Indeks 371262
e-ISSN 2657-9545
ISSN 0033-2372

Przegląd Statystyczny
Statistical Review

Vol. 72 No. 3 2025

GŁÓWNY URZĄD STATYSTYCZNY
STATISTICS POLAND

INFORMATION FOR AUTHORS

Przegląd Statystyczny. Statistical Review publishes original research papers on theoretical and empirical topics in statistics, econometrics, mathematical economics, operational research, decision science and data analysis. The manuscripts considered for publication should significantly contribute to the theoretical aspects of the aforementioned fields or shed new light on the practical applications of these aspects. Manuscripts reporting important results of research projects are particularly welcome. Review papers, shorter papers reporting on major conferences in the field, and reviews of seminal monographs are eligible for submission based on the Editor-in-Chief's decision.

Since 1st May 2019, the journal has been publishing articles in English.

Any spelling style is acceptable as long as it is consistent within the manuscript.

All works should be submitted to the journal through the Editorial System
(https://www.editorialsystem.com/pst).

For details of the submission process and editorial requirements, please visit
https://ps.stat.gov.pl/ForAuthors.

Przegląd Statystyczny
Statistical Review

Vol. 72 No. 3 2025

EDITORIAL OFFICE ADDRESS

Statistics Poland (GUS), Al. Niepodległości 208, 00-925 Warsaw, Poland

Language editing: Scientific Publications Division, Analyses and Information Dissemination Department, Statistics Poland
Technical editing and typesetting: Statistical Publishing Establishment, Statistical Computing Center – team supervised by Mariusz Męcina

Centrum Informatyki
Statystycznej

Printed and bound by: Statistical Computing Center
Al. Niepodległości 208, 00-925 Warsaw, Poland, cis.stat.gov.pl

Website: ps.stat.gov.pl

ISSN 0033-2372
e-ISSN 2657-9545
Index 371262

Information on the sales of the journal: Statistical Computing Center
Phone no.: +48 22 608 32 10, +48 22 608 38 10

Order no. 382/2025

CONTENTS

Przegląd Statystyczny. Statistical Review, 2025, vol. 72, 3, 1–27 https://doi.org/10.59139/ps.2025.03.1
Received: 14.09.2025, revised: 23.11.2025, accepted: 28.11.2025

Explaining regional wage disparities with machine learning: A SHAP-based interpretation approach

Andrzej Dudek,[a] Marcin Pełka,[b] Artur Skiba[c]

Abstract. The aim of the study is to provide an explanation for the factors that most influence the differences in wage levels between Polish powiats (equivalent to counties). This study investigates regional wage disparities in Poland by applying machine learning models enhanced by Explanatory Model Analysis techniques. Using powiat-level data from the Local Data Bank (Pol. Bank Danych Lokalnych – BDL) for 2010 and 2023, a neural network framework was developed to predict wage levels based on economic, demographic, infrastructural and environmental variables. To interpret the model, we employed the Variable Importance over Permutation (VIP) and SHapley Additive exPlanations (SHAP) approaches, which provide insights into both the global feature importance and the local contributions of individual variables. The results indicate that the share of the productive population, unemployment rates and social vulnerability remain key determinants of wage differences, although their relative influence shifts significantly over time. The SHAP analysis demonstrates how regional contexts such as the Jelenia Góra and Wrocław powiats exhibit distinct factor dynamics, with demographic and infrastructural variables playing varying roles across the studied years. The findings highlight the potential of combining machine learning with explainability methods to uncover complex, nonlinear determinants of wages, offering a more transparent analytical basis for understanding evolving regional disparities.

Keywords: deep learning, machine learning, explanatory model analysis, wage disparities
JEL: C15, C45, O150

1. Introduction

Regional wage disparities remain a central topic in labor economics, often explained by the differences in human capital endowments, sectoral structures, and spatial inequalities (Combes et al., 2008; Moretti, 2011). Traditional econometric models have been widely used to quantify these disparities, yet they frequently rely on restrictive assumptions that may not capture complex, nonlinear interactions between the explanatory factors. Recent advances in machine learning provide a powerful alternative by enabling predictive modeling that accommodates high-dimensional and

[a] Wroclaw University of Economics and Business, Faculty of Economics and Finance, Department of Financial Investments and Risk Management, ul. Komandorska 118/120, 53–345 Wrocław, Poland, e-mail: andrzej.dudek@ue.wroc.pl, ORCID: https://orcid.org/0000-0002-4943-8703.

[b] Wroclaw University of Economics and Business Branch in Jelenia Góra, Faculty of Economics and Finance, Department of Econometrics and Computer Science, ul. Nowowiejska 3, 58–500 Jelenia Góra, Poland, e-mail: marcin.pelka@ue.wroc.pl, ORCID: https://orcid.org/0000-0002-2225-5229.

[c] Polish Information Processing Society, ul. Solec 38 lok. 103, 00–394 Warszawa, Poland, e-mail: artur.skiba70@gmail.com, ORCID: https://orcid.org/0000-0003-4616-7271.

interdependent features without imposing strong functional form restrictions (Mullainathan & Spiess, 2017). However, the opacity of machine learning methods has raised concerns about interpretability, especially in policy-relevant domains such as labor markets, where transparent explanations are crucial.

To address this challenge, methods of Explainable AI (XAI) like SHapley Additive exPlanations (SHAP) have emerged as a robust framework for interpreting complex machine learning models by attributing feature importance based on the principles of the cooperative game theory (Lundberg & Lee, 2017; Masís, 2023; Molnar, 2020). Applying SHAP to wage prediction models allows for a granular understanding of how regional characteristics such as industrial composition, education levels, or urbanization contribute to the observed wage gaps. This approach bridges predictive performance with interpretability, enabling researchers and policymakers to identify the factors that matter most and see how their effects vary across regions. By combining Machine Learning with a SHAP-based interpretation, the analysis of regional wage disparities can advance beyond aggregate statistical associations toward more actionable, fine-grained insights.

2. Analysis of an explanatory model for studies on wage differences

2.1. Analysis of wage differences: A literature review

Recent literature offers numerous analyses of spatial wage differentials across various territorial levels, including Polish powiats and voivodships (equivalent to counties and provinces, respectively), and Ukrainian oblasts (equivalent to provinces) (Adamczyk et al., 2009; Bolińska & Gomółka, 2018; Dykas et al., 2020; Dykas & Misiak, 2013; Kapela & Kwiatkowski, 2023; Przekota, 2016). Theoretical frameworks typically rely on efficiency wage models. Empirical studies use such indicators as wages, labor productivity, and unemployment rates to estimate wage determinants via regression analysis. Beyond basic metrics, newer models such as those by Kapela and Kwiatkowski (2023) incorporate variables like higher education rates, technological innovation, and patent activity, while also addressing the effects of the 2020 pandemic. The applied methods include least squares, the generalized method of moments, clustering methods, and fixed effects models, which enhance the accuracy of the results. Findings show that factors like proximity to large cities, labor productivity, and human capital play crucial roles in wage disparities, while results regarding capital expenditures and industry output remain ambiguous (Adamczyk et al., 2009; Przekota, 2016).

Wage elasticity relative to unemployment remains a central theme. Many studies confirm that a negative relationship between the two exists, as seen in an earlier work by Phillips (1958) and later by Kaliski (1964), Blanchflower and Oswald (1990), though exceptions occur, such as in South Africa (Kingdon & Knight, 2006) and in some Polish powiat-level fixed-effects models (Dykas & Misiak, 2013). Modern applications of the Phillips curve continue to show relevance in different national contexts (Bartosik & Mycielski, 2015; Machuca & Cota, 2017). Other important aspects include the growing role of education, innovation, and demographic shifts in explaining wage variation (Combes et al., 2008; Kapela & Kwiatkowski, 2023). Despite the robust research at higher administrative levels, recent powiat-level studies are scarce, with the latest comprehensive analyses dating back to 2014 (Dykas & Misiak, 2013). Consequently, a renewed need emerged to reassess spatial wage dynamics at the powiat level, particularly in light of the post-pandemic developments and ongoing socio-economic changes (c.f. Luśtyk et al., 2024).

2.2. Methods of explanatory model analysis

To address the challenges described in the previous section, newly arisen methods of Explanatory Model Analysis/XAI (see Biecek & Burzykowski, 2021; Masís, 2023; Molnar, 2020), particularly through Variable Importance over Permutation (VIP) and SHAP values, offer significant advantages in analyzing economic phenomena.

VIP enables researchers to assess the relative impact of each predictor by measuring the change in model performance after randomly permuting individual variables. This model-agnostic method provides an intuitive ranking of features, highlighting the most influential economic indicators driving predictive accuracy. It supports a transparent, reproducible evaluation of variable relevance, which is essential for policy analysis and decision-making in complex economic systems.

SHAP values further enhance the explanatory power by attributing prediction contributions to individual features in a theoretically grounded manner based on the cooperative game theory. Unlike aggregate importance scores, SHAP delivers local explanations for each prediction, allowing analysts to understand heterogeneity across economic agents or regions. This granularity is particularly valuable for exploring non-linear interactions and dependencies commonly present in econometric models. Together, VIP and SHAP form a robust framework for interpreting black-box machine learning models, facilitating deeper insights into causal mechanisms and improving the credibility of data-driven economic policy recommendations.

Other methods that make explaining black box models possible are partial dependence plots (PDP), which show the marginal effect that one or two variables (features) have on the predicted outcome (Friedman, 2001; Greenwell et al., 2018). PDPs capture only the main effect of the feature and ignore the possible interactions, so it should be used with care.

Accumulated local effects (ALE) plots describe how variables influence the prediction. Moreover, ALE plots are faster than PDPs (Apley & Zhu, 2020). In the ALEs, however, an interpretation of the effect across intervals is not permissible if the features are strongly correlated. ALE effects may differ from coefficients specified in linear regression models when variables interact and are correlated. What is more, ALE plots are not accompanied by Individual Conditional Explanation (ICE) curves and can have many small ups and downs. In this case, when we reduce the number of variables, we not only make the estimates more stable, but also smooth out the complexity of the model.

A feature interaction model based on Friedman's H statistic (Friedman & Popescu, 2008) and variable interaction networks (Hooker, 2004) allow variable interactions to be taken into account in the predictions.

Another way to interpret variable importance is through functional decomposition. It can be done by: functional Analysis of Variance (ANOVA) (Hooker, 2004), generalized functional ANOVA for dependent variables (features) (Hooker, 2007), generalized additive regression modes, or ALE plots.

The permutation feature importance algorithm based on Fisher et al. (2019) measures the increase in the prediction error of the model after the variable's values are permuted, which breaks the relationship between the variable and the known (true) outcome.

The global surrogate model is another interpretable model that is trained to approximate the prediction of a black box model. The surrogate model uses a much simpler model instead of a complex one (Molnar, 2020).

The local interpretable model-agnostic explanations (LIME) is a technique that approximates any black box machine learning model with a local, interpretable model to explain each individual prediction that is described in the paper by Ribeiro et al. (2016). The main idea is that we perturb (change) the original data points, feed them into the black box model, and then observe the corresponding outcomes. Then the method weighs those new data points as a function of their proximity to the original point. Ultimately, using those sample weights, LIME fits a surrogate model, such as linear regression, on the dataset with the variations. Each original data point can then be explained with the newly trained explanation model.

2.3. VIP and SHAP methods for model explanation

Permutation-based methods like VIP, originating from the idea introduced by Breiman (2001), provide a model-agnostic approach to estimating variable importance. This is done by assessing the impact of controlled perturbations in the input data on the predictive performance. Instead of relying on the internal structure of a model, this technique treats the model as a black box and evaluates how the quality of a prediction changes when the values of a given variable are deliberately disrupted.If the variable contributes substantially to the model's predictive mechanism, permuting its values should lead to a notable decline in performance. In contrast, if the variable has little or no influence, prediction quality should remain to a large extent unaffected.

The change in performance – measured through metrics such as mean squared error, accuracy, or alternative loss functions – serves as an inverse proxy for variable importance. A larger degradation in predictive quality implies a higher significance of the variable in the decision-making process. In practice, this procedure is implemented by randomly permuting the values of a selected feature across observations in the dataset and re-evaluating the model's output. Repeating this process for each variable provides a systematic and interpretable measure of feature importance that is independent of the model specification.

This process involves what follows.

Let:

X – a dataset with m explanatory variables and n instances (objects),
Y – column vector of the observed values of the dependent variable,
$\hat{Y}$ – column vector of the predicted values of the dependent variable,
$P(\hat{Y}, X, Y)$ – performance metrics (loss function) for the model.

