (GOPT), where items arrive sequentially; their generalisation across bin sizes stems from the size-agnostic nature of the transformer. Zhao et al. [2022] trained an actor–critic network that predicts placement coordinates successively (length, width, rotation), using the critic to veto infeasible moves.

Gap in the Literature

Direct, standardised comparisons between ACO and PPO on identical 3D-BPP instances are absent from the literature. Viegas et al. [2014] benchmark ACO against tabu search but not against DRL methods. Que et al. [2023] compare

transformer-PPO against earlier DRL approaches but not against metaheuristics. The present paper fills this gap with a controlled head-to-head experiment that also quantifies the economic break-even point of PPO training.

METHODOLOGY

Problem Formulation

The benchmark uses an offline, heterogeneous-bin 3D-BPP. Each instance specifies a set of boxes with dimensions (w_i, h_i, d_i) and a set of bins with dimensions (W_k, H_k, D_k), all provided simultaneously. Boxes must be placed with edges parallel to bin faces, without mutual overlap and without exceeding bin boundaries. All six axis-aligned rotations are permitted. The ACO solver additionally prefers lower placement positions to simulate gravity. The objective is the composite score:

$$score = 0.65 \cdot p + 0.25 \cdot u - 0.30 \cdot b - 0.25 \cdot \mu w \quad (1)$$

Placement ratio p = (boxes placed)/(boxes total); volume utilisation u = (Σ box volumes placed)/(Σ bin volumes); bin penalty b = (bins used)/(bins available); mean per-bin waste $\mu w = (1/m) \cdot \Sigma_i(1-u_i)$ where m is the number of bins used. The coefficient on p (0.65) is the largest because complete placement is the primary industrial requirement; b (0.30) and μw (0.25) penalise inefficient bin usage; u (0.25) rewards density. An earlier formulation with equal weights on p and u (0.40 each) produced an agent that opened extra bins to maintain high utilisation per bin; the revised weights eliminate this incentive.

PPO Implementation

At each decision step a deterministic corner-search routine identifies the first geometrically feasible anchor point in the current bin, reducing the learning problem to selecting an (item, orientation) pair. This factorisation keeps the action space compact and avoids learning geometric feasibility, which would require far more training data [Que et al. 2023; Murdivien, Um 2023].

The actor–critic network consists of six fully connected layers with ReLU activations (layers 1–4: 256 units; layers 5–6: 128 units). Actor and critic share all six layers and diverge only at the output projection. A binary placement mask zeroes out already-packed items, so the softmax over action logits considers only unpacked boxes and their six orientations. Weight sharing halves parameter count and couples the value estimate to the same feature space used for action selection.

Training uses 2,000 episodes drawn from 700 knapsack instances [Beasley 1990–2018, MIT licence] augmented with randomly generated cases in a 1:2 ratio. PPO clipping threshold: $\varepsilon = 0.10$; discount factor: $\gamma = 0.99$; learning rate halved at episode 1,000. The step reward is the increment in composite score; a terminal bonus of +0.1 is added when all boxes are placed. Total training time: approximately 5.5 hours on the experimental CPU.

The Markov Decision Process: state s_t = [placement mask, box dimensions, bin dimensions, current score]; action a_t = index of a feasible (item, orientation) pair; transition: deterministic geometry check then state update; reward $r_t = score_t - score_{t-1}$ + completion bonus.

ACO Implementation

The solver implements MMAS [Dorigo, Stützle 2018]. Pheromone levels are stored in a hash-map keyed by (box index, extreme point) pairs, keeping memory proportional to the number of extreme points rather than to all possible placements [Silveira et al. 2013]. Each iteration: (1) evaporate all trails by factor 0.8; (2) each ant samples a box permutation proportional to pheromone$^{\alpha}$ · heuristic$^{\beta}$; (3) 3DHA decodes the permutation into placements via extreme-point first-fit; (4) the top five ants deposit pheromone proportional to their normalised score and rank. If the best score does not improve for 10% of iterations, all trails are halved (soft reset).

Colony size adapts to instance complexity via:

$$ants = clamp(round(b \cdot log_2(n+1) \cdot s), 50, 8000) \quad (2)$$

$$iterations = clamp(round(n \cdot log_2(b+1) \cdot 4s), 150, 5000) \quad (3)$$

where n = box count, b = bin count, s = 2. This contrasts with prior work that fixes colony size regardless of instance size [Viegas et al. 2014; Maia, Borenstein 2024]. The extreme-point list is capped at 1,000 entries per iteration to bound runtime on large instances. The solver is written in Rust and called from Python via PyO3 bindings, which eliminates CPython interpreter overhead in the inner loop.

Note on statistical validation of the dynamic colony formula: a controlled ablation study comparing the dynamic formula against a fixed baseline (e.g. 100 ants, 300 iterations) across instances of varying size would quantify the runtime savings and any score trade-off. Such an experiment is a necessary next step before the formula can be recommended as a general design principle; it is identified here as immediate future work.

Benchmark Design

Fifty independent instances were generated synthetically with box and bin dimensions drawn from a uniform distribution subject to the constraint that each box fits inside at least one bin. For each instance the benchmark driver ran PPO inference first (weights loaded from a .pt checkpoint), then the ACO solver on the same instance. Wall-clock CPU time was recorded separately for each algorithm. All results were written to a single CSV file. Hardware: 6-core/12-thread CPU at 3.6 GHz base clock, GPU disabled, so both algorithms face identical latency constraints. Software: Python 3.11, PyTorch 2.7.0 [Ansel et al. 2024], gym-BinPack3D environment [2024].

Statistical note: the paired Wilcoxon signed-rank test [Wilcoxon 1945] is the appropriate non-parametric test for comparing the two algorithms, because each of the 50 instances was solved by both methods under identical conditions, producing

matched pairs of scores. The test requires per-instance score pairs rather than aggregate statistics. Those data are stored in the benchmark CSV and should be analysed in the final version of the paper; the present analysis is therefore confined to descriptive statistics.

RESULTS

Overall Composite Score

Table 1 reports descriptive statistics for the composite scores. PPO's mean of 0.346 exceeds ACO's mean of 0.283 by 0.063 points (22% relative). PPO's standard deviation (0.042) is half of ACO's (0.085), indicating that PPO's performance varies less across instances. ACO's maximum (0.458) marginally exceeds PPO's (0.417), showing that the colony can occasionally find tighter packings, but ACO's minimum (0.141) is 0.064 below PPO's floor (0.205). Whether the 0.063-point gap in means is statistically significant requires a paired Wilcoxon test on per-instance scores (see Section 3.4); the descriptive evidence is consistent with a systematic advantage for PPO.

Table 1. Composite score statistics across 50 benchmark instances

Metric	Mean	Median	Std. Dev.	Min	Max
PPO score	0.346	0.350	0.042	0.205	0.417
ACO score	0.283	0.275	0.085	0.141	0.458
Δ (PPO−ACO)	0.063	0.089	0.081	−0.041	0.064

Source: own calculations

Head-to-Head Comparison

Table 2 records win/loss outcomes per instance. PPO wins on 38 of 50 instances (76%) with a mean winning margin of 0.101. ACO wins on 12 instances (24%) with a mean margin of 0.056. There are no ties. PPO's winning margin (0.101) is 1.8× larger than ACO's (0.056), meaning PPO not only wins more often but wins by larger amounts. The 12 ACO victories cluster on instances where box and bin geometry aligns with 3DHA's first-fit scanning direction (see Section 5.2).

Table 2. Win/loss record across 50 instances

Method	Wins	Ties	Win rate	Mean margin when winning
PPO	38	0	76%	0.101
ACO	12	0	24%	0.056

Source: own calculations

Placement Ratio

Both algorithms achieve a placement ratio of 1.0 on all 50 instances: every box is packed in every run. Score differences between PPO and ACO therefore reflect bin-count efficiency and void management exclusively, not item completeness.

Bin Penalty

Table 3 shows bin penalty statistics. PPO's mean bin penalty of 0.516 is 0.298 points below ACO's mean of 0.814, indicating that PPO uses on average 30 percentage points fewer of the available bins. PPO's lower standard deviation (0.155 vs. 0.198) confirms more predictable bin usage. The training diagnostics show that the agent learned to return to partially filled bins rather than opening new ones, a behaviour driven by the b coefficient in the reward function.

Table 3. Bin penalty statistics across 50 instances

	Mean	Median	Std. Dev.	Min	Max
PPO	0.516	0.500	0.155	0.200	1.000
ACO	0.814	0.800	0.198	0.400	1.000

Source: own calculations

Runtime and Break-Even Analysis

Table 4 shows inference runtimes. PPO's mean of 13.2 s and median of 10.8 s contrast sharply with ACO's mean of 359 s and median of 96.5 s. ACO's distribution is heavily right-skewed (max 1,706 s), driven by large instances where the dynamic colony formula allocates up to 8,000 ants. PPO's maximum of 57 s reflects a fixed network forward pass whose cost scales with box count but not quadratically.

Table 4. Inference runtime (seconds) across 50 instances

	Mean	Median	Std. Dev.	Min	Max
PPO	13.17	10.79	12.09	0.15	56.73
ACO	359.14	96.46	451.18	1.73	1705.72

Source: own calculations

Table 5 presents the break-even analysis. PPO's training phase required approximately 5.5 hours (19,800 s) on the experimental hardware. The mean time saving per instance relative to ACO is 359.14 − 13.17 = 345.97 s. Dividing training cost by saving per instance gives a break-even of ⌈ 19,800 / 345.97 ⌉ = 58 instances. An organisation that solves more than 58 similar packing instances recovers the PPO training investment and thereafter saves 345.97 s per instance. At the median ACO

runtime (96.5 s) the break-even rises to 241 instances, reflecting the fact that easy instances provide smaller time savings.

Table 5. Break-even analysis: PPO training cost vs. inference time saving

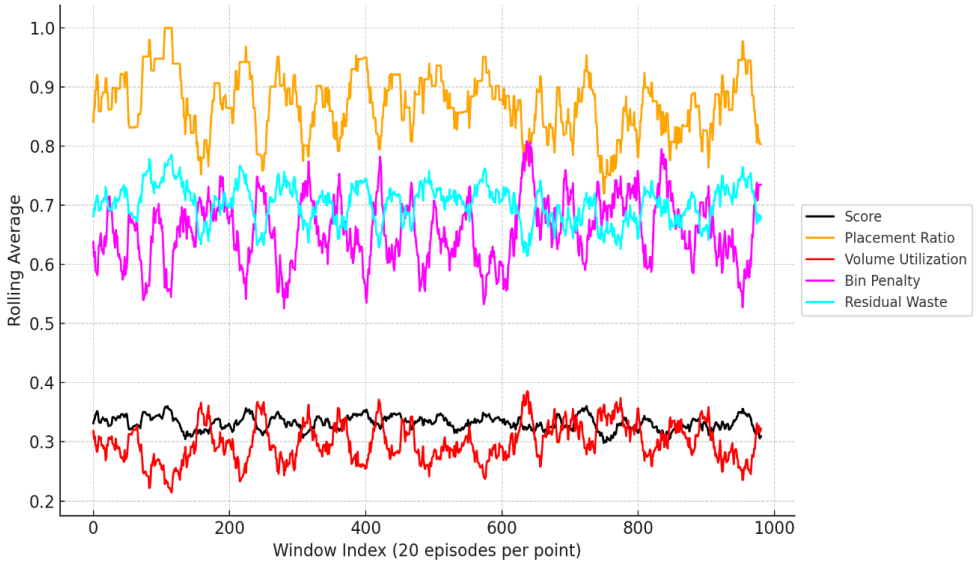
Parameter	Value
PPO training time	19,800 s (5.5 h)
Mean ACO inference time per instance	359.14 s
Mean PPO inference time per instance	13.17 s
Mean time saving per instance	345.97 s
Break-even (mean runtime)	58 instances
Break-even (median runtime, 96.5 s ACO)	241 instances

Source: own calculations. Break-even = ⌈ training time / mean saving per instance⌉ .