The procedure then involves the following steps:

1. Training the model;
2. Computing $p_0 = P^0(\hat{Y}, X, Y)$, i.e. the initial value of the loss function;
3. Shuffling (permuting) column vector X_k for given $1 < k < m$. Matrix X after permutation becomes $X^{(*k)}$;
4. Computing model predictions $\hat{Y}^{*k}$ for X^{*k};
5. Computing $p_{*k} = P(\hat{Y}^{*k}, X^{*k}, Y)$;
6. Estimating the importance for variable k in the process of prediction through $vip_k = p_{*k} - p_0$ (alternatively used in the $vip_k = \frac{p_{*k}}{p_0}$ form).

The Shapley values, another technique of Explanatory Model Analysis, originating from the cooperative game theory, provide a rigorous framework for quantifying the joint contribution of explanatory variables to model predictions. In Shapley's (1953) original formulation, the method determined each player's marginal contribution to the overall payoff obtained by a coalition. Transposed into model interpretation, the 'players' are the variables, and the 'payoff' corresponds to the model's prediction. Thus, Shapley values measure how the estimated outcome changes when a specific variable is added to the different subsets of predictors involved in generating the prediction.

The final attribution is obtained as a weighted average of these marginal contributions across all possible subsets. The weighting scheme depends on the size of the subsets: variables added to very small or nearly complete subsets receive higher weights, whereas those added to medium-sized subsets are assigned lower weights. This ensures fairness in attributing contributions across all possible coalitions of variables. The resulting SHAP provides a consistent and theoretically grounded measure of variable importance at both the global (model-wide) and local (instance-specific) levels.

The algorithm for finding the SHAP values for a certain object explained and a certain variable may be stated as follows.

Let:

X – dataset with m explanatory variables and n instances (objects);

Y – column vector of the observed values of the dependent variable;

$\hat{Y}$ – column vector of the predicted values of the dependent variable;

l – object (instance) index for which the analysis is conducted;

k – feature (variable) index for which the analysis is conducted.

The procedure then involves the following steps:

1. Training the model;
2. Calculating $\widehat{Y_0} = \frac{\sum_{i=1}^{n} \widehat{Y_i}}{n}$, i.e. the average prediction value over the dataset (and initial explanation estimation);
3. Let:

$$V_{-k} = \{1,2,\dots,\mathrm{m}\}\backslash\{\mathrm{k}\} \tag{1}$$

(The set of all variable indices with k excluded);

4. For each s in *0,1, …, m-1*;
5. For all subsets S of V_{-k} of size s, calculating:
   - $\widehat{(Y_l)}^{*S}$ average prediction for the dataset for which variables' $X_i\colon i \in S$ values in the whole dataset are set to the values of object X_l;
   - $\widehat{(Y_l)}^{*S\cup\{k\}}$ be the average prediction for the dataset for which variables' $X_i\colon i \in S$ and variable's X_k values in the whole dataset are set to the values of object X_l;

and the Shapley value:

$$SHAP_S = \frac{s!\cdot(m-s-1)!}{m!}\left(\widehat{Y_l}^{*S\cup\{k\}} - \widehat{Y_l}^{*S}\right); \tag{2}$$

6. Summing all the $SHAP_S$ values.

The SHAP method was originally introduced by Štrumbelj and Kononenko (2010, 2014) and later popularized by Lundberg and Lee (2017). Its widespread application stems from a solid theoretical foundation and the reliability of its explanatory power.

3. Factors determining wage disparities. Research based on data from the Local Data Bank for 2010 and 2023

The analysis has been conducted using data that describe the average compensation level in Polish powiats in the years 2010 and 2023. The data were acquired directly from the Local Data Bank (Pol. Bank Danych Lokalnych – BDL), which is Statistics Poland's official repository, through webservices, and contained variables which describe economic (labor market), sociological, demographical, infrastructural and environmental phenomena. The description of dependent and exogenous variables along with BDL identifiers is presented in the Table.

Table. Description of variables used in the research

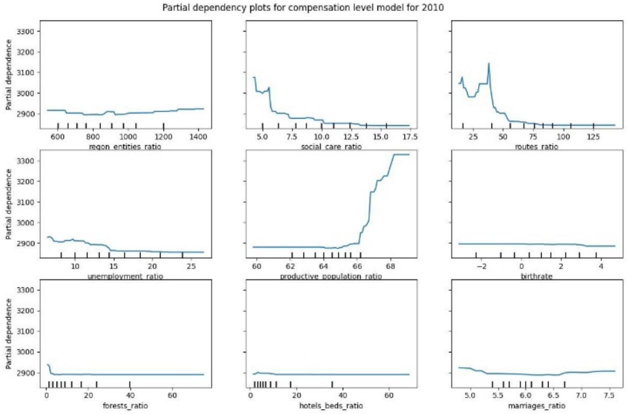
Variable ID	Internal Name	Type of variable	Description (English)
64428	compensation_level	Dependent variable (economic)	Average gross monthly wages in PLN
60530	regon_entities_ratio	Labor market	Business entities with registered REGON per 10,000 population
60270	unemployment_ratio	Labor market	Registered unemployment rate (overall)
458700	social_care_ratio	Sociological	Beneficiaries of social assistance by place of residence as the percentage of the total population
60566	productive_population_ratio	Demographical	The percentage share of the working-age population in the total population
450551	birthrate	Demographical	Natural increase (births minus deaths) per 1,000 population
450543	marriages_ratio	Demographical	Marriages per 1,000 population
60300	hotels_beds_ratio	Touristic	Bed places per 1,000 population
395404	routes_ratio	Infrastructural	Gmina (Polish equivalent to municipality) and powiat hard surface roads in km per 10,000 population
1646059	forests_ratio	Environmental	Municipal forest area in m2 *per capita*

Source: Local Data Bank (https://bdl.stat.gov.pl).

To find the most influenced factors for wages level modelling, we have built the eXtreme Gradient Boosting (Chen & Guestrin, 2016) model based on 319 objects describing powiats. The distinct models have been built for both studied years.

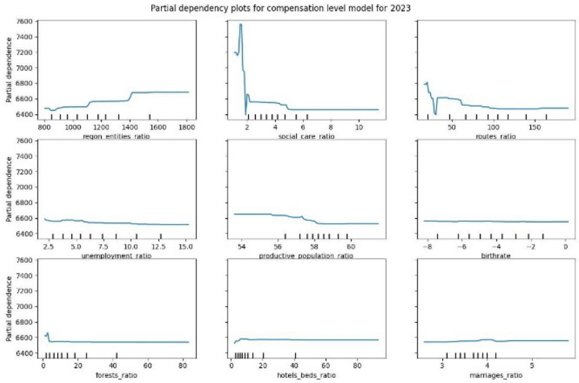
The python code implementing this procedure is included in Appendix 1. The full text results are presented in Appendix 2. The partial dependency plots presented in Figures 1 and 2 demonstrate that both models' convergence is stable.

Figure 1. Partial dependency plots for explanatory variables for wage levels in powiats in the 2010 model



Source: authors' calculations (the code is presented in Appendix 1).

Figure 2. Partial dependency plots for explanatory variables for wage levels in powiats in the 2023 model

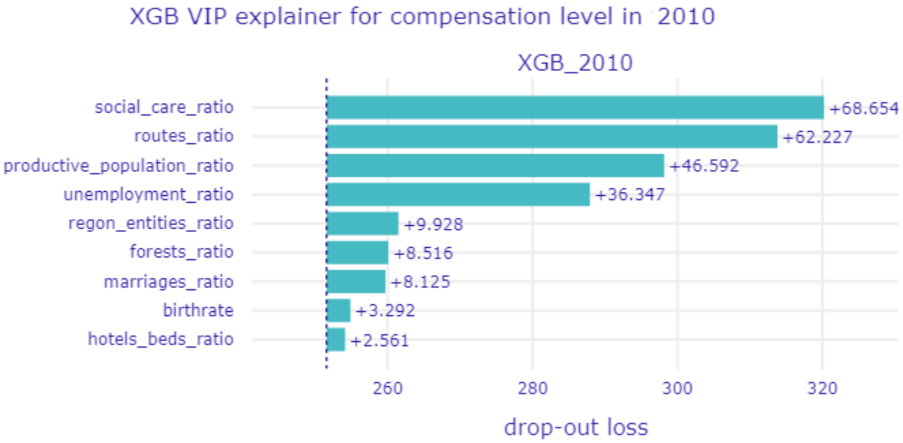


Source: authors' calculations (the code is presented in Appendix 1).

The model shows solid learning on training data ($R^2 = 0.626$). The test performance is positive and reasonable ($R^2 = 0.290$), indicating it captures useful predictive relationships. The gap between 0.626 and 0.290 suggests some degree of overfitting, but not severe, which is typical and acceptable for many socioeconomic datasets. The model generalizes moderately well and is reliable enough to proceed with interpretation (VIP, SHAP).

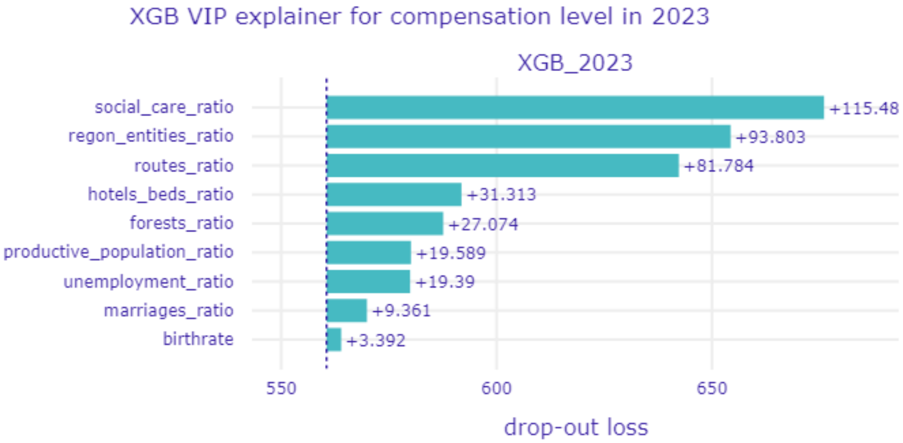
The VIP method is used to evaluate the influence of explanatory variables on the explained phenomena (wage level in powiats). The results for the models for 2010 and 2023 are presented in Figures 3 and 4.

Figure 3. Variable importance plot for exogenous variables for wage levels in powiats in 2010



Source: authors' calculations (the code is presented in Appendix 2).

Figure 4. Variable importance plot for exogenous variables for wage levels in powiats in 2023



Source: authors' calculations (the code is presented in Appendix 1).

The VIP results for 2010 indicate that the most influential variable is the *social_care_ratio*, with the highest dropout loss equal to 320.23. This means that removing the variable causes the strongest deterioration in model performance, suggesting that the social-assistance burden was a key structural determinant of compensation levels in 2010. The next highly influential variables are the *routes_ratio* (313.80) and *productive_population_ratio* (298.16), both of which significantly worsen prediction when excluded, showing that transportation accessibility and the working-age population share are critical factors.

Further in the ranking, variables such as the *unemployment_ratio* (287.92), *regon_entities_ratio* (261.50), and *forests_ratio* (260.09) still contribute substantially to model accuracy, but their influence is more moderate. Their dropout losses imply that labor-market structure, business density, and environmental context affect compensation prediction, but to a lesser degree than factors related to social services and transport. These mid-ranked variables form a secondary explanatory layer that stabilizes the model.

At the lower end of the importance distribution, the predictors with the smallest dropout losses, namely the *marriages_ratio* (259.70), *birthrate* (254.86), and the *hotels_beds_ratio* (254.13) exerted the least influence in 2010. Removing them increases error only slightly, suggesting they contain comparatively limited independent information for determining compensation differences. In this year, demographic and tourism indicators appear marginal relative to the socioeconomic structure and accessibility.

The VIP analysis of the 2023 wage prediction model for Polish powiats highlights the relative strength of diverse structural, demographic, and environmental determinants.

In 2023, the variable importance structure shifts noticeably, with the *social_care_ratio* again emerging as the most influential predictor. This time, it shows an even higher dropout loss of 676.01, making it the dominant factor in the model. The next influential variables are the *regon_entities_ratio* (654.33) and *routes_ratio* (642.31), both showing large performance drops when removed. This highlights the growing importance of business density and transportation infrastructure for explaining compensation levels in 2023.

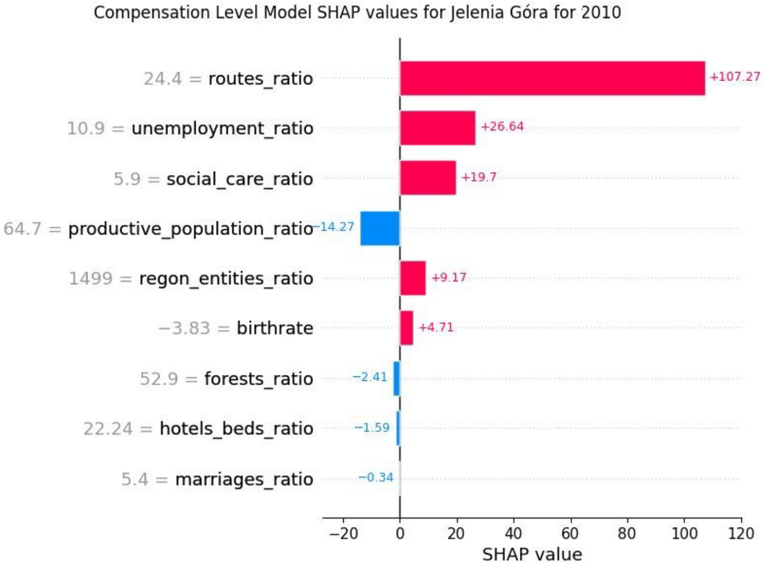
The middle tier of variables, including the *hotels_beds_ratio* (591.84), *forests_ratio* (587.60), and *productive_population_ratio* (580.12) also carry substantial explanatory weight. Their dropout losses show that tourism capacity, environmental features, and demographic composition meaningfully support model predictions. Compared to 2010, these secondary predictors become more informative, suggesting a more complex structure of the determinants.

The least influential predictors are the *unemployment_ratio* (579.92), *marriages_ratio* (569.89), and *birthrate* (563.92), whose dropout losses are closer to the full model but nevertheless in the lower range of importance. Although still impactful, the demographic and labor-market indicators exert smaller marginal effects compared to structural and institutional features. The 2023 importance pattern therefore portrays a landscape where social-service load, enterprise density, and infrastructure dominate compensation prediction, while demographic variables play a supportive yet reduced role.

The explanatory model analysis method allows a deeper insight into factors determining the analyzed phenomenon (compensation level). The analysis covers not only the general model explanation but also most influential factors in individual cases.

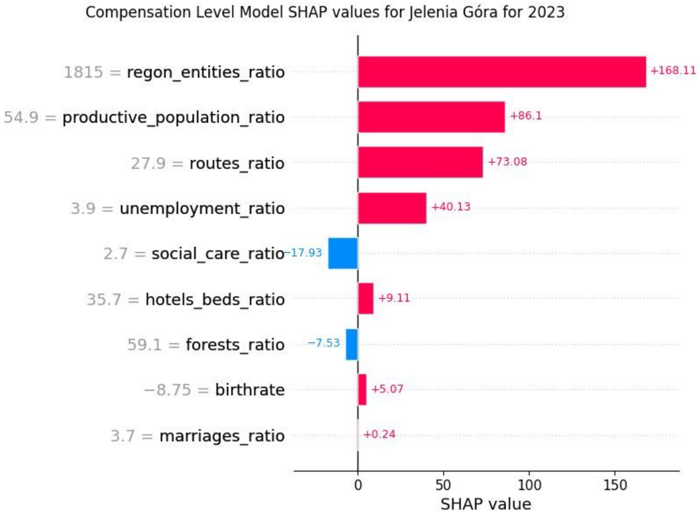
To better understand the influence of the given phenomena on overall compensation differences at the local level, a SHAP analysis is conducted. The SHAP values for the 2010 and 2023 models for the Jelenia Góra powiat are presented in Figures 5 and 6.

Figure 5. SHAP plot for exogenous variables for wage levels in the Jelenia Góra powiat in 2010



Source: authors' calculations (the code is presented in Appendix 1).

Figure 6. SHAP plot for exogenous variables for the wage levels in the Jelenia Góra powiat in 2023



Source: authors' calculations (the code is presented in Appendix 2).

For the Jelenia Góra powiat, in 2010, the strongest SHAP contributor was the *routes_ratio*, with a positive effect of 107.27 at a value of 24.40. This highlights the powiat's relative transport accessibility as a major factor supporting its compensation prediction. The next significant variables are the *unemployment_ratio* (+26.64 at 10.90) and *social_care_ratio* (+19.70 at 5.90), indicating that despite relatively high unemployment and social-care indicators, these conditions still contribute positively within the model structure.

Negative contributions also proved to play an essential role. The *productive_population_ratio* (–14.27 at 64.70) pulls the prediction downward, suggesting demographic or economic strain associated with the powiat's working-age population share. The *forests_ratio* (–2.41), *hotels_beds_ratio* (–1.59), and *marriages_ratio* (–0.34) also reduce the prediction slightly, implying that environmental and tourism indicators contribute less positively for Jelenia Góra compared to other powiats.

A few variables exert small positive influences. The *regon_entities_ratio* (+9.17 at 1,499) and *birthrate* (+4.71 at –3.83) add a marginal upward pressure on salaries. The overall SHAP structure for 2010 reflects a mix of strong transport infrastructure effects and modest socioeconomic constraints, with demographic features moderating the powiat's predicted compensation level.

For Jelenia Góra in 2023, the *regon_entities_ratio* became the strongest positive contributor, with a SHAP value of +168.11 at 1,815 entities. This signals the increasing importance of local business density for salary levels. The *productive_population_ratio* (+86.10 at 54.90) and *routes_ratio* (+73.08 at 27.90) also strongly elevate the prediction, with transportation accessibility remaining a key structural advantage.

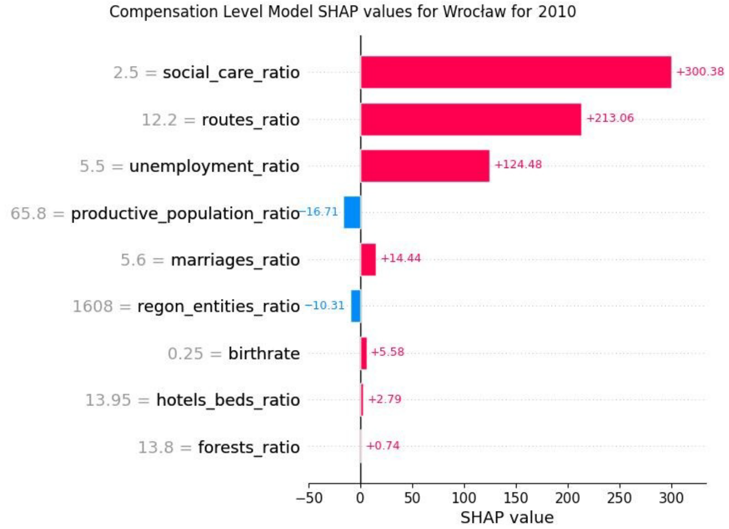
Additional positive contributions derive from the *unemployment_ratio* (+40.13 at 3.90) and *hotels_beds_ratio* (+9.11 at 35.70), indicating that tourism infrastructure played a more supportive role in 2023 than in 2010. Meanwhile, the *social_care_ratio* shows a negative impact (−17.93), which suggests an increasing sensitivity of the model to social-assistance burdens. The *forests_ratio* also contributes negatively (−7.52), moderating the positive effects of other variables.

Smaller contributions come from *birthrate* (+5.07) and the *marriages_ratio* (+0.24), which have a limited influence. Overall, the SHAP profile for 2023 indicates that Jelenia Góra's salary structure is shaped by a combination of economic density, demographic composition, and improved labor-market indicators, with structural accessibility continuing to reinforce compensation predictions.

For Jelenia Góra, the SHAP comparison between 2010 and 2023 shows a clear shift in the structure of factors influencing compensation levels. In 2010, the main positive driver was the *routes_ratio* (+107.27 at 24.40), supported by the *unemployment_ratio* (+26.64) and *social_care_ratio* (+19.70), while the *productive_population_ratio* (−14.27) exerted a negative influence and the remaining variables had only small effects. In 2023, however, the leading factor becomes the *regon_entities_ratio* (+168.11 at 1815), accompanied by strong positive contributions from the *productive_population_ratio* (+86.10) and *routes_ratio* (+73.08). This indicates a transition from an 'infrastructure-driven' model to a more 'economic-demographic' one. The role of the *social_care_ratio* also changes, from a small positive effect in 2010 (+19.70) to a clearly negative effect in 2023 (−17.93), suggesting the model became more sensitive to social-assistance burdens.

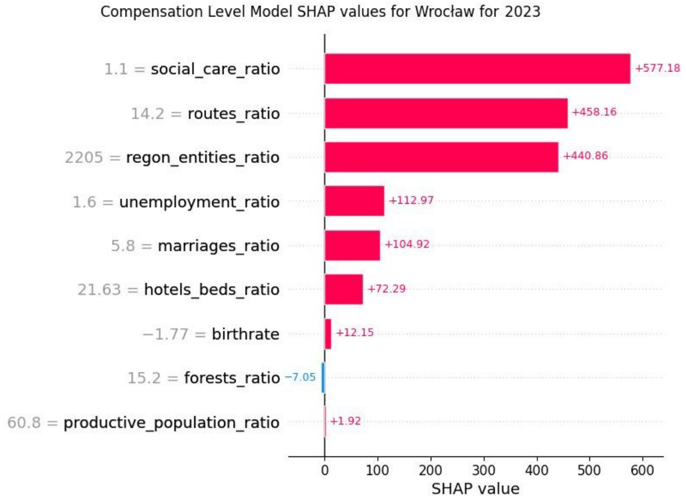
Jelenia Góra is a representative of medium-sized powiats. To broaden the analysis, the SHAP values have been estimated for a representative of larger powiats, like the Wrocław powiat with results presented in Figures 7 and 8.

Figure 7. SHAP plot for exogenous variables for wage levels in the Wrocław powiat in 2010



Source: authors' calculations (the code is presented in Appendix 1).

Figure 8. SHAP plot for exogenous variables for the wage levels in the Wrocław powiat in 2023



Source: authors' calculations (the code is presented in Appendix 1).

In the Wrocław powiat (2010), the SHAP analysis highlights the *social_care_ratio* as the dominant positive driver, contributing 300.38 units to the prediction at a value of 2.50. This indicates that Wrocław's low social-care burden is interpreted by the model as strongly favorable for compensation levels. Similarly, the *routes_ratio*

(SHAP = 213.06, value = 12.20) exerts a substantial positive impact, reflecting Wrocław's well-developed transport networks.

Another strong contributor is the *unemployment_ratio*, adding 124.48 units at a relatively low level of 5.50, suggesting that lower unemployment aligns with higher predicted salaries. The *marriages_ratio* also shows a smaller but positive impact (+14.44), hinting at demographic vitality. In contrast, the *regon_entities_ratio* (SHAP = –10.31 at 1,608 entities) slightly reduces the prediction, which may reflect saturation or diminishing marginal returns in areas with very high business density.

Most remaining variables contribute modestly. *Birthrate* (+5.58), the *hotels_beds_ratio* (+2.79), and *forests_ratio* (+0.74) collectively reinforce the positive prediction but with relatively small effects. Their limited magnitude suggests that Wrocław's compensation structure in 2010 was driven far more by social infrastructure, transportation connectivity, and labor-market conditions than by tourism capacity or environmental features.

In 2023, the Wrocław powiat showed significantly larger SHAP magnitudes than in 2010. The strongest contributor was still the *social_care_ratio*, this time with an even more extreme value of +577.18 at a feature value of 1.10, reinforcing the model's interpretation of a low social-care burden as a strong positive salary determinant. The *routes_ratio* follows with 458.16 at 14.20, highlighting substantial benefits from transport connectivity.

A major upward contribution also comes from the *regon_entities_ratio*, adding 440.86 at a high value of 2,205, implying that in 2023, business density exerted a far stronger positive effect than in 2010. The *unemployment_ratio* (+112.97) and *marriages_ratio* (+104.92) further elevated the compensation prediction, linking favorable labor-market and demographic conditions to higher wages.

Lesser yet notable effects included the *hotels_beds_ratio* (+72.29), *birthrate* (+12.15), and a small negative influence from the *forests_ratio* (–7.05). The *productive_population_ratio* contributed only +1.92, indicating minimal effect. Overall, the SHAP profile revealed that in 2023, Wrocław's compensation structure was strongly shaped by socioeconomic advantage, business density, and infrastructure, with demographic indicators reinforcing but not dominating the signal.

For Wrocław, the comparison of 2010 and 2023 reveals an increase in the strength of the main predictive factors and a shift in the importance of several of them. In 2010, the model was dominated by the *social_care_ratio* (+300.38 at 2.50) and *routes_ratio* (+213.06), with a notable but smaller effect from the *unemployment_ratio* (+124.48), while the *regon_entities_ratio* was even slightly negative (–10.31). In 2023, all major 2010 factors remained influential: the *social_care_ratio* (+577.18), *routes_ratio* (+458.16), and especially the *regon_entities_ratio* (+440.86 at 2205), indicating that business density became a key advantage for the city. At the same time, the *marriages_ratio* (+104.92) and *hotels_beds_ratio* (+72.29) gained significantly more importance than in 2010, while the effect of the *productive_population_ratio* decreased and became nearly neutral (+1.92). This shows that in 2023, compensation levels in Wrocław were primarily shaped by a combination of institutional-infrastructural strengths and high economic activity.