Training Dynamics

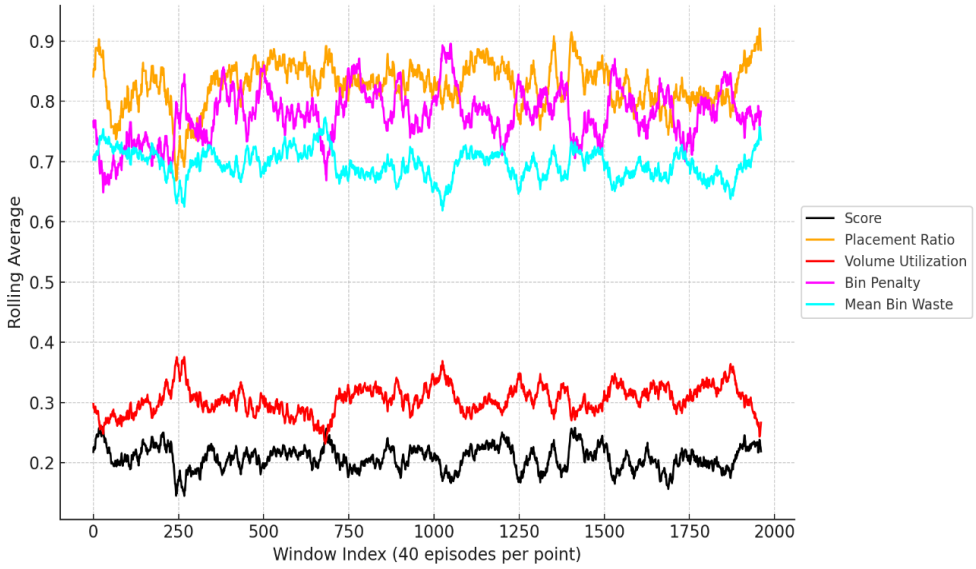
Figures 1–3 show rolling-average training curves for the five metrics (p, u, b, μw, composite score) across three reward formulations. In the first formulation (equal weights 0.40 on p and u), the agent raised placement ratio rapidly but did not reduce bin count, because opening a new bin had no cost that outweighed utilisation gains. Replacing the aggregate waste term with per-bin waste μw and increasing the bin-penalty coefficient shifted the equilibrium: the final training run (Figure 3) shows b declining gradually while p stays above 0.95, confirming that the agent learned to consolidate items into fewer bins rather than maximising per-bin fill in isolation.

Figure 1. Rolling averages (20-episode window) of training metrics, initial reward formulation, 1,000 episodes



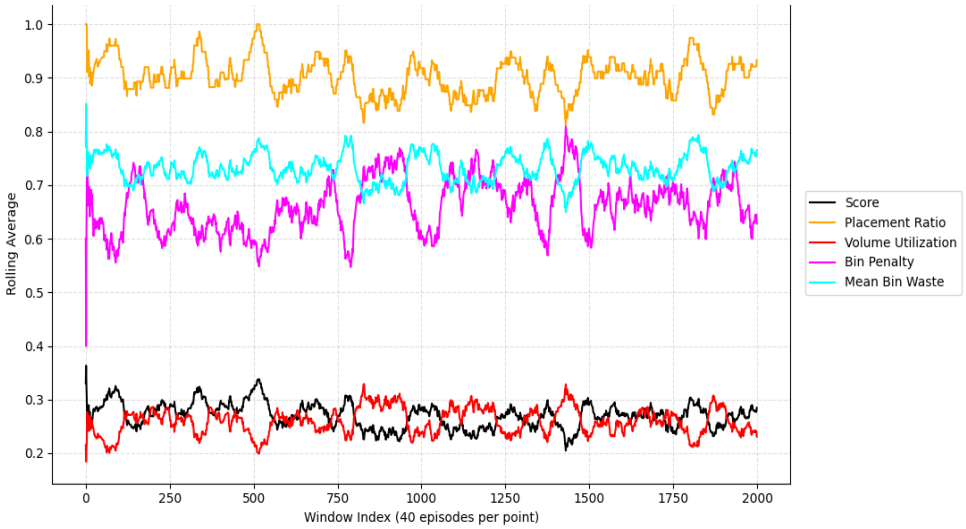
Source: own preparation

Figure 2. Rolling averages (40-episode window), intermediate formulation, 2,000 episodes.



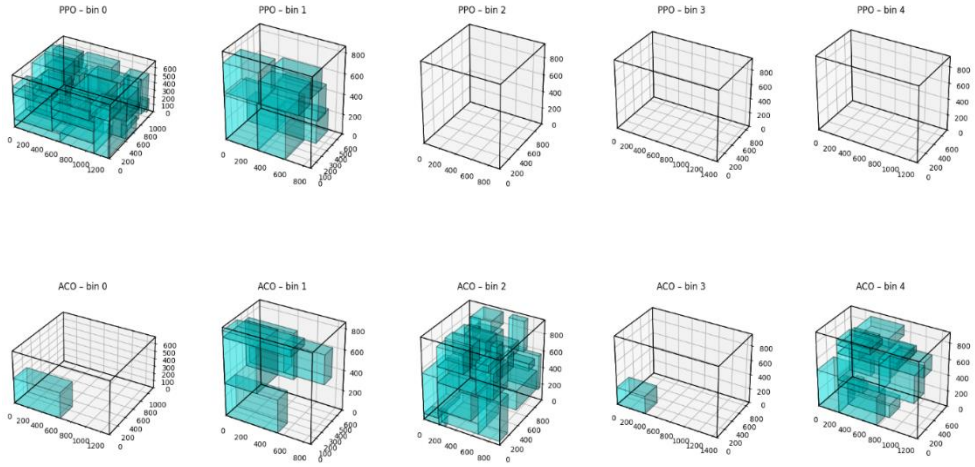
Source: own preparation

Figure 3. Rolling averages (40-episode window), final reward formulation, 2,000 episodes



Source: own preparation

Figure 4. Visual representation of packing solutions: PPO (left) and ACO (right) on a representative instance



Source: own preparation

DISCUSSION

Why PPO Scores Higher

Three implementation choices contributed to PPO's advantage. First, the action-space factorisation (deterministic corner search + learned item–orientation selection) eliminates infeasible placements before they reach the network. This keeps the gradient signal clean and reduces the number of training episodes needed to learn a useful policy [Que et al. 2023; Murdivien, Um 2023]. Second, the step reward $r_t = \text{score}_t - \text{score}_{t-1}$ provides a non-zero learning signal at every step, avoiding the sparse-reward problem that would arise if reward were given only at episode end. Third, the shared actor–critic backbone couples the value function to the same feature representation used for action selection, which stabilises advantage estimation and reduces variance in policy gradient updates.

Why ACO Scores Lower on Average

ACO's higher variance stems from two sources. First, the dynamic colony-size formula may under-allocate ants on instances that appear small by box and bin count but have a complex feasible-placement geometry, causing pheromone trails to converge prematurely on a sub-optimal sequence. Second, 3DHA places each box at the first feasible extreme point in a fixed scanning order; if that order does not match the geometry of a specific instance, no amount of additional ant iterations can recover the missed placement. The 12 ACO victories are consistent with instances where 3DHA's scanning order happens to align with the optimal packing direction, yielding a tight arrangement that the PPO policy, trained on a broader distribution, does not replicate.

Practical Deployment Criteria

The break-even analysis in Table 5 translates the algorithmic comparison into an operational decision rule. PPO is cost-effective when: (a) the organisation solves more than 58 instances of comparable size and structure before the policy requires retraining, and (b) inference latency below 60 s is compatible with the dispatch scheduling window. ACO is preferable when: (a) fewer than 58 instances are expected, (b) the box–bin size distribution changes rapidly (e.g. new product lines introduced weekly), making retraining impractical, or (c) no GPU or training infrastructure is available.

The complementary performance profiles also suggest a practical hybrid: use PPO to generate an initial packing within seconds, then run ACO for a fixed time budget as a local search phase. On the 12 instances where ACO currently wins, its margin averages 0.056; whether an ACO post-processing pass on a PPO solution can capture similar improvements is an empirical question for future work.

Limitations

Four limitations constrain the generalisability of these results. First, all instances were generated from a single synthetic distribution; the algorithms have not been tested on real warehouse order data where box-size distributions are typically skewed by fast-moving SKUs. Second, the lack of per-instance score data in the current thesis prevents a formal paired Wilcoxon test; the 0.063-point mean difference and 76% win rate are descriptively convincing but not inferentially confirmed. Third, both algorithms delegate placement to a deterministic heuristic (corner search for PPO, 3DHA for ACO); end-to-end learned placement such as GOPT [Xiong et al. 2024] may close or reverse the observed gap. Fourth, the dynamic colony-size formula has not been validated against a fixed-size baseline; it is possible that 100 ants with 300 iterations would match or exceed the dynamic formula on most instances at lower computational cost.

SUMMARY

This paper compared a PPO agent and an ACO solver on fifty offline heterogeneous-bin 3D-BPP instances under identical scoring and hardware conditions. The main findings are as follows.

PPO achieves a mean composite score of 0.346 vs. 0.283 for ACO (22% higher) and wins on 76% of instances with a mean margin of 0.101, compared to ACO's mean margin of 0.056 in its 24% of wins. Both algorithms achieve a perfect placement ratio of 1.0 on all instances; the score gap is driven by bin-count efficiency (PPO mean bin penalty 0.516 vs. ACO 0.814). PPO resolves instances in a mean of 13.2 s vs. 359 s for ACO; the PPO training cost of 19,800 s is recovered after 58 instances at mean ACO runtime.

Three actions are required before the results can be considered fully rigorous: (1) extract per-instance score pairs from the benchmark CSV and compute a paired Wilcoxon signed-rank test; (2) run an ablation study comparing the dynamic colony-size formula against a fixed baseline (100 ants, 300 iterations) across instances of varying size; (3) evaluate both algorithms on real warehouse order data to assess external validity.

Subject to these qualifications, the results support the following deployment recommendation: PPO is the preferred solver for operations that process more than approximately 60 packing instances of stable composition, because training cost is recovered within the first 60 runs and inference remains below 60 s on commodity hardware. ACO is preferable for low-volume or rapidly-changing environments where policy retraining is impractical. The complementary win profiles of the two methods suggest that a PPO–ACO hybrid merits empirical investigation.

REFERENCES

Ansel J., Yang E., He H., Gimelshein N., Jain A., Voznesensky M., … Chintala S. (2024) PyTorch 2: Faster Machine Learning through Dynamic Python Bytecode Transformation and Graph Compilation. Proceedings of ASPLOS '24. ACM. https://doi.org/10.1145/3620665.3640366

Christensen H. I., Khan A., Pokutta S., Tetali P. (2016) Multidimensional Bin Packing and Other Related Problems: A Survey. Computer Science Review, 24.

Dorigo M., Gambardella L. M. (1997) Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem. IEEE Transactions on Evolutionary Computation, 1(1), 53-66.

Dorigo M., Stützle T. (2018) Ant Colony Optimization: Overview and Recent Advances. [in:] Handbook of Metaheuristics. Springer, 311-351.

Dyckhoff H. (1990) A Typology of Cutting and Packing Problems. European Journal of Operational Research, 44(2), 145-159.

Dyckhoff H., Finke U. (1992) Cutting and Packing in Production and Distribution: A Typology and Bibliography. Springer.

Elhedhli S., Gzara F., Yildiz B. C. (2019) Three-Dimensional Bin Packing and Mixed-Case Palletization. INFORMS Journal on Optimization, 1(4), 323-352.

Garey M. R., Johnson D. S. (1979) Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman.

gym-BinPack3D [Source Code] (2024) https://github.com/ylchan87/gym-BinPack3D

Gzara F., Elhedhli S., Yildiz B. C. (2020) The Pallet-Loading Problem: Three-Dimensional Bin Packing with Practical Constraints. European Journal of Operational Research, 287, 545-565.

Hu H., Zhang X., Yan X., Wang L., Xu Y. (2017) Solving a New 3D Bin Packing Problem with Deep Reinforcement Learning Method. arXiv preprint arXiv:1708.05930.

Kanna S. K., Jaisree A. D., Balasundaram K., Kumar S. B. (2015) Optimization of 3D Constrained Rectangular Bin Packing Problem using Recursive Ant Colony Algorithm. IOSR Journal of Mechanical and Civil Engineering, 12(4), 65-70.

Karaboga D., Basturk B. (2008) On the Performance of Artificial Bee Colony (ABC) Algorithm. Applied Soft Computing, 8(1), 687-697.

Lampropoulos A. S., Tsihrintzis G. A. (2015) Machine Learning Paradigms: Applications in Recommender Systems. Springer International Publishing.

Levine J., Ducatelle F. (2004) Ant Colony Optimization and Local Search for Bin Packing and Cutting Stock Problems. Journal of the Operational Research Society, 55(7), 705-716.