4. Conclusions

The results demonstrate that machine learning, when combined with interpretability methods, can capture the complexity of regional wage disparities beyond the scope of traditional econometric approaches. While labor market and demographic indicators consistently emerge as the strongest determinants, their relative importance evolves in response to broader socio-economic changes. The observed shifts between 2010 and 2023 underline the dynamic nature of regional wages formation, where structural conditions such as productive population ratios and enterprise density interact with local demographic and infrastructural contexts in non-linear ways.

Importantly, a SHAP-based analysis allows for a nuanced understanding of these dynamics by revealing how the same variable can contribute differently across powiats and time periods. This local interpretability enhances the practical value of predictive modeling for policymakers, offering insights that extend beyond aggregate associations. The findings suggest that data-driven approaches, when complemented with robust explanatory tools, provide not only accurate predictions but also meaningful guidance for regional development strategies aimed at mitigating wage inequalities.

References

Adamczyk, A., Tokarski, T., & Włodarczyk, R. W. (2009). Regional Wage Differences in Poland. *Gospodarka Narodowa. The Polish Journal of Economics*, *234*(9), 87–108. https://doi.org/10.33119/GN/101248.

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *82*(4), 1059–1086. https://doi.org/10.1111/rssb.12377.

Bartosik, K., & Mycielski, J. (2015). *Dynamika płac a długotrwałe bezrobocie w polskiej gospodarce* (INE PAN Working Paper Series no. 38). https://www.inepan.pl/images/pliki/Working_Papers/WorkingPapers_38.pdf.

Biecek, P., & Burzykowski, T. (2021). *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models*. CRC Press.

Blanchflower, D. G., & Oswald, A. J. (1990). The Wage Curve. *The Scandinavian Journal of Economics*, *92*(2), 215–235. https://doi.org/10.2307/3440026.

Bolińska, M., & Gomółka, A. (2018). Determinanty przestrzennego zróżnicowania płac w obwodach Ukrainy Zachodniej w latach 2004–2015. *Modern Management Review*, *23*, 31–44. https://doi.prz.edu.pl/pl/publ/zim/341.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In B. Krishnapuram & M. Shah (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785.

Combes, P.-P., Duranton, G., & Gobillon, L. (2008). Spatial Wage Disparities: Sorting Matters!. *Journal of Urban Economics*, *63*(2), 723–742. https://doi.org/10.1016/j. jue.2007.04.004.

Dykas, P., & Misiak, T. (2013). Determinanty przestrzennego zróżnicowania wybranych zmiennych makroekonomicznych. In M. Trojak & T. Tokarski (Eds.), *Statystyczna analiza przestrzennego zróżnicowania rozwoju ekonomicznego i społecznego Polski* (pp. 67–80). Wydawnictwo Uniwersytetu Jagiellońskiego.

Dykas, P., Misiak, T., & Tokarski, T. (2020). Determinants of spatial differentiation of labour markets in Ukraine. *Przegląd Statystyczny. Statistical Review*, *67*(1), 33–50. https://doi.org/10.5604/01.3001.0014.1784.

Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but *Many* are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, *20*, 1–18. https://jmlr.org/papers/volume20/18-760/18-760.pdf.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistic*s, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*(3), 916–954. https://doi.org/10.1214/07-AOAS148.

Greenwell, B. M., Bradley, C. B., & McCarthy, A. J. (2018). *A simple and effective model-based variable importance measure*. https://doi.org/10.48550/arXiv.1805.04755.

Hooker, G. (2004). Discovering additive structure in black box functions. In W. Kim, R. Kohavi, J. Gehrke & W. DuMouchel, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 575–580). Association for Computing Machinery. https://doi.org/10.1145/1014052.1014122.

Hooker, G. (2007). Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, *16*(3), 709–732. https://doi.org/10.1198/106186007X237892.

Kaliski, S. F. (1964). The Relation Between Unemployment and the Rate of Change of Money Wages in Canada. *International Economic Review*, *5*(1), 1–33. https://doi.org/10.2307/2525631.

Kapela, M., & Kwiatkowski, E. (2023). Regional Wage Differentiation and Qualitative Determinants of Economic Development: Evidence from Poland. *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie. Cracow Review of Economics and Management*, (3), 47–65. https://doi.org/10.15678/ZNUEK.2023.1001.0303.

Kingdon, G. G., & Knight, J. (2006). How Flexible Are Wages in Response to Local Unemployment in South Africa?. *ILR Review*, *59*(3), 471–495. https://doi.org/10.1177/001979390605900308.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (pp. 4765–4774). https://papers.nips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Luśtyk, A., Połeć, A., & Voznyuk, I. (2024). Wage Differences in Poland at the County Level and their Determinants. *Central European Economic Journal*, *11*(58), 447–460. https://doi.org/10.2478/ceej-2024-0028.

Machuca, J. A. L., & Cota, J. E. M. (2017). Salarios, desempleo y productividad laboral en la industria manufacturera mexicana. *Ensayos Revista de Economía, 36*(2), 185–228. https://ensayos.uanl.mx/index.php/ensayos/issue/view/10/17.

Masís, S. (2023). *Interpretable machine learning with Python*. Packt Publishing.

Molnar, C. (2020). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Leanpub. https://originalstatic.aminer.cn/misc/pdf/Molnar-interpretable-machine-learning_compressed.pdf.

Moretti, E. (2011). Local Labor Markets. In O. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics* (Vol. 4B, pp. 1237–1313). Elsevier. https://doi.org/10.1016/S0169-7218(11)02412-9.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87.

Phillips, A. W. (1958). The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957. *Economica*, *25*(100), 283–299. https://doi. org/10.1111/j.1468-0335.1958.tb00003.x.

Przekota, G. (2016). Ocena poziomu i przyczyn zróżnicowania wynagrodzeń w Polsce. *Roczniki Ekonomiczne Kujawsko-Pomorskiej Szkoły Wyższej w Bydgoszczy*, (9), 386–403. https://kpsw.edu.pl/pobierz/wydawnictwo/re9/przekota2.pdf.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?". Explaining the Predictions of Any Classifier. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778.

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn, A. & W. Tucker (Eds.), *Contributions to the Theory of Games* (Vol. 2, pp. 307–317). Princeton University Press. https://doi.org/10.1515/9781400881970-018.

Štrumbelj, E., & Kononenko, I. (2010). An Efficient Explanation of Individual Classifications using Game Theory. *Journal of Machine Learning Research*, *11*(1), 1–18. https://doi.org/10.1145/1756006.1756007.

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x.

Appendix 1.

The code used in the research study is presented below. The data were acquired directly from BDL through the webservices. To repeat the analysis for years other than 2010 and 2023 (assuming that data are available in the repository for the chosen years), the only line that requires change is ‘for YEAR in [2010,2023]:’.

```
import requests
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import shap
import dalex as dx

from sklearn.model_selection import train_test_split, RepeatedKFold, cross_validate
from sklearn.preprocessing import StandardScaler
# from sklearn.neural_network import MLPRegressor
from sklearn.inspection import PartialDependenceDisplay
from sklearn.metrics import r2_score, mean_squared_error
from xgboost import XGBRegressor

base_url = "https://bdl.stat.gov.pl/api/v1/data/by-variable/"
params = {
 'format': 'jsonapi',
 'unit-level': 5,
 'page-size': 100,
}

def get_data_by_variable(variable_id, variable_name, year):
 ids = []
 values = []

 for page in range(4):
 params['page'] = page
 params['year'] = year
 response = requests.get(f"{base_url}{variable_id}", params=params)
 data = response.json()

 for item in data['data']:
```

```
 attributes = item['attributes']
 id_ = item['id']
 val_data = attributes['values']

 if val_data:
 val = val_data[0]['val']
 ids.append(id_)
 values.append(val)

 return pd.DataFrame({variable_name: values}, index=ids)


for YEAR in [2010, 2023]:
 df_vars = {
 64428: 'compensation_level',
 60530: 'regon_entities_ratio',
 458700: 'social_care_ratio',
 395404: 'routes_ratio',
 60270: 'unemployment_ratio',
 60566: 'productive_population_ratio',
 450551: "birthrate",
 1646059: "forests_ratio",
 60300: "hotels_beds_ratio",
 450543: "marriages_ratio"
 }

 df = None
 for key, val in df_vars.items():
 df_current = get_data_by_variable(key, val, YEAR)
 if df is None:
 df = df_current
 else:
 df = df.join(df_current)

 # Basic dataset summary
 X = df.drop(columns=['compensation_level'])
 y = df['compensation_level']
 n_obs, n_features = X.shape
 print(f"\n=== YEAR {YEAR} ===")
 print(f"Number of observations: {n_obs}")
 print(f"Number of predictors: {n_features}")
 print(f"Observation-to-predictor ratio: {n_obs / n_features:.2f}")

 # Train–test split BEFORE scaling to avoid leakage
 X_train, X_test, y_train, y_test = train_test_split(
 X, y, test_size=0.05, random_state=42
 )

 xgb = XGBRegressor(
 n_estimators=200,
 learning_rate=0.01,
 max_depth=3,
 subsample=0.8,
 colsample_bytree=0.8,
 reg_lambda=1.0,
 random_state=42,
 objective="reg:squarederror"
 )

 cv = RepeatedKFold(n_splits=5, n_repeats=5, random_state=42)
 cv_results = cross_validate(
 xgb,
 X_train,
 y_train,
 cv=cv,
 scoring=['r2', 'neg_root_mean_squared_error'],
 return_train_score=True
 )

 print(f"CV mean test R²: {np.mean(cv_results['test_r2']):.3f}")
 print(f"CV mean test RMSE: {-
np.mean(cv_results['test_neg_root_mean_squared_error']):.3f}")

 xgb.fit(X_train, y_train)
 y_pred_train = xgb.predict(X_train)
 y_pred_test = xgb.predict(X_test)
 train_r2 = r2_score(y_train, y_pred_train)
 test_r2 = r2_score(y_test, y_pred_test)
```

```
train_rmse = mean_squared_error(y_train, y_pred_train)**.5
test_rmse = mean_squared_error(y_test, y_pred_test)**.5

print(f"Train R²: {train_r2:.3f}, RMSE: {train_rmse:.3f}")
print(f"Test R²: {test_r2:.3f}, RMSE: {test_rmse:.3f}")

model_explainer = dx.Explainer(xgb, X, y, label=f"XGB_{YEAR}")
vi = model_explainer.model_parts(N=10000, random_state=42)
print(vi.result)
fig_vi = vi.plot(show=False, title=f"XGB VIP explainer for compensation level in
year {YEAR}")
fig_vi.write_image(f"vip_plot_{YEAR}.png")
# SHAP analysis
explainer = shap.Explainer(xgb.predict, X, feature_names=X.columns)
shap_values = explainer(X)

# Jelenia Góra
index_jg = df.index.get_loc("030210161000")
shap_df_jg = pd.DataFrame({
'Feature': X.columns,
'SHAP Value': shap_values[index_jg].values,
'Feature Value': shap_values[index_jg].data
})
print("SHAP values for county Jelenia Góra (030210161000):")
print(shap_df_jg.sort_values(by='SHAP Value', key=abs,
ascending=False).head(10))

plt.figure(figsize=(18, 6))
plt.suptitle(f"Compensation Level Model SHAP values for Jelenia Góra for year
{YEAR}")
shap.plots.bar(shap_values[index_jg], max_display=10, show=False,
show_data=True)
plt.tight_layout(pad=1.0)
plt.savefig(f"Jelenia_shap_{YEAR}.jpg")
plt.show()

# Wrocław
index_wr = df.index.get_loc("030210564000")
shap_df_wr = pd.DataFrame({
```

```
'Feature': X.columns,
'SHAP Value': shap_values[index_wr].values,
'Feature Value': shap_values[index_wr].data
})
print("SHAP values for county Wrocław (030210564000):")
print(shap_df_wr.sort_values(by='SHAP Value', key=abs,
ascending=False).head(10))

plt.figure(figsize=(12, 6))
shap.plots.bar(shap_values[index_wr], max_display=10, show=False,
show_data=True)
plt.suptitle(f"Compensation Level Model SHAP values for Wrocław for year
{YEAR}")
plt.tight_layout(pad=1)
plt.savefig(f"Wroclaw_shap_{YEAR}.jpg")
plt.show()

# PDP plots (using scaled data from final model)
fig, ax = plt.subplots(figsize=(12, 8))
PartialDependenceDisplay.from_estimator(
xgb,
X,
features=list(range(X.shape[1])),
feature_names=X.columns,
ax=ax
)
plt.suptitle(f"Partial dependency plots for compensation level model for year
{YEAR}")
plt.tight_layout()
plt.savefig(f"PDP_{YEAR}.jpg")
plt.show()
```

Appendix 2.

The full results obtained after the execution of the code presented in Appendix 1 are as follows:

```
=== YEAR 2010 ===
Number of observations: 379
```

Number of predictors: 9
Observation-to-predictor ratio: 42.11
CV mean test R^2: 0.299
CV mean test RMSE: 340.347
Train R^2: 0.629, RMSE: 254.324
Test R^2: 0.369, RMSE: 192.135
Preparation of a new explainer is initiated

```
 -> data : 379 rows 9 cols
 -> target variable : Parameter 'y' was a pandas.Series. Converted to a numpy .ndarray.
 -> target variable : 379 values
 -> model_class : xgboost.sklearn.XGBRegressor (default)
 -> label : XGB_2010
 -> predict function : <function yhat_default at 0x000002C41C6A75B0> will be used (default)
 -> predict function : Accepts pandas.DataFrame and numpy.ndarray.
 -> predicted values : min = 2.72e+03, mean = 2.89e+03, max = 4.11e+03
 -> model type : regression will be used (default)
 -> residual function : difference between y and yhat (default)
 -> residuals : min = -6.88e+02, mean = 5.32, max = 1.9e+03
 -> model_info : package xgboost

A new explainer has been created!
 variable dropout_loss label
0 _full_model_ 251.572326 XGB_2010
1 hotels_beds_ratio 254.133193 XGB_2010
2 birthrate 254.864096 XGB_2010
3 marriages_ratio 259.696907 XGB_2010
4 forests_ratio 260.088788 XGB_2010
5 regon_entities_ratio 261.499846 XGB_2010
6 unemployment_ratio 287.919198 XGB_2010
7 productive_population_ratio 298.164109 XGB_2010
8 routes_ratio 313.799260 XGB_2010
9 social_care_ratio 320.226454 XGB_2010
10 _baseline_ 468.653795 XGB_2010
ExactExplainer explainer: 380it [00:52, 7.25it/s]
SHAP values for county Jelenia Góra (030210161000):
 Feature SHAP Value Feature Value
```

```
2 routes_ratio 107.268913 24.40
3 unemployment_ratio 26.644887 10.90
1 social_care_ratio 19.697117 5.90
4 productive_population_ratio -14.265568 64.70
0 regon_entities_ratio 9.167773 1499.00
5 birthrate 4.711150 -3.83
6 forests_ratio -2.413097 52.90
7 hotels_beds_ratio -1.588051 22.24
8 marriages_ratio -0.335435 5.40
SHAP values for county Wrocław (030210564000):
 Feature SHAP Value Feature Value
1 social_care_ratio 300.382093 2.50
2 routes_ratio 213.063043 12.20
3 unemployment_ratio 124.481779 5.50
4 productive_population_ratio -16.713112 65.80
8 marriages_ratio 14.444688 5.60
0 regon_entities_ratio -10.308929 1608.00
5 birthrate 5.582377 0.25
7 hotels_beds_ratio 2.794502 13.95
6 forests_ratio 0.744744 13.80

=== YEAR 2023 ===
Number of observations: 380
Number of predictors: 9
Observation-to-predictor ratio: 42.22
CV mean test R²: 0.251
CV mean test RMSE: 718.596
Train R²: 0.564, RMSE: 562.235
Test R²: -0.034, RMSE: 527.064
Preparation of a new explainer is initiated

 -> data : 380 rows 9 cols
 -> target variable : Parameter 'y' was a pandas.Series. Converted to a numpy
.ndarray.
 -> target variable : 380 values
 -> model_class : xgboost.sklearn.XGBRegressor (default)
 -> label : XGB_2023
 -> predict function : <function yhat_default at 0x000002C41C6A75B0> will be used
(default)
```

```
 -> predict function : Accepts pandas.DataFrame and numpy.ndarray.
 -> predicted values : min = 4.5e+03, mean = 6.56e+03, max = 9.46e+03
 -> model type : regression will be used (default)
 -> residual function : difference between y and yhat (default)
 -> residuals : min = -4.5e+03, mean = 2.39, max = 3.35e+03
 -> model_info : package xgboost

A new explainer has been created!
 variable dropout_loss label
0 _full_model_ 560.529352 XGB_2023
1 birthrate 563.920908 XGB_2023
2 marriages_ratio 569.890669 XGB_2023
3 unemployment_ratio 579.918983 XGB_2023
4 productive_population_ratio 580.117875 XGB_2023
5 forests_ratio 587.603146 XGB_2023
6 hotels_beds_ratio 591.842447 XGB_2023
7 routes_ratio 642.313577 XGB_2023
8 regon_entities_ratio 654.331861 XGB_2023
9 social_care_ratio 676.012299 XGB_2023
10 _baseline_ 935.037314 XGB_2023
ExactExplainer explainer: 381it [00:34, 8.13it/s]
SHAP values for county Jelenia Góra (030210161000):
 Feature SHAP Value Feature Value
0 regon_entities_ratio 168.111842 1815.00
4 productive_population_ratio 86.101887 54.90
2 routes_ratio 73.077743 27.90
3 unemployment_ratio 40.129574 3.90
1 social_care_ratio -17.931182 2.70
7 hotels_beds_ratio 9.112820 35.70
6 forests_ratio -7.528812 59.10
5 birthrate 5.067187 -8.75
8 marriages_ratio 0.236881 3.70
SHAP values for county Wrocław (030210564000):
 Feature SHAP Value Feature Value
1 social_care_ratio 577.179101 1.10
2 routes_ratio 458.159322 14.20
0 regon_entities_ratio 440.862895 2205.00
3 unemployment_ratio 112.970279 1.60
8 marriages_ratio 104.923334 5.80
```

7 hotels_beds_ratio 72.285200 21.63
5 birthrate 12.149382 -1.77
6 forests_ratio -7.046792 15.20
4 productive_population_ratio 1.921097 60.804 productive_population_ratio 3.846133 2.033800

Przegląd Statystyczny. Statistical Review, 2025, vol. 72, 3, 28–42 https://doi.org/10.59139/ps.2025.03.2
Received: 4.11.2025, revised: 31.12.2025, accepted: 4.01.2026

Empirical analysis of trade duration distributions: The WIG20 case

Agnieszka Lach[a]

Abstract. The aim of the study presented in this paper is to analyse the distributions of trade durations for WIG20 stocks using data from May 2025, with a particular focus on modelling doubly truncated data. Left-truncated distributions for trade durations have already been described in the literature, which is justified, as the values of an excessive proportion of observations were equal to zero. In this study, it is assumed that the data are also right-truncated due to time limitations between the trading sessions. Three doubly truncated continuous distributions were analysed in the study, namely the lognormal, the Pareto and the Weibull distribution. To satisfy the assumptions of stationarity and independence, the data were divided into smaller subsamples. Goodness-of-fit tests were then performed to determine which theoretical distribution best describes the empirical data. The results indicate that the quality of the fit depends on the lower truncation level – the higher the truncation threshold, the better the lognormal distribution fits the empirical trade durations.

Keywords: probability distributions, doubly truncated data, high-frequency data, trade durations
JEL: C12, C24, C41, G19

1. Introduction

The analysis of durations, defined as the waiting times between consecutive financial events, is an important aspect of market microstructure research. This analysis depends on the event type, measurement precision and the characteristics of the used data. This paper focuses on the time intervals between the successive transactions, referred to as trade durations. The timestamps of trades are recorded in milliseconds, and the research is based on tick-by-tick (intraday transaction) data.

Ni et al. (2010) distinguish between two main approaches to modelling durations: the mainstream finance approach and the econophysics approach. In the former, the most commonly used framework is the autoregressive conditional duration (ACD) model and its various extensions. In the econophysics approach, modelling is typically based on the continuous-time random walk (CTRW) framework.

A ground-breaking contribution to the mainstream finance approach was made by Engle and Russell (1998), who introduced the autoregressive conditional duration (ACD) model. Numerous modifications have since appeared in the literature,

[a] Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Department of Operations Research and Mathematical Economics, Al. Niepodległości 10, 61–875 Poznań, Poland, e-mail: agnieszka.lach@ue.poznan.pl, ORCID: https://orcid.org/0000-0002-2831-6336.

including the logarithmic ACD model, the Markov-switching ACD model and the threshold model. However, most of them were applied to low-frequency data – that is data measured with a precision of one second or longer – and to relatively old datasets (Li et al., 2023).

An overview of the second approach might be found in Ni et al. (2010). Early studies in this field suggest that inter-trade durations may be described by the power-law, modified power-law or stretched-exponential distributions. However, as asserted by Ni et al. (2010), the Weibull distribution often provides the best fit for inter-trade durations.

Recent research in this area has evolved toward modelling random variables characterised by data inconsistencies. The ongoing digitalisation of economic processes and the growing use of algorithmic trading on stock exchanges have led to order submissions and trade executions occurring within fractions of a second. Ultra-high-frequency trading algorithms now generate transactions at microsecond or even nanosecond intervals. This introduces rounding effects, resulting in a substantial proportion of observations with zero values. At the same time, not all market activity is driven by high-frequency trading. O'Hara (2015) distinguishes two broad categories of market participants: high-frequency traders and non-high-frequency traders, emphasising that both groups have to optimise their trading strategies with respect to market design and the behaviour of other traders. Moreover, O'Hara (2015) highlights the need for new analytical tools capable of capturing the evolving market microstructure, including changes in the size of trade, the trading volume, inter-trade durations, and the interdependencies among these characteristics.

In terms of inter-trade durations, the division of market activity into high-frequency and non-high-frequency trading naturally implies a distinction between durations close to zero and those of larger magnitudes. This perspective has already begun to be reflected in the literature. A high proportion of zero-valued data necessitates partitioning trade duration distributions. Two main approaches to this issue can be distinguished: truncating data close to zero or modelling the entire distribution. Empirical findings by Kızılersü et al. (2016) for the London Stock Exchange demonstrated that the durations recorded in the order book can be described by a left-truncated Weibull distribution, with the lower truncation point set at ten milliseconds. Kreer et al. (2022) analysed entire duration distributions, also for the London Stock Exchange, and found that a mixture of one exponential and one Weibull distribution models inter-trade waiting times remarkably well across all time horizons. The Weibull component captures the behaviour in the transition and tail regions, whereas the exponential component explains the excess mass at zero. Li et al. (2023) also analysed complete duration distributions and found that inter-trade durations of stocks follow bimodal distributions, driven by switching between market-making and speculative trading strategies.

In this article, the distribution of trade durations is modelled using doubly truncated distributions. Observations below a selected lower truncation point are excluded from the analysis, since durations close to zero represent a substantial proportion of the data and require separate modelling. Additionally, an upper truncation threshold is introduced to account for the daily nature of the data. Since the analysis is conducted on a day-by-day basis, it is reasonable to assume that trade durations should also be right-truncated.

The contribution of this paper to the existing knowledge is twofold. Firstly, doubly truncated distributions are applied to model trade durations. Secondly, the research is conducted using data from the Polish stock exchange.

2. Methodology

2.1. Trade durations

To formally define trade durations, a point-process framework is adopted. Trade durations represent time intervals between successive transactions. Let $\tau \in (0, \infty)$ be a variable representing physical time and let $\tau_0 = 0$. Furthermore, let $\tau_1, \tau_2, \ldots$ be non-negative random variables satisfying condition $\tau_t \leq \tau_{t+1}$ for $t = 1, 2, \ldots$. These variables indicate successive points in time at which transactions are executed on the stock exchange. A point process is defined as sequence (τ_t), $t = 1, 2, \ldots$. A trade duration is defined as a time interval between the consecutive events:

$$d_t = \begin{cases} \tau_t - \tau_{t-1}, & \text{for } t \geq 2, \\ \tau_1, & \text{for } t = 1. \end{cases} \tag{1}$$

Process (d_t), $t = 1, 2, \ldots$, is called the duration process and is referred to as the process associated with (τ_t).

2.2. Continuous distributions for trade durations

This subsection presents three distributions that are suitable for modelling doubly truncated trade durations. The same distributions – lognormal, Pareto and Weibull – were employed by Kızılersü et al. (2016) to model trade durations after left truncation. Special attention is paid to the tails of these distributions, as they determine the behaviour of extreme durations.

The probability density function of the lognormal distribution is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right], \tag{2}$$

where $x > 0$, $\mu \in \mathbb{R}$ is the location parameter, and $\sigma > 0$ is the scale parameter. The cumulative distribution function of the lognormal distribution has no closed-form expression and is determined using numerical methods. The tails of this distribution may range from light to medium-heavy (Čížek et al., 2005), where light tails denote those that decay at an exponential rate, and medium-heavy tails refer to those that decay more slowly than exponential but faster than power-law tails.