Li X., Zhang K. (2015) A Hybrid Differential Evolution Algorithm for Multiple Container Loading Problem with Heterogeneous Containers. Computers & Industrial Engineering, 90, 305-313.

Maia D. B., Borenstein D. (2024) A Hybrid Approach Combining Ant Colony Optimization and Column Generation for Solving the 3D Bin Packing Problem with Rotation. Available at SSRN 4978748 [preprint].

Manthey B., van Rhijn J., Safari A., Vredeveld T. (2025) Convergence and Running Time of Time-Dependent Ant Colony Algorithms. arXiv preprint arXiv:2501.10810 [preprint].

Murdivien S. A., Um J. (2023) BoxStacker: Deep Reinforcement Learning for 3D Bin Packing Problem in Virtual Environment of Logistics Systems. Sensors, 23(15), 6928.

Ojha V. K., Abraham A., Snášel V. (2014) ACO for Continuous Function Optimization: A Performance Analysis. Proceedings of the 14th International Conference on Intelligent Systems Design and Applications, 145-150. IEEE.

Que Q., Yang F., Zhang D. (2023) Solving 3D Packing Problem using Transformer Network and Reinforcement Learning. Expert Systems with Applications, 214, 119153.

Sangeetha V., Krishankumar R., Ravichandran K. S., Cavallaro F., Kar S., Pamucar D., Mardani A. (2021) A Fuzzy Gain-Based Dynamic Ant Colony Optimization for Path Planning in Dynamic Environments. Symmetry, 13(2), 280.

Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. (2017) Proximal Policy Optimization Algorithms. arXiv preprint arXiv:1707.06347.

Silveira M. E., Vieira S. M., Da Costa Sousa J. M. (2013) An ACO Algorithm for the 3D Bin Packing Problem in the Steel Industry. [in:] Recent Trends in Applied Artificial Intelligence: IEA/AIE 2013, 535–544. Springer Berlin Heidelberg.

Singh N. K., Baidya S. (2013) A Novel Work for Bin Packing Problem by Ant Colony Optimization. International Journal of Research in Engineering and Technology, 2(2), 71-73.

Skackauskas J., Kalganova T., Dear I., Janakiram M. (2022) Dynamic Impact for Ant Colony Optimization Algorithm. Swarm and Evolutionary Computation, 69, 100993.

Viegas J. P., Vieira S. M., Sousa J. M., Henriques E. M. (2014) Metaheuristics for the 3D Bin Packing Problem in the Steel Industry. 2014 IEEE Congress on Evolutionary Computation (CEC), 338-343. IEEE.

Viegas J. L., Vieira S. M., Henriques E. M., Sousa J. M. (2015) A Tabu Search Algorithm for the 3D Bin Packing Problem in the Steel Industry. CONTROLO 2014 – Proceedings of the 11th Portuguese Conference on Automatic Control, 355-364. Springer.

Wäscher G., Haußner H., Schumann H. (2007) An Improved Typology of Cutting and Packing Problems. European Journal of Operational Research, 183(3), 1109-1130.

Wilcoxon F. (1945) Individual Comparisons by Ranking Methods. Biometrics Bulletin, 1(6), 80-83.

Xiong H., Guo C., Peng J., Ding K., Chen W., Qiu X., … Xu J. (2024) GOPT: Generalizable Online 3D Bin Packing via Transformer-Based Deep Reinforcement Learning. IEEE Robotics and Automation Letters.

Zeineldin R. A., Morsy A. M. (2015) A Modified Artificial Bee Colony for Solving the Container Loading Problem. International Journal of Computer Applications, 114(3).

Zhao H., Zhu C., Xu X., Huang H., Xu K. (2022) Learning Practically Feasible Policies for Online 3D Bin Packing. Science China Information Sciences, 65(1), 112105.

METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XXVII/1, 2026, s. 41 – 54

Received: 11.02.2026
Accepted: 20.03.2026

ANALIZA WPŁYWU LOKALIZACJI I CECH NIERUCHOMOŚCI NA ICH CENY W POLSCE W LATACH 2023-2024

Angelika Samson
Wydział Zastosowań Informatyki i Matematyki
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
e-mail: angelika_samson@sggw.edu.pl
Monika Zielińska-Sitkiewicz https://orcid.org/0000-0003-4829-3239
Instytut Ekonomii i Finansów
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
e-mail: monika_zielinska-sitkiewicz@sggw.edu.pl

Streszczenie: W artykule przeprowadzono badanie wpływu metrażu, lokalizacji i udogodnień na ceny ofertowe mieszkań w Polsce w latach 2023-2024, ze szczególnym uwzględnieniem Warszawy, Krakowa, Łodzi i Trójmiasta. Połączono metody ekonometryczne i uczenie maszynowe, aby wskazać kluczowe czynniki cenotwórcze. W analizie zastosowano modele OLS, Lasso, GWR, k-means oraz Random Forest. Badanie oparte na próbie 49 531 ofert wykazuje wyraźną przewagę modeli nieliniowych w predykcji. Potwierdzono również dominującą rolę lokalizacji i cech fizycznych w objaśnianiu przestrzennego zróżnicowania cen na rynku.

Słowa kluczowe: wycena nieruchomości, metraż, lokalizacja, udogodnienia, modele OLS, Lasso, GWR, k-means, Random Forest

JEL classification: R31, C53, C55, R12

WSTĘP

Rynek nieruchomości stanowi jeden z kluczowych sektorów gospodarki, ściśle powiązany z cyklem koniunkturalnym, polityką pieniężną oraz sytuacją dochodową społeczeństwa. W ostatnich latach, w warunkach rosnącej inflacji, zmieniających się preferencji mieszkaniowych oraz rozwoju pracy zdalnej, temat wyceny nieruchomości zyskał szczególne znaczenie. Dostęp do rzetelnych modeli wyceny staje się dziś niezwykle istotny dla rzeczoznawców majątkowych, deweloperów, instytucji finansowych oraz gospodarstw domowych.

https://doi.org/10.22630/MIBE.2026.27.1.3

W perspektywie długookresowej polski rynek mieszkaniowy charakteryzował się systematycznym wzrostem cen, przerywanym okresami spowolnienia gospodarczego (np. globalny kryzys finansowy, pandemia) oraz zmian polityki pieniężnej i warunków finansowania [NBP 2025]. W latach 2023-2024 obserwowano łagodniejsze tempo wzrostu cen, a w I-II kw. 2025 r. sygnały stabilizacji i miejscami niewielkich korekt w części lokalizacji, co NBP wiązał m.in. z wygaszeniem programu dopłat do kredytów oraz dostosowaniami po stronie popytu i podaży [NBP 2024; NBP 2025]. Jednocześnie analizy Polskiego Instytutu Ekonomicznego dla 2024 r. wskazywały na istotne zróżnicowanie dynamiki między miastami oraz różnice pomiędzy rynkiem pierwotnym a wtórnym [PIE 2024]. Ponadto istotnym czynnikiem kształtującym koniunkturę były instrumenty publiczne – funkcjonowanie programu „Bezpieczny Kredyt 2%" w 2023 roku sprzyjało wzrostowi popytu, podczas gdy jego wygaśnięcie z końcem roku osłabiło dynamikę transakcyjną [NBP 2024; UKNF 2024; Szelągowska 2023]. Równolegle uwarunkowania podażowe - koszty gruntów i materiałów, ograniczenia planistyczne oraz dostępność terenów utrwalały istotne zróżnicowanie regionalne i lokalne [NBP 2025; PIE 2024].

W literaturze współczesna analiza rynku nieruchomości opiera się na trzech fundamentalnych podejściach wyceny: porównawczym, dochodowym i hybrydowym [Pagourtzi et al., 2003; Sirmans et al., 2005] przy czym modele hedoniczne pozostają głównym narzędziem badawczym. Teoretyczne podstawy koncepcji cen ukrytych, gdzie wartość nieruchomości jest postrzegana jako suma jej indywidualnych atrybutów, zostały ustanowione przez Rosena [1974], natomiast kompleksowy przegląd ich empirycznych zastosowań i metod selekcji zmiennych przedstawili Sirmans, Macpherson i Zietz [2005]. Jednakże, skuteczne ilościowe określenie wpływu cech fizycznych i lokalizacyjnych na cenę wymaga uwzględnienia autokorelacji przestrzennej, ponieważ zdaniem Anselina [2001] pominięcie zależności sąsiedztwa w modelach SAR lub SEM prowadzi do obciążonych estymatorów i błędnych wniosków. Ponadto, pełne zrozumienie mechanizmów rynkowych wymaga umiejscowienia ich w kontekście makroekonomicznym, gdzie sektor mieszkaniowy jest uznawany za kluczowy wiodący wskaźnik cyklu koniunkturalnego [Leamer 2007] oraz włączenia perspektywy behawioralnej. Jak zauważyli Case i Shiller [2003], dynamika cen jest często napędzana przez czynniki psychologiczne i specyficzne oczekiwania kupujących, co może prowadzić do baniek spekulacyjnych odbiegających od obiektywnych fundamentów ekonomicznych.

W badaniach nad analizą cen nieruchomości szczególną wagę przywiązuje się do metod ilościowych. Tradycyjne indeksy cen często nie są w stanie w pełni oddać dynamiki rynku, gdyż nie uwzględniają zmienności cech nieruchomości [Foryś, 2015]. W celu ograniczenia tych niedoskonałości stosuje się modele oparte na regresji hedonicznej, dekomponujące cenę na wartości poszczególnych atrybutów, takich jak powierzchnia, rok budowy czy lokalizacja. Badania Foryś [2015] nad rynkiem szczecińskim potwierdziły skuteczność tej metody, wykazując dominujący

wpływ powierzchni użytkowej na ceny mieszkań. Metoda ta posiada jednak ograniczenia w przypadku gwałtownych zmian rynkowych i wymaga dużych, homogenicznych zbiorów danych.
Z drugiej strony Dąbrowski [2010] udokumentował, że podejście oparte wyłącznie na lokalnych cechach nieruchomości jest niewystarczające do opisu złożonych zjawisk rynkowych. Wprowadzenie do modeli tzw. atrybutów globalnych (wskaźników makroekonomicznych i społeczno-gospodarczych, np. inflacji, produkcji przemysłowej) znacząco poprawia trafność prognoz. Dąbrowski wykazał silne powiązania między sytuacją gospodarczą kraju a poziomem cen, postulując potrzebę stosowania wielowymiarowych analiz dynamicznych.
Współczesna literatura coraz częściej odchodzi od klasycznej ekonometrii w stronę zaawansowanych algorytmów. Naji [2024], badając rynek nieruchomości, zaprezentował podejście z wykorzystaniem uczenia maszynowego (m.in. Random Forest, Lasso, Stacking Regressor). Wyniki jego pracy dowiodły, że modele zespołowe (ensemble models) osiągają najwyższą dokładność predykcyjną dzięki efektywnej integracji danych z różnych źródeł (w tym web scrapingu i analityki przestrzennej). Uczenie maszynowe pozwala na lepsze radzenie sobie ze złożonym i nieliniowym charakterem rynków, precyzyjniej oceniając wpływ lokalizacji, metrażu i udogodnień.

W kontekście zmian rynkowych i ewolucji metodologicznej, analiza empiryczna mechanizmów wyceny na polskim rynku zyskuje szczególne znaczenie. Celem artykułu jest empiryczne zbadanie stopnia, w jakim metraż, lokalizacja oraz wybrane udogodnienia determinują ceny ofertowe mieszkań w Polsce, ze szczególnym uwzględnieniem największych rynków: Warszawy, Krakowa, Łodzi i Trójmiasta w latach 2023-2024. Badanie łączy tradycyjne podejście ekonometryczne z nowoczesnymi metodami uczenia maszynowego, aby zidentyfikować czynniki o największej sile wyjaśniającej w warunkach silnego zróżnicowania przestrzennego i makroekonomicznego polskiego rynku mieszkaniowego.

METODOLOGIA I DANE

Materiał badawczy stanowił pozyskany z platformy Kaggle publiczny zbiór 49 531 ogłoszeń z polskiego rynku nieruchomości [Jamróz 2024]. Na potrzeby realizacji celu naukowego zbiór ograniczono wyłącznie do ofert sprzedaży opublikowanych w okresie od sierpnia 2023 r. do czerwca 2024 r., wykluczając rekordy dotyczące najmu. Takie zawężenie pozwoliło na precyzyjną analizę czynników kształtujących ceny rynkowe oraz badanie ich zróżnicowania przestrzennego. Przed przystąpieniem do modelowania dane poddano procedurom oczyszczania, standaryzacji i transformacji, co umożliwiło ich efektywne wykorzystanie w analizach statystycznych.

W badaniu uwzględniono pełne spektrum zmiennych pozwalających na wielowymiarowy opis nieruchomości. Zmienną zależną była cena ofertowa mieszkania wyrażona w PLN (price). Zbiór zmiennych niezależnych obejmował:

- parametry fizyczne i strukturalne: powierzchnia użytkowa w m² (squareMeters), rok oddania budynku do użytku (buildYear), piętro, na którym znajduje się lokal (floor), oraz całkowita liczba kondygnacji w budynku (floorCount);
- atrybuty budynkowe: typ zabudowy (buildingType, np. blok, kamienica) oraz materiał konstrukcyjny (buildingMaterial, np. cegła, żelbeton);
- zmienne lokalizacyjne i przestrzenne: miasto zakodowane binarnie (city, m.in. Warszawa, Kraków, Gdańsk i Łódź), współrzędne geograficzne (latitude, longitude), odległość od ścisłego centrum miasta w kilometrach (centreDistance) oraz wskaźnik dostępności usług mierzony jako średnia odległość do najbliższych punktów użyteczności publicznej (avgPOIDistance);
- udogodnienia (zmienne binarne): posiadanie miejsca parkingowego (parking_space), balkonu (balcony), windy (elevator), ochrony (security) oraz komórki lokatorskiej (storage_room);
- czas: miesiąc (month) oraz rok (year) publikacji ogłoszenia.

Proces badawczy zrealizowano w środowisku Python, wykorzystując biblioteki pandas, numpy, scikit-learn, statsmodels, matplotlib oraz geopandas. Obejmował on następujące etapy:

1. Eksploracyjna analiza danych (EDA) - przeprowadzono winsoryzację wartości odstających w rozkładach cen i powierzchni. Obliczono podstawowe miary statystyczne oraz macierz korelacji Pearsona. Zidentyfikowano i usunięto zmienne silnie skorelowane w celu redukcji kolinearności [Ligas & Czaja 2010; Śleszyński 2020; Chaim & Łukasik 2024; Rana & Singhal 2015].
2. Modelowanie regresyjne - estymowano model klasycznej regresji liniowej (OLS) w celu identyfikacji globalnego wpływu predyktorów na zmienną zależną. Równolegle zastosowano regresję Lasso (Least Absolute Shrinkage and Selection Operator) wykorzystując regularyzację L_1 w celu jednoczesnej estymacji i selekcji zmiennych [Berry & Feldman 1985; Bun & Harrison 2019; Fatih 2024]
3. Segmentacja rynku (klasteryzacja) - wykorzystano algorytm k-means do identyfikacji homogenicznych grup nieruchomości. Optymalna liczba klastrów została określona metodą „łokcia" (elbow method), co zapewniło kompromis między liczbą grup a ich spójnością [MacQueen 1967; Arthur & Vassilvitskii 2007; Huang & Lai 2023].
4. Analityka przestrzenna (GWR) - zastosowano lokalną regresję ważoną geograficznie (Geographically Weighted Regression – GWR) z adaptacyjnym pasmem umożliwiającym estymację współczynników lokalnych dla poszczególnych lokalizacji. Dodatkowo zweryfikowano występowanie

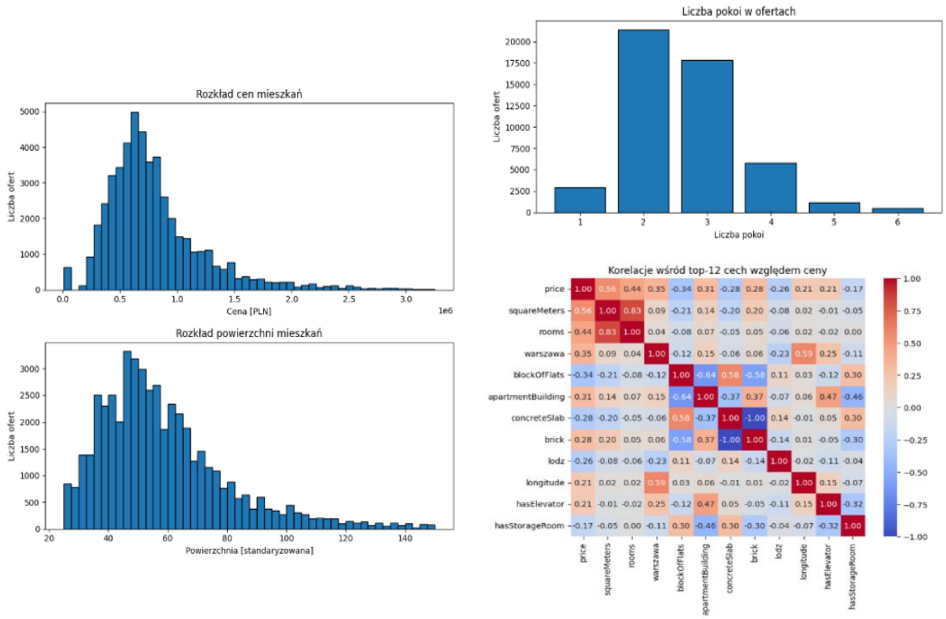
autokorelacji przestrzennej przy użyciu indeksu Morana (Moran's I) [Brunsdon, Fotheringham & Charlton 1996, 2000, 2006; Matthews & Yang 2012].

5. Modelowanie nieliniowe (Random Forest) - w celu uchwycenia złożonych i nieliniowych zależności między atrybutami a ceną, zbudowano model lasu losowego z wykorzystaniem pięciokrotnej walidacji krzyżowej. Dokonano oceny jakości predykcji na zbiorze testowym (miary R^2 i RMSE), a także przeanalizowano ważność poszczególnych cech [Breiman 2001].
6. Analiza porównawcza aglomeracji - dla czterech największych aglomeracji w Polsce opracowano odrębne modele OLS i Random Forest. Porównano wartości współczynników β, miary dopasowania R^2 oraz błędy RMSE.

WYNIKI BADAŃ

Eksploracyjną analizą danych scharakteryzowano rozkłady najważniejszych zmiennych oraz zidentyfikowano potencjalne odchylenia od typowych wartości rynkowych. Na podstawie całego zbioru obliczono medianę ceny - 525 000 PLN oraz medianę powierzchni mieszkania 48 m². Dane te potwierdzają, że analizowany zbiór obejmuje w przeważającej mierze mieszkania średniej wielkości, o przeciętnej wartości transakcyjnej.

Rysunek 1. Charakterystyka rozkładów i korelacja badanych zmiennych

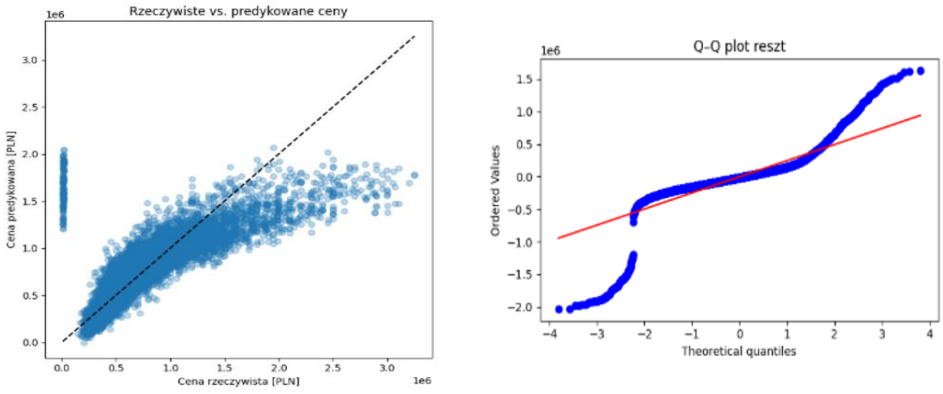


Źródło: Obliczenia i opracowanie własne

Na rysunku 1 przedstawiono rozkłady cen ofertowych oraz powierzchni charakteryzujące się wyraźną prawoskośnością. Występowanie dodatniej asymetrii świadczy o rynkowej dominacji lokali o niższych i średnich parametrach oraz obecności nielicznej grupy ofert o skrajnie wysokich wartościach cenowych i powierzchniowych, przy jednoczesnej przewadze rynkowej mieszkań dwu- i trzypokojowych. Na podstawie analizy korelacji (heatmapa na rys. 1) stwierdzono, że cena jest najsilniej powiązana z metrażem (r=0,84), liczbą pokoi (r=0,60) oraz obecnością windy (r=0,55). W celu wyeliminowania multikolinearności, wynikającej z silnej zależności między zmiennymi squareMeters i rooms oraz tożsamości cech brick i concreteSlab, przeprowadzono redukcję wymiarów. Usunięcie zmiennej rooms (na rzecz squareMeters o wyższej sile objaśniającej) oraz concreteSlab pozwoliło na zapewnienie stabilności numerycznej i poprawę interpretowalności estymowanych modeli.

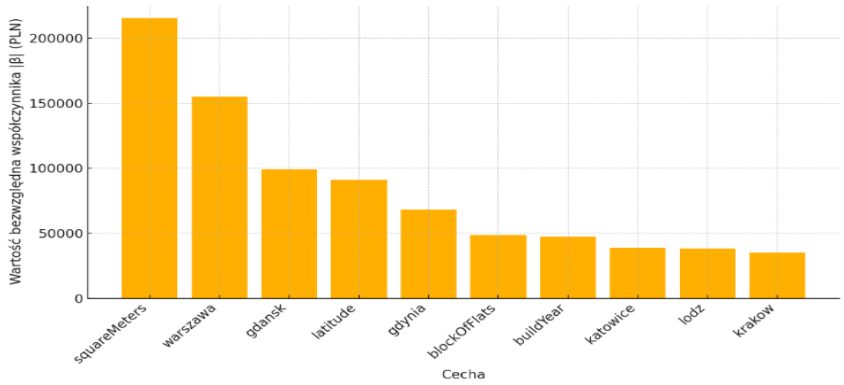
W celu oszacowania determinant cen ofertowych zastosowano klasyczny model najmniejszych kwadratów (OLS) oraz regresję Lasso z regularyzacją L_1. Model OLS uzyskał dopasowanie na poziomie $R^2 = 0{,}57$ oraz RMSE $\approx$ 286 379 PLN, poprawnie odwzorowując środek rozkładu, przy jednoczesnej tendencji do niedoszacowania wartości najdroższych nieruchomości. Analiza reszt wykazała występowanie heteroskedastyczności oraz odstępstwa od rozkładu normalnego, co jest zjawiskiem typowym dla danych rynkowych o nieliniowych zależnościach (Rysunek 2). Wyniki regresji Lasso ($\alpha \approx 0{,}015$) okazały się zbliżone do OLS: $R^2 = 0{,}565$, RMSE $\approx$ 286 718 PLN, co potwierdziło stabilność oszacowań. Zidentyfikowano, że kluczowym predyktorem ceny jest metraż mieszkania (squareMeters), a wśród zmiennych lokalizacyjnych największy wpływ wykazują aglomeracje Warszawy, Gdańska i Gdyni (Rysunek 3). Istotnymi czynnikami różnicującymi ceny okazały się również parametry geograficzne (latitude), rok budowy (buildYear) oraz typ zabudowy.