The Pareto distribution has the following probability density and cumulative distribution functions:

$$f(x) = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}, \tag{3}$$

$$F(x) = 1 - \left(\frac{\theta}{x+\theta}\right)^\alpha, \tag{4}$$

where $x > 0$, $\alpha > 0$ is the shape parameter, and $\theta > 0$ is the scale parameter. The Pareto distribution is classified as a heavy-tailed distribution (Klugman et al., 2008), meaning that its tail decays at a power-law rate.

Finally, the probability density function and the cumulative distribution function of the Weibull distribution are given by:

$$f(x) = \alpha\beta^{-\alpha}x^{\alpha-1} \exp\left[-\left(\frac{x}{\beta}\right)^\alpha\right], \tag{5}$$

$$F(x) = 1 - \exp\left[-\left(\frac{x}{\beta}\right)^\alpha\right], \tag{6}$$

where $x > 0$, $\alpha > 0$ is the shape parameter, and $\beta > 0$ is the scale parameter. The Weibull distribution has a heavy tail (Haas & Pigorsch, 2009).

The parameters of all distributions examined in this study were estimated using the maximum likelihood method.

2.3. Doubly truncated distributions

According to the classification of sample types proposed by Cohen (1991), this study deals with doubly truncated samples with known truncation points. If we denote the observations by x_i, $i = 1, \dots, n$, then for each observation in a doubly truncated sample the following holds: $T_1 \leq x_i \leq T_2$, where T_1 and T_2 are known truncation points.

Now the cumulative distribution function and the probability density function of a doubly truncated random variable are defined. Let X be a random variable with cumulative distribution function F and probability density function f, and let $Y = X | T_1 \leq X \leq T_2$ denote the doubly truncated random variable. Then the cumulative distribution function F^* and the probability density function f^* of the doubly truncated random variable are given by (Krysicki et al., 2004):

$$F^*(y) = \frac{F(y) - F(T_1)}{F(T_2 + 0) - F(T_1)} \text{ for } T_1 \leq y \leq T_2, \tag{7}$$

$$f^*(y) = \frac{f(y)}{F(T_2) - F(T_1)} \text{ for } T_1 \leq y \leq T_2. \tag{8}$$

The expressions above outline the general formulation of doubly truncated random variables. In empirical applications, closed-form expressions for the probability density or cumulative distribution functions in the doubly truncated case are often highly complex and therefore typically evaluated numerically.

2.4. Stationarity and independence tests

The goodness-of-fit between empirical distributions and selected continuous distributions can be assessed only when the stochastic processes under consideration are stationary and the observations are independent and identically distributed. Trade durations are typically stationary; however, they frequently attest to significant autocorrelation (Doman, 2011). Stationarity here means weak (covariance) stationarity, where the process has constant mean, variance and covariance over time.

In order to verify the stationarity of the time series, the Augmented Dickey–Fuller test was conducted. The null hypothesis of this test assumes that the analysed time series is non-stationary, while the alternative hypothesis presumes that the series is stationary.

The independence of the examined time series was tested with the Ljung-Box test. Under the null hypothesis, the observations are independent, i.e. all autocorrelation coefficients equal zero, while the alternative hypothesis assumes that at least one autocorrelation coefficient is nonzero.

2.5. Goodness-of-fit tests

This subsection presents the statistical tests used to evaluate how well the theoretical distributions fit the empirical trade duration data. Two classical goodness-of-fit tests were applied: the Kolmogorov-Smirnov and the Cramér-von Mises tests. The Anderson-Darling test, which is also popular, was intentionally omitted. The power of this test stems from its weighting function, which assigns greater weight to the tails of the distribution. When the distribution is doubly truncated, as is the case in this study, the weighting function would need to be modified accordingly, which is beyond the scope of this analysis.

Let us assume that we have a sample $\boldsymbol{X} = (X_1, \dots, X_n)'$ of i.i.d. random variables with an unknown distribution function F. To construct a goodness-of-fit test, both the empirical cumulative distribution function and the theoretical cumulative distribution function are required.

The empirical cumulative distribution function is defined as (Krzyśko, 2004):

$$F_n(x; \boldsymbol{X}) = \frac{\#\{1 \leq j \leq n : X_j \leq x\}}{n}, \tag{9}$$

where $x \in \mathbb{R}$, $\boldsymbol{X} \in \mathbb{R}^n$ and # denote the number of elements satisfying the condition. The theoretical cumulative distribution functions are presented in Subsection 2.2, and their doubly truncated forms are defined in Subsection 2.3.

The null and alternative hypotheses to be tested are as follows:

$$\begin{aligned} H_0 &: F = F_0, \\ H_1 &: F \neq F_0, \end{aligned} \tag{10}$$

where F_0 denotes the assumed theoretical distribution function.

The two goodness-of-fit statistics used in this study are defined below. The Kolmogorov-Smirnov statistic is defined as:

$$D_n = \sup_{\mathrm{x} \in \mathbb{R}} |F_n(x; \boldsymbol{X}) - F_0(x)|, \tag{11}$$

and the Cramér-von Mises statistic as:

$$W_n^2 = n \int_{-\infty}^{\infty} [F_n(x; \boldsymbol{X}) - F_0(x)]^2 dF_0(x). \tag{12}$$

In the empirical part of this study, the parameters of the theoretical distributions are estimated from the data, therefore composite hypotheses are tested. As noted by Pewsey (2018), in such cases the sampling distributions of the test statistics depend on several factors – including the form of the theoretical distribution, the estimated parameters, the estimation method and the sample size. Consequently, the exact distributions of the Kolmogorov-Smirnov and Cramér-von Mises statistics are unknown, and their sampling distributions were approximated using bootstrap methods.

The decision regarding the null hypothesis is based on the value of the test statistic. If the calculated value of the statistic exceeds the corresponding critical value, null hypothesis H_0 is rejected. Otherwise, there is no sufficient evidence to reject H_0.

2.6. The choice of truncation points

The selection of the left truncation point corresponds to identifying a boundary between high-frequency and non-high-frequency trading activity. This boundary is not fixed and has evolved over time with advances in computational speed and other operational and technological factors (O'Hara, 2015). Empirical evidence (Li et al., 2023) suggests that trade durations are characterised by bimodal distributions, with modes occurring at the millisecond and second levels; the former commonly associated with high-frequency traders and the latter with non-high-frequency traders. Accordingly, the boundary is assumed to lie between these regimes, and four left truncation points were considered (in seconds): 0.001, 0.01, 0.1, 1.

In addition to left truncation, right truncation was applied for practical reasons. When analysing daily data, trade durations cannot exceed the length of the trading session itself. Therefore, a right truncation point of 28,800 seconds was imposed, corresponding to eight hours. This choice does not substantially alter the modelling assumptions but yields favourable computational properties by mitigating some numerical issues.

Overall, the adopted truncation pattern reflects both the empirical characteristics of trade durations and the practical considerations related to data resolution and numerical estimation.

2.7. Description of the empirical study

The methodological framework was adapted from Kızılersü et al. (2016). The analysis was carried out separately for each company according to the following procedure:

1. Data collection and preliminary processing, which involved removing records with no trading volume and observations recorded outside the regular trading hours of the stock exchange (9:00 a.m. to 4:50 p.m.). Additionally, it was assumed that the time between the stock exchange's closing and reopening was equal to zero;
2. Calculation of trade durations according to formula (1);
3. Truncation of data below a specified lower threshold T_1, where $T_1 \in \{0.001, 0.01, 0.1, 1\}$, and above the upper threshold $T_2 = 28{,}800$ (equivalent to 8 hours in seconds);
4. Division of the doubly truncated sample obtained in step 3 into smaller samples of n = 125 observations each. Subsamples containing fewer than 125 observations (typically the last one) were excluded;
5. Testing for stationarity and independence in the subsamples obtained in step 4;
6. Conducting goodness-of-fit tests for doubly truncated lognormal, Pareto and Weibull distributions (applied only to stationary and independent subsamples).

This systematic approach ensured consistency across all the analysed companies and allowed a reliable comparison of goodness-of-fit results between different truncation levels.

3. Empirical study

3.1. Data

The study was carried out for 20 companies listed on the Warsaw Stock Exchange (WSE), included in the WIG20 index as of 31st May 2025. These were: Alior Bank (ALRR), Allegro (ALEP), Bank Pekao (PEO), Budimex (BDXP), CCC (CCCP), CD Projekt (CDR), Dino Polska (DNP), KGHM (KGH), Kęty (KTY), Kruk (KRU), LPP (LPPP), mBank (MBK), Orange Polska (OPL), Pepco (PCOP), PGE (PGE), PKN Orlen (PKN), PKO Bank Polski (PKO), PZU (PZU), Santander Bank Polska (SPL1) and Żabka (ZAB). The time and sales data for these companies over the period of May 2025 were obtained from the Eikon Refinitiv database. The time data were measured with a millisecond accuracy. All the results presented in this section are sorted in an ascending order by company market capitalisation as of 30th December 2025.

Descriptive statistics of trade durations are presented in Table 1. Although all the stocks included in the WIG20 index are a part of the same benchmark, they differ in terms of liquidity. For the most liquid company in the index, buy-sell transactions occur on average every 1.9 seconds, whereas for the least liquid one, approximately every 14.6 seconds. In general, higher market capitalisation is typically associated with higher liquidity. All the distributions are right-skewed and leptokurtic. The share of zero values in the total number of observations ranges from 38.4% to 50.7%. The high proportion of such observations suggests that zero-inflated models might be appropriate, meaning that observations close to zero can be treated differently from the remaining data. The following analysis focuses exclusively on the part of the dataset that remains after excluding the observations with values close to zero.

Table 1. Descriptive statistics for trade durations

Ticker[a]	N	Zero values (in %)	Min (s)	Max (s)	Mean (s)	St. Dev. (s)	Skewness	Excess kurtosis
KTY	36,559	42.7	0.0	1,106.0	14.6	42.2	6.2	68.1
CCCP	79,770	46.7	0.0	566.4	6.7	20.6	6.7	76.3
KRU	81,668	38.4	0.0	374.5	6.5	15.4	5.0	43.8
OPL	68,045	40.4	0.0	843.4	7.8	25.4	8.1	117.8
ALRR	122,392	47.8	0.0	362.8	4.4	14.7	6.8	71.6
BDXP	77,738	48.6	0.0	694.4	6.9	19.9	5.8	62.2
PCOP	133,632	43.1	0.0	281.8	4.0	11.4	5.5	48.0
PGE	108,336	41.6	0.0	363.0	4.9	14.4	5.6	50.1
ZAB	102,632	42.7	0.0	454.7	5.2	15.8	6.9	78.1
CDR	127,531	48.4	0.0	302.9	4.2	12.3	5.5	46.5
ALEP	216,991	43.7	0.0	267.1	2.5	7.5	6.4	67.3
LPPP	38,159	40.1	0.0	1,033.7	14.0	43.6	7.0	84.6
DNP	135,978	43.4	0.0	378.1	3.8	12.2	6.7	73.7
MBK	59,185	39.3	0.0	898.3	9.0	28.6	6.9	79.3
PEO	169,257	45.8	0.0	308.0	3.2	10.6	6.8	73.9
KGH	167,696	50.7	0.0	234.2	3.2	9.4	5.7	51.5
SPL1	74,748	47.9	0.0	610.6	7.1	22.2	6.1	60.1
PZU	123,814	47.8	0.0	309.8	4.3	11.9	5.1	40.4
PKO	247,259	47.4	0.0	278.0	2.2	7.1	7.1	85.6
PKN	279,210	43.6	0.0	168.9	1.9	5.1	5.7	57.5

a Tickers are ordered in an ascending order according to the company market capitalisation as of 30th December 2025.

Source: author's calculations.

3.2. Results

First, the stationarity and independence of the entire samples were examined separately for each company. The augmented Dickey-Fuller and Ljung-Box test results revealed that all the analysed samples are non-stationary and show significant autocorrelation. This justified dividing the samples into smaller subsamples.

To determine the appropriate subsample size, the data for the PGE company, after being doubly truncated, were sequentially divided into subsamples of the sizes of 30, 50, 75, 100, 125 and 150 observations. Stationarity and independence tests were then performed on each subsample, and the sample size boasting the highest passing rates for both tests simultaneously was selected. Kızılersü et al. (2016) adopted a sample size of 30 observations in their study; however, for the PGE data, the simultaneous pass rates for both tests did not exceed 16% for subsamples of this size. Subsequent sample sizes were chosen arbitrarily, starting from 50 and increasing by 25 each time. At the size of 150, the passing rates began to decrease again. Ultimately, the full samples were split into subsamples of 125 observations each, as this division ensured a reasonable proportion of stationary and independent subsamples. The number of subsamples of the size of 125 corresponding to various lower truncation thresholds for each company is presented in Table 2.