Rysunek 2. Zależność pomiędzy cenami rzeczywistymi a predykowanymi (OLS) oraz analiza reszt modelu OLS



Źródło: Obliczenia i opracowanie własne

Rysunek 3. Dziesięć cech o największych modułach współczynników (Lasso)

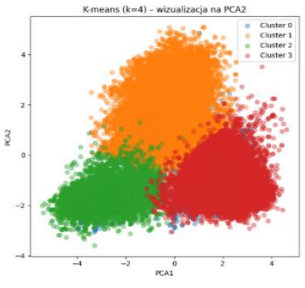


Źródło: Obliczenia i opracowanie własne

Identyfikując homogeniczne grupy ofert o zbliżonej charakterystyce rynkowej zastosowano algorytm k-means. Optymalną liczbę klastrów (k=4) wyznaczono metodą „łokcia", analizując spadek inercji wewnątrzklasowej. Wyniki segmentacji, przedstawione w tabeli 1, pozwoliły na wyodrębnienie zróżnicowanych podgrup mieszkań różniących się istotnie pod względem parametrów cenowych, metrażu oraz lokalizacji. Klaster 0 identyfikuje segment mieszkań o charakterze ekonomicznym, w którym dominują lokale o relatywnie niższych metrażach i cenach ofertowych, zlokalizowane głównie w blokach mieszkalnych w niewielkiej odległości od punktów usługowych. Klaster 1 reprezentuje segment ekonomiczny o najniższej średniej cenie (606 tys. PLN), natomiast Klaster 3 grupuje lokale o najwyższym standardzie cenowym (980 tys. PLN) i największej powierzchni średniej (63,11 m^2). Z kolei Klaster 2 charakteryzuje się najmniejszą średnią odległością od punktów usługowych (0,364 km), co sugeruje lokalizacje w ścisłych centrach miast. Separację i strukturę wyodrębnionych segmentów potwierdza wizualizacja w przestrzeni dwóch głównych składowych (PCA), wskazująca na wyraźne zróżnicowanie między rynkiem masowym a segmentem premium.

Tabela 1 Charakterystyka klastrów rynku mieszkaniowego

Klaster	Średnia cena (PLN)	Mediana ceny (PLN)	Średni metraż m^2	Średnia odległość od POI	Średnia dl. geogr. E	Średnia szer. geogr. N
0	870499,61	775000	58,85	0,638	18,62	54,37
1	605897,55	580000	52,46	0,586	19,62	51,76
2	857588,93	735000	62,14	0,364	19,25	51,82
3	979864,74	875000	63,11	0,720	19,89	51,76



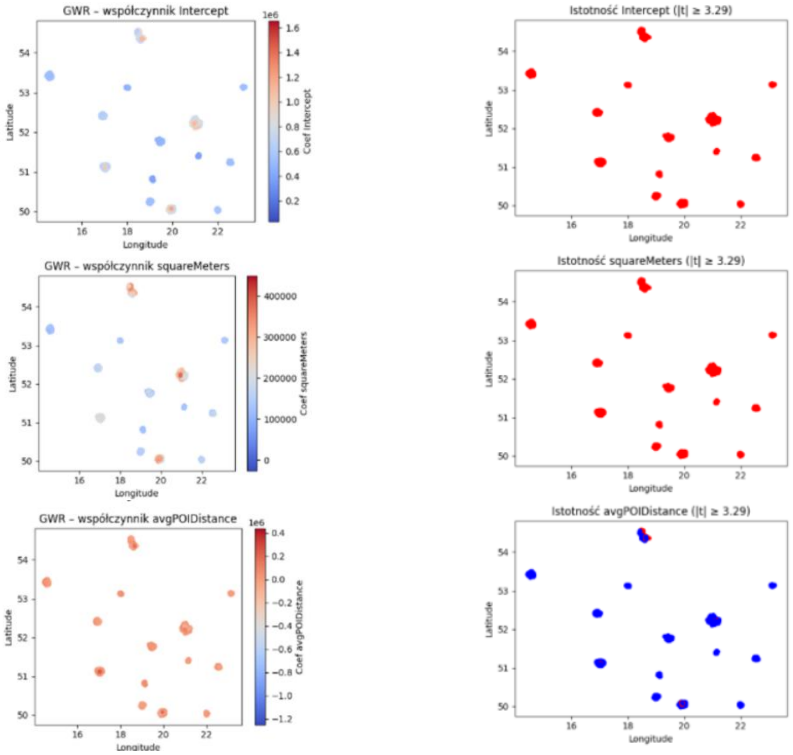
Źródło: Obliczenia własne

Uwzględnienie niestacjonarności przestrzennej mechanizmów kształtowania cen za pomocą lokalnej regresji ważonej geograficznie GWR pozwoliło na zidentyfikowanie najwyższej ceny bazowej (wyrazu wolnego) w centralnych obszarach aglomeracji oraz najsilniejszego wpływu metrażu (squareMeters) na premię cenową w strefach o najwyższym stopniu zurbanizowania. Zaobserwowano również regionalną zmienność wpływu dostępności usług (avgPOIDistance), którego oddziaływanie na cenę jest wyraźniejsze w południowo-zachodniej części badanego obszaru. Weryfikacja istotności statystycznej lokalnych współczynników regresji ($\alpha = 0{,}001$, $|t| \geq 3{,}29$) umożliwiła precyzyjne określenie przestrzennego zasięgu oddziaływania poszczególnych zmiennych. Stwierdzono, że lokalny wyraz wolny zachowuje istotność niemal w całym obszarze badań, co potwierdza powszechność lokalnego efektu cenowego, występującego niezależnie od cech fizycznych nieruchomości. Z kolei wpływ metrażu koncentruje się wokół głównych ośrodków miejskich i ich przedmieść. Istotność dostępności usług wykazuje natomiast charakter punktowy i ogranicza się głównie do południowo-zachodniej części obszaru, co sugeruje, że w tym regionie bliskość punktów usługowych stanowi krytyczny czynnik kształtujący ceny. Wyniki te dowodzą, że choć większość cech wykazuje znaczenie globalne, siła ich oddziaływania ulega istotnemu zróżnicowaniu przestrzennemu, osiągając najwyższą stabilność statystyczną w rejonach o dużej intensywności procesów rynkowych (Rysunek 4). Diagnostyka modelu przy użyciu statystyki Morana wykazała niewielką, choć istotną autokorelację reszt (Moran's $I = 0{,}04$ $p<0{,}01$), co przy tak niskiej wartości współczynnika potwierdza skuteczność modelu GWR w eliminowaniu efektów sąsiedztwa i poprawnym ujęciu struktury przestrzennej danych.

Wykorzystanie modelu lasu losowego umożliwiło uwzględnienie nieliniowych zależności oraz złożonych interakcji zachodzących między predyktorami. Algorytm wytrenowany przy użyciu 5-krotnej walidacji krzyżowej (z parametrami n_estimators = 500, max_depth = None oraz max_features = 'sqrt') wykazał wysoką zdolność predykcyjną i stabilność, osiągając współczynnik determinacji $R^2 = 0{,}742$ na zbiorze testowym oraz błąd RMSE = 220 729 PLN. Minimalna rozbieżność między wynikami walidacji ($0{,}745 \pm 0{,}01$) a zbiorem testowym, przy błędzie OOB na poziomie 0,75, potwierdza wysoką zdolność generalizacji i brak przeuczenia modelu. Analiza ważności zmiennych ujawniła dominującą rolę metrażu (squareMeters) w kształtowaniu cen ofertowych. Kluczowe znaczenie przypisano również atrybutom przestrzennym (współrzędne geograficzne, odległość od centrum, lokalizacja w Warszawie) oraz strukturalnym (buildYear) (Rysunek 5).

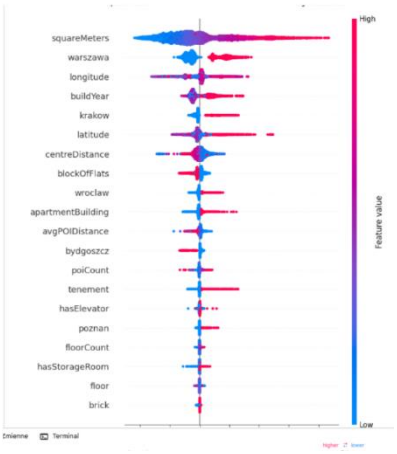
Analiza porównawcza czterech kluczowych rynków — Warszawy, Trójmiasta, Krakowa i Łodzi — pozwoliła na identyfikację istotnych różnic w strukturze cenowej oraz hierarchii determinant wartości mieszkań. Podczas gdy Warszawa i Trójmiasto charakteryzują się najwyższymi medianami cen oraz znaczną heterogenicznością powierzchniową, rynki w Krakowie i Łodzi wykazują relatywnie większą jednorodność ofert.

Rysunek 4 Rozkład lokalnych współczynników regresji ważonej geograficznie (wyraz wolny, metraż, odległość od POI) wraz z mapami istotności parametrów.



Źródło: Obliczenia własne

Rysunek 5 Ważność cech w modelu Random Forest



Źródło: obliczenia własne

W czterech badanych miastach odnotowano prawoskośność rozkładów cen i metrażu, typową dla sektora mieszkaniowego o przewadze lokali o mniejszej powierzchni. Porównanie efektywności modeli OLS i Random Forest (RF) jednoznacznie wskazuje na wyższą zdolność predykcyjną algorytmów zespołowych, które pozwoliły na redukcję błędu RMSE oraz wzrost współczynnika R^2 średnio o 0,25–0,30, przy czym największą poprawę dopasowania odnotowano dla Warszawy i Trójmiasta.

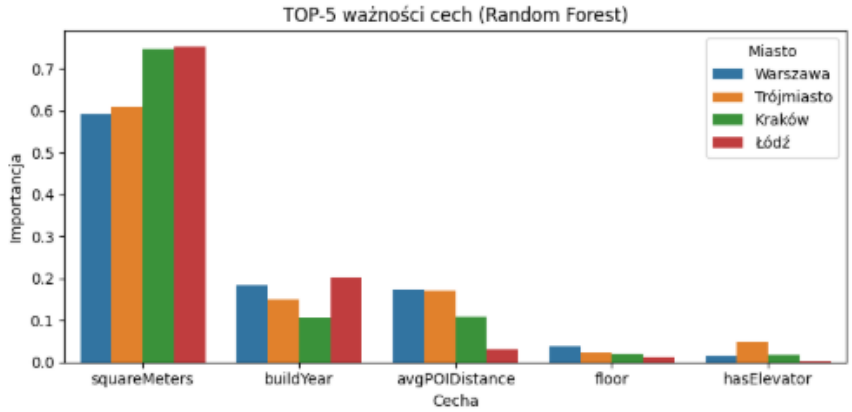
Tabela 2 Wyniki R^2 i RMSE dla modeli OLS i Random Forest

Miasto	N ofert	R^2 OLS	RMSE OLS	R^2 RF	RMSE RF
Warszawa	17481	0,285	426 389	0,608	314 922
Trójmiasto	5368	0,621	247 463	0,828	165 927
Kraków	7494	0,646	239 267	0,781	202 131
Łódź	4414	0,781	76 108	0,885	58 109

Źródło: obliczenia własne

Mimo różnic lokalnych, metraż (squareMeters) pozostaje najsilniejszym predyktorem ceny w każdym z analizowanych miast. Interesujące zróżnicowanie zaobserwowano w przypadku dostępności infrastruktury usługowej - o ile w Warszawie, Trójmieście i Krakowie bliskość usług generuje istotną premię cenową (ujemny współczynnik β), o tyle w Łodzi zależność ta przyjmuje kierunek dodatni, co sugeruje odmienną specyfikę przestrzenną tego ośrodka. Analiza ważności cech w modelach RF (Rysunek 6) potwierdza dominację metrażu, wskazując jednocześnie na istotną, choć zróżnicowaną rolę roku budowy (buildYear) oraz dostępności punktów usługowych, przy marginalnym wpływie pozostałych atrybutów fizycznych, takich jak piętro czy obecność windy.

Rysunek 6 Ważność cech w modelu Random Forest dla czterech miast



Źródło: obliczenia własne

PODSUMOWANIE

Celem pracy było empiryczne zbadanie, w jakim stopniu metraż, lokalizacja i wybrane udogodnienia determinują ceny ofertowe mieszkań w Polsce oraz w wybranych dużych miastach (Warszawa, Trójmiasto, Kraków, Łódź). Analiza została przeprowadzona na zbiorze 49 531 unikalnych ofert z lat 2023–2024, po uprzednim oczyszczeniu danych (eliminacja duplikatów, imputacja braków, winsoryzacja 1% skrajnych wartości) oraz podziale na zbiory uczący i testowy. Wykazano prawoskośność rozkładów cen i metraży, dominację mieszkań 2–3-pokojowych oraz różnice cen między typami budynków. Redukcja cech ograniczyła multikolinearność (m.in. rezygnacja z concreteSlab oraz pozostawienie squareMeters zamiast rooms).

Wyniki potwierdziły, że metraż jest najsilniejszym predyktorem ceny (korelacja Pearsona z ceną r = 0,84; dominująca ważność w Random Forest). Wskazano jednocześnie zróżnicowanie przestrzenne znaczenia determinant: w Warszawie przeważają czynniki lokalizacyjne (m.in. odległość od centrum, gęstość usług), podczas gdy w Łodzi większą rolę odgrywają wielkość i cechy budynku. Modele globalne (OLS, Lasso) wyjaśniały ok. 57% wariancji cen ($R^2 \approx 0{,}57$; RMSE ≈ 286 tys. PLN), z typową dla danych rynkowych heteroskedastycznością reszt. Uwzględnienie niestacjonarności przestrzennej w GWR podniosło dopasowanie do $R^2 = 0{,}62$ (RMSE ≈ 268 tys. PLN), co potwierdziło istotność lokalnych efektów. Najwyższą trafność predykcji uzyskano w Random Forest (R^2 test = 0,742; RMSE ≈ 220 729 PLN; CV $R^2 = 0{,}745 \pm 0{,}010$; OOB ≈ 0,75), co wskazuje, że nieliniowe zależności i interakcje są na tym rynku znaczące. Segmentacja k-means wyodrębniła cztery spójne klastry (od segmentu ekonomicznego po premium), potwierdzając heterogeniczność struktury popytowo-podażowej.

Uzyskane wyniki empirycznie potwierdzają mechanizmy formowania się cen mieszkań w Polsce w latach 2023–2024 oraz wskazują, że połączenie klasycznych metod ekonometrycznych z modelami uczenia maszynowego (zwłaszcza Random Forest) stanowi skuteczne podejście do analizy złożonych, przestrzennie zróżnicowanych danych rynkowych.

BIBLIOGRAFIA

Anselin L. (2001) Spatial Econometrics [in:] B. H. Baltagi (red.) A Companion to Theoretical Econometrics, Chapter 14, 310-330, Blackwell Publishing Ltd.

Arthur D., Vassilvitskii S. (2007) K-means++: The Advantages of Careful Seeding, SODA ’07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, New Orleans Louisiana, 1027-1035.

Berry W. D., Feldman S. (1985) Multiple Regression in Practice. Sage University Paper series on Quantitative Applications in the Social Sciences, Series No. 07-050, Sage, Newbury Park.

Breiman L. (2001) Random Forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

Brunsdon C., Fotheringham A. S., Charlton M. (1996) Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. Geographical Analysis, 28(4), 281-298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Brunsdon C., Fotheringham A. S., Charlton M. (2000) Geographically Weighted Regression as a Statistical Model. University of Newcastle-upon-Tyne: Newcastle upon Tyne, UK.

Bun M. J. G., Harrison T. D. (2019) OLS and IV Estimation of Regression Models Including Endogenous Interaction Terms. Econometric Reviews, 38(7), 814-827. https://doi.org/10.1080/07474938.2018.1427486

Case K. E., Shiller R. J. (2003) Is There a Bubble in the Housing Market? Brookings Papers on Economic Activity, 34(2), 299-362.

Chaim A., Łukasik N. (2024) Analiza korelacji w stosunkach międzynarodowych na przykładzie wybranych aspektów stosunków polsko-niemieckich. Krakowskie Studia Małopolskie, 3, 83-110

Charlton M., Fotheringham A. S. Brunsdon C. (2006) Geographically Weighted Regression: NCRM Methods Review Papers/NCRM/006. http://eprints.ncrm.ac.uk/90/ (dostęp: 12.01.2026)

Dąbrowski J. (2010) Zastosowanie wybranych metod statystycznych do analizy rynku nieruchomości. StatSoft Polska. http://www.statsoft.pl/Portals/0/Downloads/Zast_met_stat_analizy_rynku_nieruchomosci.pdf.

Fatih C. (2024) Lasso Regression Model [Technical Report]. University Ferhat Abbas of Sétif. https://doi.org/10.13140/RG.2.2.23949.14563

Foryś I. (2015) Indeks hedoniczny na wtórnym rynku mieszkań spółdzielczych na przykładzie wybranego osiedla w Szczecinie. Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania Uniwersytetu Szczecińskiego, 42(1), 149-159.

Huang Z., Lai, G (2023) A House Price Prediction Model Based on K-means Clustering and Random Forest in Guangzhou. Frontiers in Business, Economics and Management, 10(2), 377-381. https://doi.org/10.54097/fbem.v10i2.11077

Jamróz K. (2024) Apartment prices in Poland [zbiór danych]. Kaggle. https://www.kaggle.com/datasets/krzysztofjamroz/apartment-prices-in-poland

Leamer E. E. (2007) Housing IS the Business Cycle. NBER Working Paper 13428. https://doi.org/10.3386/w13428

Ligas M., Czaja, J. (2010) Zaawansowane metody analizy statystycznej rynku nieruchomości. Studia i Materiały Towarzystwa Naukowego Nieruchomości, 18(1), 7-19.

Matthews S. A., Yang T.-C. (2012) Mapping the Results of Local Statistics: Using Geographically Weighted Regression. Demographic Research, 26, 151-166. https://doi.org/10.4054/DemRes.2012.26.6

MacQueen J. B. (1967) Some Methods for Classification and Analysis of Multivariate Observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1: Statistics, University of California Press, Berkeley, 281-297. http://projecteuclid.org/euclid.bsmsp/1200512992.

Naji M. (2024) Real estate market analysis and prediction using machine learning. Master's thesis, Faculty of Science, Department of Applied Mathematics. https://www.esrilebanon.com/content/dam/esrisites/en-us/education/higher-education/Masters/Projects/2024/Naji/Thesis_Report_MohamadNaji.pdf

NBP (2024) Informacja o cenach mieszkań i sytuacji na rynku nieruchomości w II i IV kw.2024 r. https://nbp.pl/wp-content/uploads/2025/03/Informacja-o-cenach-mieszkan-w-IV-2024.pdf

NBP (2025) Informacja o cenach mieszkań i sytuacji na rynku nieruchomości mieszkaniowych i komercyjnych w Polsce w I kw. 2025 r. https://nbp.pl/wp-content/uploads/2025/06/Informacja-o-cenach-mieszkan-i-sytuacji-na-rynku-nieruchomosci-mieszkaniowych-i-komercyjnych-w-Polsce-w-I-kw.-2025-r.pdf?

Pagourtzi E., Assimakopoulos V., Hatzichristos T., French N. (2003) Real Estate Appraisal: A Review of Valuation Methods. Journal of Property Investment & Finance, 21(4), 383-401. https://doi.org/10.47772/IJRISS.2026.100300057

PIE – Polski Instytut Ekonomiczny (2024) Analiza rynku mieszkaniowego – IV kw. 2024 r. https://pie.net.pl/wp-content/uploads/2025/02/Rynek-mieszk-4_kwartal_2024.pdf

Rana R. K., Singhal R. (2015) Chi-square Test and its Application in Hypothesis Testing. Journal of the Practice of Cardiovascular Sciences, 1(1), 69-71. https://doi.org/10.4103/2395-5414.157577

Rosen S. (1974) Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy, 82(1), 34-55.

Sirmans G. S., Macpherson D. A., Zietz E. N. (2005) The Composition of Hedonic Pricing Models. Journal of Real Estate Literature, 13(1), 3-43.

Szelągowska A. (2023) Bezpieczny kredyt 2% oraz konto mieszkaniowe jako nowe instrumenty polityki mieszkaniowej w Polsce. Studia BAS, 4, 55-84. https://doi.org/10.31268/StudiaBAS.2023.30

Śleszyński Z. (2020) Wyznaczanie współczynników korelacji liniowej – podstawy. Wiadomości Statystyczne, 65(6), 69-87.

UKNF (2024) Informacja na temat sytuacji sektora bankowego w 2023 r. Pobrane z:
https://www.knf.gov.pl/knf/pl/komponenty/img/Informacja_na_temat_sytuacji_sektora_bankowego_w_2023_91099.pdf

ANALYSIS OF THE IMPACT OF LOCATION AND PROPERTY CHARACTERISTICS ON REAL ESTATE PRICES IN POLAND IN 2023–2024

Abstract: This article examines the impact of floor space, location, and amenities on the asking prices of apartments in Poland between 2023 and 2024, with a particular focus on Warsaw, Krakow, Lodz, and the Tricity. The study combines econometric methods and machine learning to identify key price-forming factors. The analysis employs OLS, Lasso, GWR, k-means, and Random Forest models. Based on a sample of 49,531 listings, the results demonstrate a clear superiority of non-linear models in price prediction. The

findings also confirm the dominant role of location and physical characteristics in explaining spatial price differentiation in the real estate market.

Keywords: real estate valuation, floor space, location, amenities, OLS, Lasso, GWR, k-means, Random Forest

JEL classification: R31, C53, C55, R12

QUANTITATIVE METHODS IN ECONOMICS
Volume XXVII, No. 1, 2026, pp. 55 – 65

Received: 01.03.2026
Accepted: 01.04.2026

A MULTIDIMENSIONAL QUANTITATIVE ASSESSMENT OF LABOR MARKET DYNAMICS IN TÜRKIYE: METHODOLOGICAL DIVERGENCES, HUMAN CAPITAL BOTTLENECKS, AND SPATIAL INEQUALITIES

Umut Kocabaş https://orcid.org/0009-0005-2128-4756
Faculty of Economics and Administrative Sciences,
Cukurova University, Adana, Türkiye
e-mail: umutkocabas01@gmail.com

Abstract: This study provides a multidimensional quantitative examination of the Turkish labor market. The primary objective is to evaluate the structural efficiency of employment by analyzing the statistical gap between official headline unemployment figures and broadly defined labor underutilization. Utilizing a descriptive statistical framework, the research analyzes 2024 cross-sectional data alongside historical time-series data sourced from the Turkish Statistical Institute (TÜİK) and the International Labour Organization (ILO). Empirical data reveals significant statistical divergences. While historical headline rates showed declines to 9.4% in 2023, the broad unemployment rate (labor slack) reached 26.7% in 2024, creating an 18.0 percentage point gap. Youth unemployment remains structurally high at 16.3%, with a persistent gender gap (22.3% for female youth vs. 13.1% for male youth). Spatial analysis identifies severe polarization, with unemployment peaking in the TRB2 region at 19.2%. Standard indicators mask structural fragilities such as informal employment and human capital mismatch. Effective mitigation necessitates data-driven Active Labor Market Policies (ALMPs) and targeted regional investments.

Keywords: Labor Economics, Quantitative Methods, Regional Disparities, Broad Unemployment, Türkiye

JEL classification: J21, J24, J64, R23

https://doi.org/10.22630/MIBE.2026.27.1.4

INTRODUCTION

Despite maintaining a dynamic and growing economy, the Turkish labor market exhibits significant structural vulnerabilities. Historically, macroeconomic assessments have primarily focused on official headline unemployment rates to gauge labor market performance. However, such aggregate indicators often mask underlying systemic fragilities, including structural underemployment, persistent youth unemployment, significant gender-based gaps, and severe regional polarization.

The primary limitation of relying on official headline unemployment figures is their strict adherence to a narrow classification of job search activity. To be officially counted as unemployed, an individual must have actively searched for work within the past four weeks and be ready to start within two weeks. This methodology naturally excludes discouraged workers—those who have abandoned job searches due to long-term failure—as well as time-related underemployed individuals. Consequently, official headline rates systematically diverge from the actual labor slack in the economy, creating a gap between lived experiences and official statistics.

This study investigates the structural complexities behind these figures. The novelty of this paper lies in its quantitative operationalization of theoretical labor market disparities. Unlike previous studies that evaluate the Turkish labor market solely through the lens of descriptive headline unemployment trends, this research contributes to the existing literature in three ways:

1. It integrates recent 2024 data to calculate precise quantitative "gap" indicators to measure absolute labor slack.
2. It bridges theoretical frameworks—such as the Dual Labor Market Theory and the Spatial Mismatch Hypothesis—with empirical regional data to explain persistent sub-national rigidities.
3. It contextualizes the informal economy's role as a structural buffer in emerging growth regimes.

To address these challenges, this study evaluates the labor market to answer the following research questions:

1. What is the precise quantitative gap between headline unemployment and broad labor underutilization in Türkiye?
2. Which specific demographic cohorts are most exposed to severe labor market exclusion?
3. Which NUTS-2 regions show the highest vulnerability and structural rigidity?