Table 2. Number of samples of the size of 125 for various lower truncation thresholds

Ticker[a]	Lower truncation threshold			
	0.001 s	0.01 s	0.1 s	1 s
KTY	145	131	111	89
CCCP	307	285	231	186
KRU	370	340	298	253
OPL	287	243	200	154
ALRR	460	398	320	226
BDXP	284	258	210	171
PCOP	556	497	403	308
PGE	454	385	317	242
ZAB	439	405	333	246
CDR	481	440	355	277
ALEP	887	778	615	434
LPPP	158	144	120	90
DNP	552	481	371	270
MBK	253	221	175	132
PEO	670	590	466	302
KGH	603	541	438	342
SPL1	277	250	197	145
PZU	468	420	350	282
PKO	958	844	633	436
PKN	1,170	1,049	887	599

[a] Tickers are ordered in an ascending order according to the company market capitalisation as of 30th December 2025.

Source: author's calculations.

The passing rates of the stationarity and independence tests for the samples of the size of 125 are presented in Table 3. Depending on the lower truncation threshold, the proportion of subsamples that passed the stationarity test ranges from 89.5% to 98.6%, while for the independence test from 73.4% to 93.7%. The proportion of the subsamples that passed both the stationarity and independence tests at the same time varies between 69.6% and 89.8%.

Table 3. Passing rates of stationarity and independence tests for the samples of the size of 125 (in %)

Ticker[a]	Lower truncation threshold											
	0.001 s			0.01 s			0.1 s			1 s		
	(1)[b]	(2)[c]	(3)[d]	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
KTY	98.6	82.8	82.1	97.7	78.6	77.1	97.3	88.3	86.5	96.6	86.5	83.2
CCCP	94.5	81.4	77.9	89.5	81.4	74.0	93.5	83.6	78.4	95.7	83.9	81.2
KRU	97.3	86.8	84.3	97.9	88.5	87.4	97.7	86.9	84.9	98.0	88.9	87.8
OPL	94.4	85.0	82.2	93.0	84.4	80.3	90.0	80.0	77.0	95.5	81.8	80.5
ALRR	94.4	86.3	82.8	96.2	87.4	84.2	94.7	84.1	80.6	96.9	92.0	89.8
BDXP	94.0	84.5	80.3	94.6	81.0	78.3	96.2	81.4	80.5	97.7	85.4	84.2
PCOP	96.0	83.3	81.3	96.6	81.5	79.1	95.8	80.9	78.2	96.1	82.1	79.9
PGE	95.4	81.7	79.3	94.6	81.6	79.0	95.3	83.3	82.0	97.5	87.6	86.8
ZAB	93.9	78.6	74.9	92.4	79.3	75.6	94.0	76.9	74.5	94.7	82.9	80.1
CDR	94.8	86.1	83.0	93.9	85.2	81.6	92.4	87.0	83.7	96.0	91.0	88.8
ALEP	94.1	83.7	80.1	94.3	83.6	80.3	93.3	78.4	75.1	95.6	84.1	81.3
LPPP	94.9	85.4	82.9	91.0	81.9	76.4	91.7	76.7	73.3	93.3	84.4	78.9
DNP	96.4	84.1	82.3	92.5	82.3	78.8	94.1	83.0	79.5	96.3	91.5	88.2
MBK	94.9	93.7	89.7	97.3	88.2	87.3	96.0	85.7	82.9	96.2	86.4	85.6
PEO	96.1	84.2	82.2	94.9	84.1	80.5	94.0	83.5	81.1	96.4	89.4	86.8
KGH	95.7	82.3	80.1	95.9	82.8	80.2	95.4	79.9	76.9	95.9	84.2	81.0
SPL1	96.8	86.6	85.2	95.2	86.8	84.4	94.9	83.3	80.2	98.6	89.7	89.0
PZU	94.0	84.6	81.4	94.8	83.1	80.5	95.1	82.9	81.4	95.4	84.0	81.2
PKO	97.2	83.8	81.7	96.8	83.8	81.5	96.4	83.3	81.4	96.1	89.5	87.2
PKN	93.6	81.1	78.0	93.0	78.1	74.8	90.6	73.4	69.6	96.8	88.2	86.6

a Tickers are ordered in an ascending order of company market capitalisation as of 30th December 2025. b Augmented Dickey-Fuller test for stationarity. c Ljung-Box test for independence. d Both tests (b and c).
Source: author's calculations.

The next step involved conducting goodness-of-fit tests, applied only to samples that had successfully passed the stationarity and independence checks. The tests were performed using the Kolmogorov-Smirnov and Cramér-von Mises statistics. Given the similarity of the results, only the passing rates from the Cramér-von Mises test have been reported (Table 4). The interpretation depends on the value of the lower truncation point. For the smallest truncation point of 0.001 s, the highest passing rates were observed for the Weibull distribution (for 17 out of the 20 companies analysed), although the passing rates for the lognormal distribution were also notably high. At the next truncation level, 0.01 s, both the lognormal and Weibull distributions recorded the highest passing rates for 10 firms each. For the remaining truncation levels, i.e. 0.1 s and 1 s, the passing rates for the lognormal distribution reached 100% across all firms. The Pareto distribution represented the weakest overall fit.

Table 4. Passing rates of Cramér-von Mises test for doubly truncated samples of the size of 125 (in %)

Ticker[a]	Lower truncation threshold											
	0.001 s			0.01 s			0.1 s			1 s		
	(1)[b]	(2)[c]	(3)[d]	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
KTY	73.1	0.0	96.5	90.1	3.1	90.8	100.0	13.5	86.2	100.0	56.2	79.3
CCCP	88.9	1.3	96.7	92.3	10.9	95.1	100.0	18.2	84.4	100.0	68.1	82.4
KRU	98.1	3.2	85.6	100.0	14.1	75.5	100.0	50.3	78.3	100.0	79.7	82.9
OPL	58.9	1.4	99.3	93.4	6.6	96.3	100.0	18.0	82.7	100.0	59.5	82.3
ALRR	74.6	3.3	100.0	93.2	5.3	98.7	100.0	26.6	90.9	100.0	80.1	87.3
BDXP	85.9	1.8	98.2	95.4	7.0	91.5	100.0	23.9	82.3	100.0	64.7	79.5
PCOP	91.4	0.9	98.4	96.0	8.7	94.7	100.0	36.2	81.0	100.0	87.3	84.6
PGE	69.4	2.2	99.6	97.4	4.2	95.8	100.0	16.4	80.1	100.0	79.8	82.2
ZAB	96.1	2.3	91.0	93.6	5.9	87.6	100.0	23.7	75.3	100.0	72.5	83.2
CDR	90.2	1.5	96.6	98.4	7.1	91.1	100.0	40.0	76.1	100.0	84.5	86.6
ALEP	89.3	2.5	97.9	99.1	8.9	96.2	100.0	54.3	88.5	100.0	92.4	88.0
LPPP	75.3	1.9	99.4	90.3	4.9	100.0	100.0	14.2	93.3	100.0	50.0	86.1
DNP	71.0	2.7	99.6	90.6	14.8	99.0	100.0	24.3	88.1	100.0	87.4	90.1
MBK	63.6	1.6	100.0	85.5	10.0	100.0	100.0	21.7	86.3	100.0	61.4	85.0
PEO	85.2	2.7	98.9	95.9	11.0	97.1	100.0	35.8	89.7	100.0	87.0	83.0
KGH	90.7	1.7	98.3	97.6	8.3	93.1	100.0	49.1	82.7	100.0	91.7	89.9
SPL1	70.0	1.4	100.0	85.6	9.6	98.4	100.0	15.8	75.0	100.0	61.1	83.9
PZU	94.4	1.5	97.4	98.3	5.0	91.4	100.0	37.8	82.3	100.0	85.4	81.4
PKO	90.3	2.0	98.5	95.9	14.6	96.0	100.0	42.2	90.5	100.0	95.9	84.5
PKN	99.1	1.2	94.2	99.9	6.1	91.5	100.0	75.3	91.0	100.0	99.7	93.8

[a] Tickers are ordered in ascending order of company market capitalisation as of 30th December 2025. [b] Lognormal distribution. [c] Pareto distribution. [d] Weibull distribution.

Source: author's calculations.

For comparison purposes, the Kolmogorov-Smirnov and Cramér-von Mises tests were also applied to the left-truncated samples, and again only to those which had successfully passed the stationarity and independence tests. As before, only the passing rates from the Cramér-von Mises test have been reported (Table 5). Several observations can be made on their basis. For the left-truncated samples with left truncation points set at 0.001 s or 0.01 s, the highest passing rates in the majority of cases were reported for the Weibull distribution. Notably, at the lowest left truncation point of 0.001 s, the passing rates for the lognormal distribution were very low. For higher left truncation levels, namely 0.1 s and 1 s, the results were similar to those obtained for doubly truncated samples, with the lognormal distribution showing the best overall fit. In general, the passing rates for doubly truncated samples were higher than those for left-truncated samples.

Table 5. Passing rates of Cramér-von Mises test for left truncated samples of the size of 125 (in %)

Ticker[a]	Lower truncation threshold											
	0.001 s			0.01 s			0.1 s			1 s		
	(1)[b]	(2)[c]	(3)[d]	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
KTY	10.3	4.8	20.3	17.6	9.2	25.2	94.6	17.3	19.6	98.9	77.1	58.2
CCCP	2.6	1.0	61.4	28.8	10.0	64.6	98.7	17.8	55.6	99.4	70.6	67.3
KRU	7.3	6.2	51.0	56.5	20.4	44.7	100.0	55.3	65.7	100.0	91.3	76.2
OPL	7.0	16.9	44.2	28.0	9.4	54.6	97.5	14.7	48.8	100.0	62.7	71.5
ALRR	9.8	7.5	58.7	37.7	6.4	86.4	87.5	11.6	63.7	99.5	82.8	77.2
BDXP	6.0	2.5	51.8	24.8	9.2	48.2	99.1	30.2	58.6	100.0	75.4	68.2
PCOP	9.0	1.1	66.5	53.9	13.7	79.9	99.8	30.5	65.7	100.0	92.1	75.1
PGE	7.3	8.3	63.6	43.4	4.7	72.3	99.4	17.8	55.7	100.0	80.0	69.4
ZAB	12.8	3.0	70.4	43.2	18.2	69.2	96.7	22.0	39.5	99.1	70.1	58.5
CDR	13.3	2.3	68.6	44.3	10.9	65.8	99.7	36.9	61.0	100.0	88.8	80.8
ALEP	13.4	3.9	72.2	71.7	14.7	89.6	100.0	39.1	75.5	100.0	90.0	80.3
LPPP	4.4	2.6	36.5	18.1	10.5	48.3	81.7	13.9	43.9	92.2	59.4	59.0
DNP	10.3	7.3	54.6	41.4	16.6	90.6	99.2	19.0	68.5	100.0	87.9	84.3
MBK	7.5	11.5	39.8	24.4	12.8	54.1	96.0	18.3	34.3	96.8	66.0	69.9
PEO	17.6	4.6	63.8	53.2	13.6	92.4	95.3	23.2	68.3	100.0	85.1	76.9
KGH	9.0	3.0	73.0	52.0	12.1	72.8	100.0	46.1	71.4	100.0	95.2	84.1
SPL1	11.2	5.5	53.3	21.6	13.8	55.2	90.8	11.4	30.9	100.0	80.0	73.1
PZU	8.6	2.1	69.8	47.9	5.1	55.8	99.7	33.9	70.4	100.0	93.8	77.3
PKO	13.4	2.3	60.7	62.3	16.9	91.6	100.0	32.9	77.1	100.0	94.2	74.1
PKN	17.5	1.5	84.9	95.5	34.4	82.5	100.0	56.5	85.3	100.0	100.0	85.7

a Tickers are ordered in ascending order according to the company market capitalisation as of 30th December 2025. b Lognormal distribution. c Pareto distribution. d Weibull distribution.

Source: author's calculations.

4. Limitations of the study

The analysis presented in this paper is exploratory in nature. Its main limitations are discussed below, along with directions for future research.

To start with, this study focuses exclusively on trade durations, which represent only one, albeit important, component of market microstructure. Other relevant characteristics, including price dynamics, trade size and trading volume, should be examined in further research.