LITERATURE REVIEW

The dynamics of the Turkish labor market have been continuously shaped by globalization, structural transformations, and technological shifts. Understanding these forces requires a careful review of theoretical frameworks and empirical findings. As noted by McLaren (2017), globalization has introduced deep labor market shifts, creating new competitive dynamics and structural turbulence [DiPrete & Nonnemaker 1997]. In the digital era, Ross et al. [2024] emphasize that labor-augmenting technological changes can lead to systemic impacts, causing a displacement of low-skilled labor while increasing demand for higher skill sets, often resulting in critical skill shortages.

To understand the persistence of underutilization in emerging economies such as Türkiye, it is essential to evaluate the core dynamics of Dual Labor Market Theory. As conceptualized by Doeringer and Piore [1971], labor markets are bifurcated into primary (secure, high-wage) and secondary (informal, precarious) segments. In the Turkish context, the widening gap between headline unemployment and broad underutilization highlights a substantial structural displacement of the workforce into this secondary tier. This segmentation prevents the economy from achieving optimal equilibrium, trapping vulnerable cohorts in low-productivity cycles [Fields 2009].

These patterns manifest uniquely in Türkiye's developing market. Saglam and Gunalp [2012] utilized the Beveridge Curve to analyze Turkish labor dynamics, revealing structural shifts and profound inefficiencies in the job matching process. Furthermore, interventions such as minimum wage policies have had heterogeneous effects, disproportionately impacting informal workers and distinct regional areas [Işık et al. 2020]. Janta et al. [2015] highlight that structural economic changes frequently transform overall employment conditions, resulting in distinct vulnerabilities for peripheral demographic groups.

Demographic divides are rigorously debated in Turkish literature. Buğra and Yakut-Cakar [2010] established that the structural evolution of the social policy environment places significant institutional constraints on female employment. In a comprehensive assessment from 2000 to 2024, Tekgüç [2025] highlights that while urban female participation has risen, the broader employment rate remains structurally depressed. This paper builds on these foundations by moving beyond aggregate descriptive trends to explicitly measure these entrenched structural barriers.

MATERIAL AND METHODS

Data Sources, Frequency, and Time Scope

The empirical analysis relies on official aggregated data obtained from the Household Labour Force Survey (HLFS) conducted by the Turkish Statistical

Institute [TÜİK 2024]. These datasets are compiled monthly and aggregated annually using methodologies aligned with ILO standards.

To ensure analytical consistency and accurately capture structural rigidities, this study employs a defined, dual time-scope approach. The primary period of analysis is the contemporary post-pandemic recovery phase (2023–2025), utilizing cross-sectional data from 2024 to calculate current quantitative disparity indicators. Selective historical time-series data (1988–2011 for gender and 2013–2023 for aggregate trends) are integrated as critical contextual baselines to prove these disparities are deeply entrenched institutional features.

Statistical Framework and Quantitative Indicators

The study utilizes a descriptive and comparative statistical methodology. To systematize the analysis, four distinct mathematical indicators were formulated:

1. **Broad Underutilization Gap (Gap_{broad}):** Measures the absolute difference between the broad labor underutilization rate (UR_{broad}) and the official headline unemployment rate ($UR_{headline}$).

$$Gap_{broad} = UR_{broad} - UR_{headline}$$

2. **Youth Unemployment Gap (Gap_{youth}):** Measures the disproportionate burden on young workers.

$$Gap_{youth} = UR_{youth(15-24)} - UR_{total}$$

3. **Gender Unemployment Gap (Gap_{gender}):** Assesses demographic exclusion between female and male unemployment rates.

$$Gap_{gender} = UR_{female} - UR_{male}$$

4. **Regional Unemployment Gap ($Gap_{regional}$):** Evaluates spatial polarization between the highest and lowest NUTS-2 regional rates.

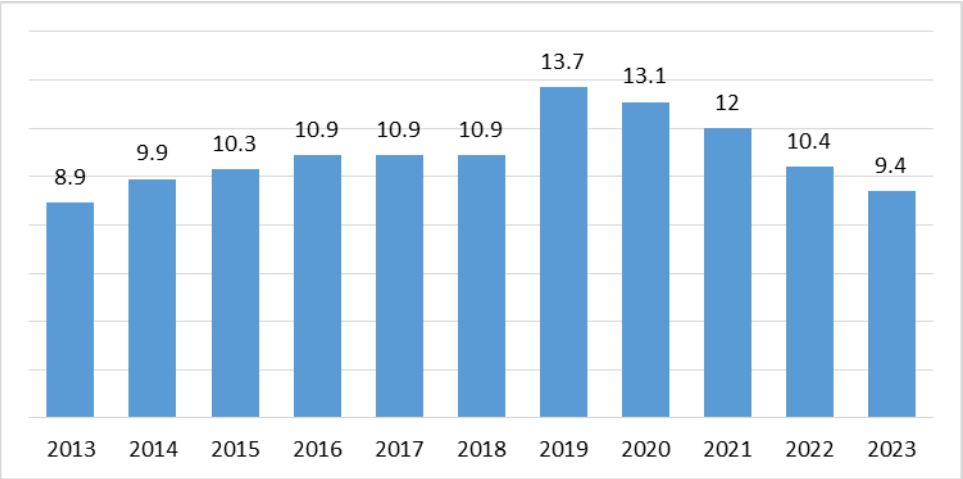
$$Gap_{regional} = UR_{max\_region} - UR_{min\_region}$$

RESULTS AND DISCUSSION

Statistical Divergences and Labor Underutilization

To properly contextualize the magnitude of labor slack, a historical macroeconomic assessment is required. As illustrated in Figure 1, the headline unemployment rate experienced a significant peak at 13.7% in 2019, reflecting the delayed impacts of severe currency shocks. Following this peak, the headline rate recorded a gradual, apparent decline to 9.4% in 2023.

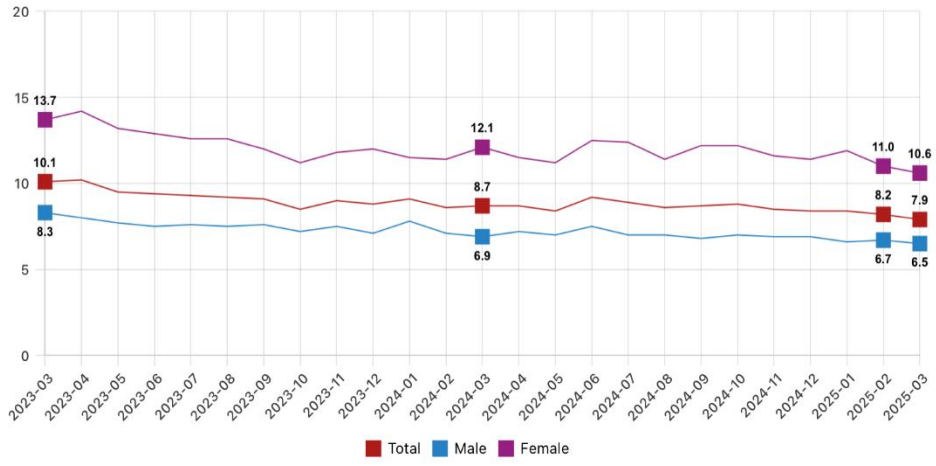
Figure 1. Headline Unemployment Rate (2013-2023) (%)



Source: own preparation

However, this apparent recovery in headline figures masks deeper inefficiencies and increasing broad underutilization. Focusing on the primary period of analysis, Figure 2 demonstrates the seasonally adjusted monthly rates from March 2023 to March 2025. Rather than a steady recovery, the line graph exposes persistent short-term monthly volatility. More importantly, the total headline rate consistently hovers strictly between the 8.5% and 9.5% threshold. This visually establishes a rigid "structural floor" for employment creation; despite periods of economic growth, the economy fundamentally struggles to absorb this baseline level of unemployment.

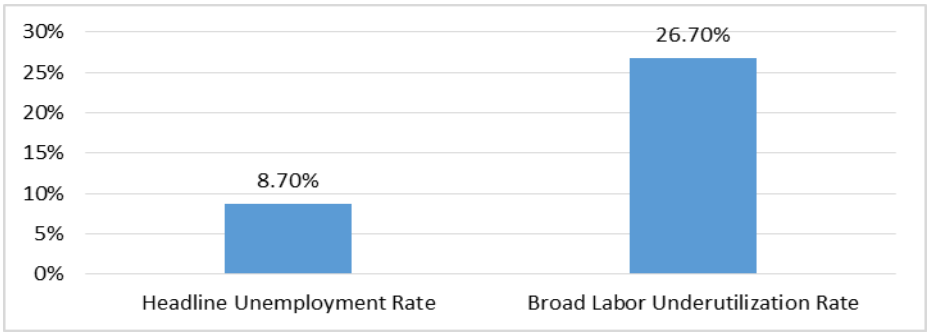
Figure 2. Seasonally adjusted unemployment rate, March 2023 - March 2025



Source: own preparation

The most critical analytical finding emerges when evaluating the broad labor underutilization indicator for 2024. As visually conceptualized in Figure 3, while the official headline rate rests at 8.7%, the Broad Labor Underutilization Rate towers at an alarming 26.7%. The resulting gap of 18.0 percentage points explicitly represents millions of discouraged and underemployed workers. This direct comparison demonstrates that a massive segment of the workforce is structurally locked out of primary employment and pushed into the secondary/informal sector (Boeri et al., 2019).

Figure 3. Comparison of Headline and Broad Labor Underutilization in Türkiye (2024)

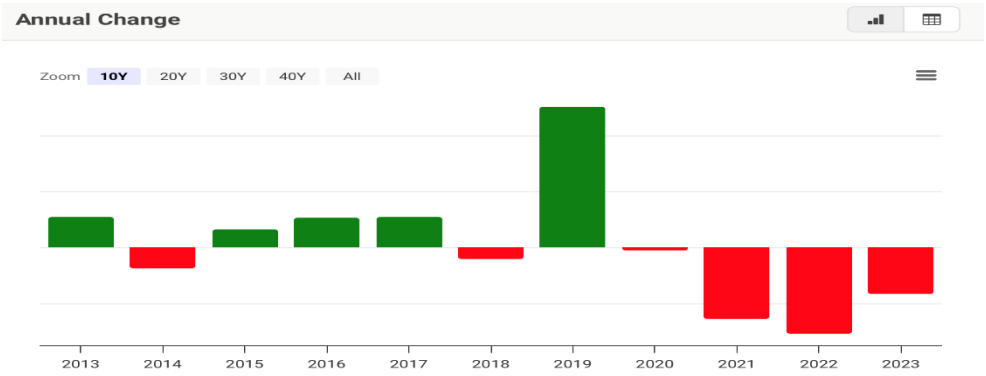


Source: own preparation

4.2. Demographic Disparities: Youth and Gender Gaps

The youth cohort (ages 15–24) exhibits extreme structural vulnerability. Figure 4 displays the annual change in youth unemployment, highlighting its hyper-sensitivity to macroeconomic business cycles. The severe amplitude of these shifts proves that young workers act as an economic "shock absorber". They endure a "last in, first out" dynamic, facing disproportionate job losses during economic downturns.

Figure 4. Annual Change in Youth Unemployment



Source: own preparation

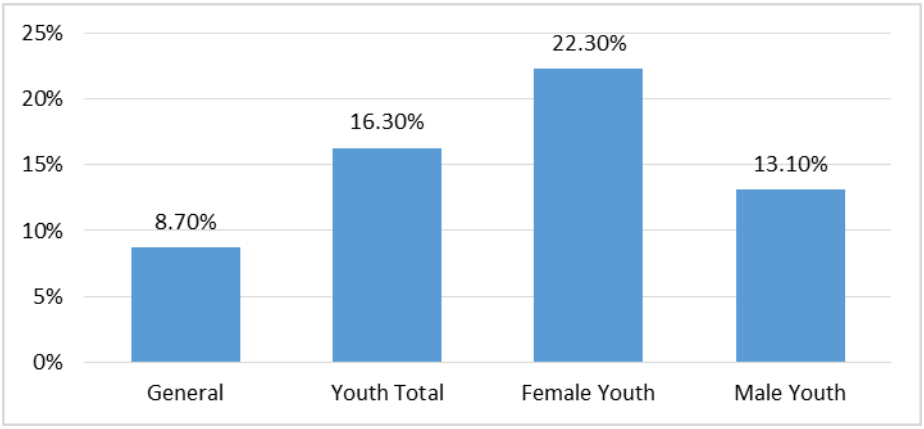
Table 1 complements this visual by detailing the absolute percentage of the youth labor force that remains unemployed. The tabulated data demonstrates that youth unemployment remains structurally elevated, peaking at 24.98% in 2019 and maintaining levels consistently above 15% even during recovery phases, such as the 17.61% recorded in 2023. Utilizing the cross-sectional data for 2024, the youth unemployment rate remains structurally high at 16.3%, ensuring that the youth gap remains at 7.6 percentage points above the national average.