Moreover, due to the precision of the available data, recorded at a millisecond level, the analysis is restricted to non-high-frequency trading activity. In this setting, all zero-valued durations can be clearly attributed to high-frequency traders who account for approximately 40–50% of the observations. Although attempts have been made in the literature to model such mixed data at a millisecond precision using exponential and Weibull distributions mixed together (Kreer et al., 2022), such an approach is considered to involve substantial simplification. Accurately modelling the distributional mass close to zero would require data recorded with a higher time precision, which were not available to the author at the time of conducting this study. Consequently, the analysis focuses on modelling the left-truncated part of the distribution.

Also, this study should be regarded as an initial step towards more comprehensive analyses. The data span was arbitrarily chosen and limited to a single month, serving primarily as an illustrative sample. Future research should aim to investigate the potential differences across time scales and to identify calendar-related effects such as variations across months, weeks of the month or days of the week.

It must also be remembered that the studied data referred only to 20 companies from the WIG20 index, so the scope of the sectoral analysis was limited. Most companies in the sample belonged to different sectors, according to the classification provided by the WSE. Moreover, no clear patterns were identified with respect to the market capitalisation of the companies analysed in this study. A more detailed analysis would require a broader set of companies.

Finally, the static model presented here may be embedded in dynamic models. Such attempts have already been reported in the literature, for example in Li et al. (2023), where static distributional components were combined with dynamic mechanisms. Extending the analysis in this direction in future research might also be worthwhile.

5. Conclusions

The comparative analysis indicates that the observed differences in trade duration distributions are driven by both market-specific factors and truncation choices. At higher left truncation levels, the lognormal distribution is preferred regardless of whether left-truncated or doubly truncated samples are examined. This finding distinguishes the WSE from the London Stock Exchange, where a consistent preference for the Weibull distribution was shown by previous studies.

For smaller values of the left truncation point, the results depend on whether left or double truncation is applied. In the case of the left truncation, the Weibull distribution provides the best fit, whereas in the case of the double truncation, the Weibull and lognormal distributions provide a similar fit.

Overall, the passing rates of the goodness-of-fit test are generally higher for doubly truncated samples than for left-truncated ones, suggesting that explicitly accounting for the natural upper bound imposed by the trading session length may lead to a more appropriate modelling of trade durations. These conclusions should be interpreted in the light of the data and modelling limitations discussed in the preceding section.

References

Cohen, A. C. (1991). *Truncated and Censored Samples. Theory and Applications.* CRC Press. https://doi.org/10.1201/b16946.

Čížek, P., Härdle, W. K., & Weron, R. (Eds.). (2005). *Statistical Tools for Finance and Insurance.* Springer. https://doi.org/10.1007/b139025.

Doman, M. (2011). *Mikrostruktura giełd papierów wartościowych.* Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu.

Engle, R. F., & Russell, J. R. (1998). Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. *Econometrica*, *66*(5), 1127–1162. https://doi.org/10.2307/2999632.

Haas, M., & Pigorsch, C. (2009). Financial Economics, Fat-Tailed Distributions. In M. A. Meyers (Ed.), *Encyclopedia of Complexity and Systems Science* (pp. 3404–3435). Springer. https://doi.org/10.1007/978-0-387-30440-3_204.

Kızılersü, A., Kreer, M., Thomas, A. W., & Feindt, M. (2016). Universal behaviour in the stock market: Time dynamics of the electronic orderbook. *Physics Letters A*, *380*(33), 2501–2512. https://doi.org/10.1016/J.PHYSLETA.2016.05.035.

Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2008). *Loss Models: From Data to Decisions.* John Wiley & Sons. https://doi.org/10.1002/9780470391341.

Kreer, M., Kızılersü, A., & Thomas, A. W. (2022). Censored expectation maximization algorithm for mixtures: Application to intertrade waiting times. *Physica A: Statistical Mechanics and its Applications*, *587*, 1–11. https://doi.org/10.1016/J.PHYSA.2021.126456.

Krysicki, W., Bartos, J., Dyczka, W., Królikowska, K., & Wasilewski, M. (2004). *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach: część 1. Rachunek prawdopodobieństwa.* Wydawnictwo Naukowe PWN.

Krzyśko, M. (2004). *Statystyka matematyczna*. Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Li, Z., Chen, X., & Xing, H. (2023). A multifactor regime-switching model for inter-trade durations in the high-frequency limit order market. *Economic Modelling*, *118*, 1–42. https://doi.org/10.1016/J.ECONMOD.2022.106082.

Ni, X.-H., Jiang, Z.-Q., Gu, G.-F., Ren, F., Chen, W., & Zhou, W.-X. (2010). Scaling and memory in the non-Poisson process of limit order cancelation. *Physica A: Statistical Mechanics and its Applications*, *389*(14), 2751–2761. https://doi.org/10.1016/J.PHYSA.2010.02.040.

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, *116*(2), 257–270. https://doi.org/10.1016/J.JFINECO.2015.01.003.

Pewsey, A. (2018). Parametric bootstrap edf-based goodness-of-fit testing for sinh-arcsinh distributions. *TEST*, *27*(1), 147–172. https://doi.org/10.1007/S11749-017-0538-2.

Przegląd Statystyczny. Statistical Review, 2025, vol. 72, 3, 43–46 https://doi.org/10.59139/ps.2025.03.3

Report from the 43rd National Scientifc Conference named after him. Professor Władysław Bukietyński 'Methods and Applications of Operations Research' – MZBO 2025

Przemysław Szufel[a]

The 43rd National Scientific Conference named after him. Professor Władysław Bukietyński 'Methods and Applications of Operations Research' (Pol. XLIII Ogólnopolska Konferencja Naukowa im. Profesora Władysława Bukietyńskiego Metody i Zastosowania Badań Operacyjnych – MZBO 2025) was held on 12th–14th October 2025 at the SGH Warsaw School of Economics, Poland. The conference was organised by the Decision Analysis and Support Unit, Collegium of Economic Analysis at the SGH Warsaw School of Economics. Detailed information about the conference can be found at the following address: https://mzbo2025.sgh.waw.pl/.

The Organising Committee was chaired by Małgorzata Wrzosek, PhD, Assistant Professor at the SGH Warsaw School of Economics, while the Scientific Committee by Przemysław Szufel, PhD, DSc, Associate Professor at the SGH Warsaw School of Economics.

The conference was held under the patronage of the Committee of Statistics and Econometrics of the Polish Academy of Sciences and the Polish Chapter of INFORMS with the support of the Polish National Agency for Academic Exchange under the Strategic Partnerships programme, grant number BPI/PST/2024/1/00129.

The conference focused on the methodological and application aspects of operations research:

- modelling of capital investments;
- optimisation in banking and insurance;
- stock-market analysis;
- optimisation in transport and inventory management;
- consumer-preference studies;
- time-series analysis;
- classical operational-research methods;
- evolutionary and ant-colony algorithms;

[a] SGH Warsaw School of Economics, Collegium of Economic Analysis, Institute of Econometrics, Al. Niepodległości 162, 02–554 Warszawa, Poland, e-mail: pszufe@sgh.waw.pl, ORCID: https://orcid.org/0000-0001-9525-3497.

- neural networks;
- multi-criteria analysis;
- stochastic dominance;
- non-linear programming algorithms;
- the chaos theory.

The 43rd edition of MZBO sought to address, as comprehensively as possible, issues of importance to the Polish operations research community. The long-established list of conference topics was expanded to include subjects which had emerged from technological progress and the increased computational capabilities of modern hardware. The conference featured sessions devoted to practical challenges and the application of quantitative methods to real-world economic and managerial problems (with particular emphasis on finance and logistics), alongside sessions focusing primarily on theoretical contributions. Special attention was given to contemporary methods of data analysis (including data mining, big data analytics and deep learning) and to the use of advanced and/or large-scale computational approaches in decision support. A detailed description of the thematic coverage and the structure of each session is available on the conference website.

This year's meeting was expanded to include a poster session aimed directly at undergraduate and doctoral students interested in decision analysis, operations research and any related areas. Its purpose was to provide early-career researchers with an attractive opportunity to acquire knowledge and experience and to build professional connections within the operations research community.

The conference gathered 57 participants. This group consisted of faculty members or doctoral students of universities from Poland and Canada, including: Dalhousie University (Halifax, Canada), AGH University of Krakow, Bialystok University of Technology, Krakow University of Economics, Maria Curie-Skłodowska University in Lublin, Nicolaus Copernicus University in Toruń, Poznań University of Economics and Business, Poznań University of Life Sciences, SGH Warsaw School of Economics, University of Economics in Katowice, University of Lodz, University of Szczecin, Wroclaw University of Economics and Business, WSB Merito University Poznan, University of Warsaw, Wrocław University of Science and Technology, Poznan University of Technology, the State University of Applied Sciences in Jaroslaw, Systems Research Institute Polish Academy of Sciences, Institute of Economics of the Polish Academy of Sciences, and the Polish Academy of Sciences. Practitioners from PKO Bank Polski, GSK, Roche and Diuna Group were also present.

During the conference, 32 papers and six posters were presented, focusing on various theoretical and practical aspects of operational research methods and data analytics. Additionally, two special sessions were held. The first one was organised by the Operational Research Section of the Committee on Statistics and Econometrics of the Polish Academy of Sciences and chaired by Marcin Anholcer, PhD, DSc, Associate Professor at Poznań University of Economics and Business. During this session, Paweł Kropiński delivered a lecture on the optimisation of investment strategies. The other special session was organised by the Polish Chapter of INFORMS and chaired by Ewa Roszkowska, PhD, DSc, ProfTit from the Białystok University of Technology. Business practitioners, Witold Fidos (PKO Bank Polski) and Jakub Witkowski (Roche), presented various practical applications of AI and Machine Learning.

The remaining sessions were chaired by Krzysztof Echaust, Michał Jakubczyk, Ignacy Kaliszewski, Bogumił Kamiński, Jerzy Michnik, Józef Stawicki, Tomasz Szapiro, Grzegorz Tarczyński, and Tadeusz Trzaskalik.

There were two keynote addresses. The first one, delivered by Stan Matwin, PhD, Professor Emeritus at Dalhousie University (Halifax, Canada), was titled 'Artificial Intelligence – A serious and personal perspective'. Professor Matwin discussed the role of AI in supporting decision-making processes. The second keynote lecture, 'Double agency and co-evolution for two-mode networks, with an application to corporate interlocks and firms' environmental performance', was presented by Beata Łopaciuk-Gonczaryk, PhD, DSc, Associate Professor at the University of Warsaw, and focused on the Stochastic Actor-Oriented Model for two-mode networks.

During the thematic sessions of the conference, the presentations were authored or co-authored by Marcin Anholcer, Maciej Bartkowiak, Milena Bieniek, Tomasz Brzęczek, Krzysztof Dmytrów, Krzysztof Echaust, Marzena Filipowicz-Chomko, Maciej Fronc, Dorota Górecka, Michał Jakubczyk, Małgorzata Just, Ignacy Kaliszewski, Daniel Kaszyński, Łukasz Kraiński, Adam Kucharski, Konrad Kułakowski, Aleksandra Łuczak, Anna Łyczkowska-Hanćkowiak, Elżbieta Majewska, Jerzy Michnik, Monika Niegowska-Postek, Mariusz Połeć, Ewa Roszkowska, Karolina Sobczak-Marcinkowska, Michał Stasiak, Małgorzata Szałucka, Marek Szopa, Grzegorz Tarczyński, Krzysztof Targiel, Rafał Weron, Aleksandra Wójcicka-Wójtowicz, Małgorzata Wrzosek, Piotr Zaborowski, Sebastian Zając, Mateusz Zawisza. The presentation titles and abstracts can be found on the conference website at https://mzbo2025.sgh.waw.pl/en/program.

The poster session included six posters authored or co-authored by Jakub Karnowski, Adam Kasiński, Magdalena Ligus, Dawid Linek, Antoni Łopacz, Piotr Peternek, Przemysław Szufel, and Michał Wójcik.

Traditionally, the Best Conference Paper competition was held during the event. The results were as follows:

- **First Prize:** Karolina Sobczak-Marcinkowska, PhD, Assistant Professor from Poznań University of Economics and Business, for her presentation titled 'Dynamics of Pro-Environmental Consumer Behaviour in the Framework of a Differential-Equations Model';
- **Second Prize:** Daniel Kaszyński, PhD, Assistant Professor from the SGH Warsaw School of Economics, for his presentation on 'Algorithmic Bias in Creditworthiness Assessment';
- **Third Prize:** Milena Bieniek, PhD, Assistant Professor from Maria Curie-Skłodowska University in Lublin, for her presentation titled 'The Impact of an Exchange Mechanism on Pricing and Logistics Decisions in a Supply Chain with Stochastic Power-Law Demand and a Multi-Criteria Decision Structure'.

The next MZBO conference will be organised by the Department of Operations Research and Mathematical Economics at the Poznań University of Economics and Business. The event will be held on 18th–20th October 2026. Information about the conference is available at: https://mzbo2026.ue.poznan.pl/.