Table 1. Turkey Youth Unemployment Rate (Percent of Total Labor Force)

2023	17.61%
2022	19.26%
2021	22.34%
2020	24.88%
2019	24.98%
2018	19.94%
2017	20.36%
2016	19.26%
2015	18.20%
2014	17.56%
2013	18.30%

Furthermore, the intersection of age and gender reveals profound, compounded exclusion. Figure 5 breaks down the 2024 demographic data, visually exposing this intersectionality. The female youth unemployment rate of 22.3% severely outpaces the male youth rate of 13.1%. Applying the specific gap methodology, this translates to a staggering gender gap of 9.2 percentage points within the youth cohort alone.

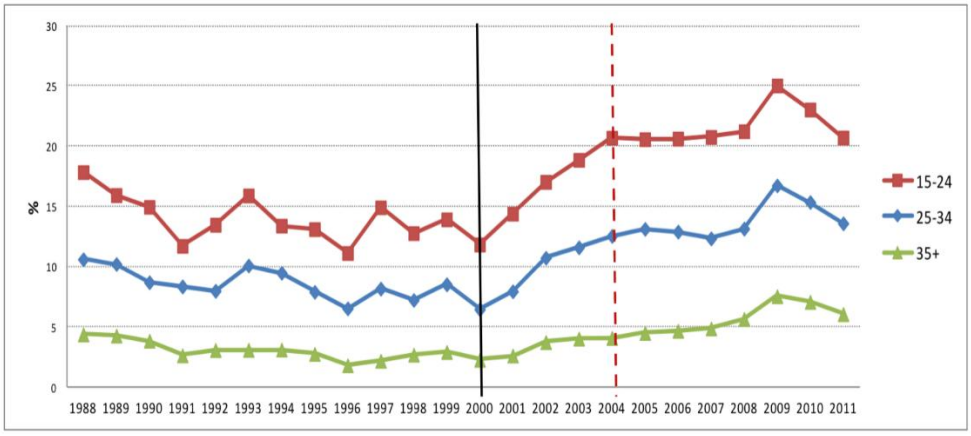
Figure 5. Demographic and Gender Disparities in Unemployment in Türkiye (2024)



Source: own preparation

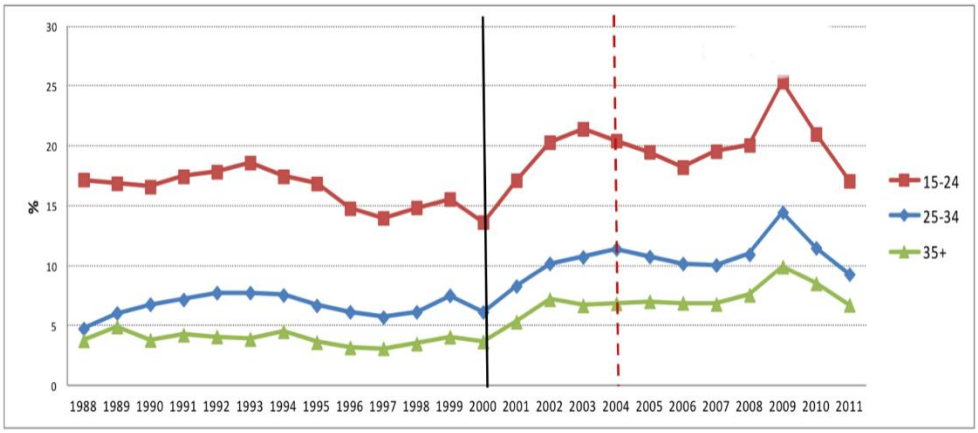
To confirm that this gap is an entrenched institutional characteristic rather than a temporary fluctuation, Figures 6 and 7 provide a necessary historical baseline from 1988 to 2011. Figure 6 shows that multi-decade female unemployment rates (especially for the 15-24 age group) have historically remained structurally high and highly resistant to periods of economic growth. This rigidity reflects long-standing institutional barriers and a persistent lack of formal labor market entry for women. Conversely, Figure 7 demonstrates that male age-based unemployment rates show a much higher alignment with macroeconomic business cycles, responding directly to periods of economic contraction and recovery.

Figure 6. Female Unemployment Rates by Age Group (1988-2011)



Source: Kaynak: HİA, 1988-2011

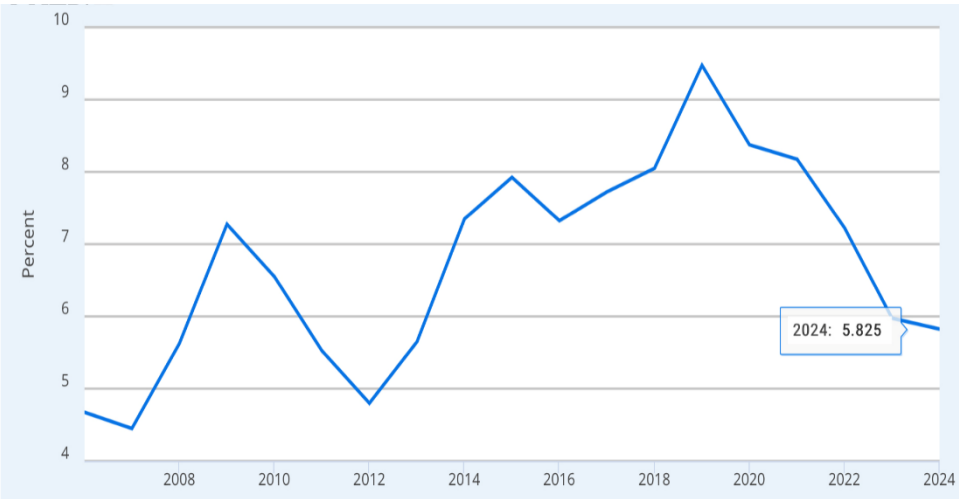
Figure 7. Male Unemployment Rates by Age Group (1988-2011)



Source: Kaynak: HİA, 1988-2011

In stark contrast, older demographics display entirely different structural trajectories. Figure 8 isolates the data for older males (ages 55-64) from 2008 to 2024. Unlike the youth cohort, this specific demographic shows a pronounced and unique decline, dropping to 5.8% in 2024. This sharp downward trend suggests prevalent early exits from the formal labor force, either through retirement patterns or unrecorded informal activities.

Figure 8. Unemployment Rate Male: From 55 to 64 Years for Turkiye

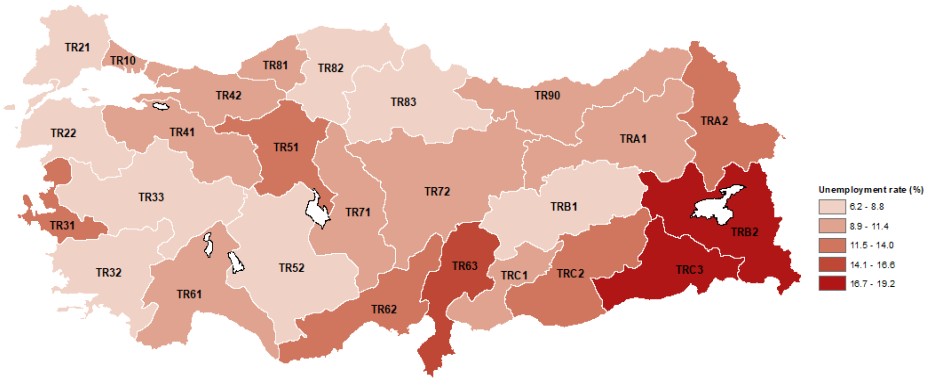


Source: Organization for Economic Co-operation and Development via FRED®

4.3. Spatial Polarization: Regional Divergences

Finally, spatial analysis of NUTS-2 regions exposes a severe dual-economy geographic structure. Figure 9 provides a geospatial heatmap of unemployment across Türkiye, clearly delineating an extreme "East-West" divide. The TRB2 region in the southeast recorded the highest absolute unemployment rate at 19.2%. This sharply contrasts with the heavily industrialized TR82 region in the north, which recorded a rate of 6.2%.

Figure 9. Regional Unemployment Disparities in Türkiye (NUTS-2 Breakdown, 2024)



Applying the regional indicator yields a gap of 13.0 percentage points, providing direct empirical proof of the Spatial Mismatch Hypothesis originally developed by Kain [1968]. As visually confirmed by Figure 9, industrial investments are heavily concentrated in the West, while eastern regions face structural isolation. Because internal labor mobility is constrained by high migration costs, peripheral regions remain perpetually locked into a low-employment equilibrium (Partridge & Rickman, 2006).

5. CONCLUSIONS AND POLICY IMPLICATIONS

This study systematically evaluated structural imbalances in the Turkish labor market. By calculating precise quantitative gaps, the empirical analysis answers the core research questions: the labor market suffers from an 18.0 pp broad underutilization gap, severe demographic exclusion (evidenced by a 9.2 pp youth gender gap), and extreme spatial polarization (a 13.0 pp regional gap). The findings confirm that conventional metrics consistently obscure underlying systemic vulnerabilities, particularly regarding the informal economy and skills mismatch. Effective structural mitigation requires:

- **Active Labor Market Policies (ALMPs):** Vocational training (İŞKUR, 2024) strictly aligned with private-sector technological demands to alleviate the human capital mismatch.
- **Regional Redistribution:** Aggressive place-based policies and fiscal incentives for isolated regions like TRB2 to halt human capital flight.
- **Institutional Support for Women:** Expanding public child care infrastructure is essential to overcome the socio-cultural barriers driving the gender unemployment gap.

LIMITATIONS AND FUTURE RESEARCH

This study relies on descriptive statistics and does not empirically test direct causal relationships. Future research must apply advanced panel data econometric models to regional labor market data in Türkiye to evaluate the causal impacts of macroeconomic shocks on localized employment.

REFERENCES

Boeri T., Garibaldi P., Ribeiro M. (2019) The Informal Sector, Explicit and Implicit Taxes, and Unemployment. Labour Economics, 58, 130-141.

Buğra A., Yakut-Cakar B. (2010) Structural Change, the Social Policy Environment and Female Employment in Turkey. Development and Change, 41(3), 517-538.

DiPrete T. A., Nonnemaker K. L. (1997) Structural Change, Labor Market Turbulence, and Labor Market Outcomes. American Sociological Review, 62(3), 386-404.

Doeringer P. B., Piore M. J. (1971) Internal Labor Markets and Manpower Analysis. M. E. Sharpe.

Fields G. S. (2009) Segmented Labor Market Models in Developing Countries. Oxford Development Studies, 37(4), 433-444.

Işık E., Orhangazi Ö., Tekgüç H. (2020) Heterogeneous Effects of Minimum Wage on Labor Market Outcomes: A Case Study from Turkey. IZA Journal of Labor Policy, 10(1), 1-24.

Janta B., Ratzmann N., Ghez J., Khodyakov D. (2015) Employment and the Changing Labour Market. RAND Corporation.

Kain J. F. (1968) Housing Segregation, Negro Employment, and Metropolitan Decentralization. The Quarterly Journal of Economics, 82(2), 175-197.

McLaren J. (2017) Globalization and Labor Market Dynamics. Annual Review of Economics, 9(1), 177-200.

Partridge M. D., Rickman D. S. (2006) The Geography of American Poverty: Is There a Role for Place-Based Policies?. W.E. Upjohn Institute for Employment Research.

Ross A. G., McGregor P. G., Swales J. K. (2024) Labour Market Dynamics in the Era of Technological Advancements. Technology in Society, 76, 102434.

Saglam B. B., Gunalp B. (2012) The Beveridge Curve and Labour Market Dynamics in Turkey. Applied Economics, 44(9), 1101-1111.

Tekgüç H. (2025) The Labor Market in Turkey, 2000–2024. IZA World of Labor, 450, 1-12.

Turkish Employment Agency (İŞKUR) (2024) Statistical Yearbook and Active Labor Market Policies Data. Ankara.

Turkish Statistical Institute (TÜİK) (2024) Labour Force Statistics, Regional Employment Data. Ankara, Türkiye.