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

QUANTITATIVE METHODS IN ECONOMICS

METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH

Volume XXVI, No. 4

Warsaw 2025

ISSN 2082-792X
e-ISSN 2543-8565

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

Warsaw 2025, Volume XXVI, No. 4
Journal homepage: https://qme.sggw.edu.pl

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

CONTENTS

QUANTITATIVE METHODS IN ECONOMICS
Volume XXVI, No. 4, 2025, pp. 141 – 154

Received: 13.08.2025
Accepted: 01.10.2025

UNSUPERVISED CLASSIFICATION OF JAPANESE CANDLESTICKS: TECHNICAL ANALYSIS VS MACHINE LEARNING

Maciej Janowicz https://orcid.org/0000-0002-1584-2089
Institute of Information Technology
Warsaw University of Life Sciences – SGGW, Poland
e-mail: maciej_janowicz@sggw.edu.pl
Luiza Ochnio https://orcid.org/0000-0001-8885-7945
Institute of Economics and Finance
Warsaw University of Life Sciences – SGGW, Poland
e-mail: luiza_ochnio@sggw.edu.pl

Abstract: The unsupervised clusterization of a set of Japanese Candlesticks generated by the currency pair prices on the Forex market has been performed. Several algorithms that do not require the number of clusters in advance have been used. It turns out that different algorithms give glaringly different numbers and description of clusters. Comparison with well-established candlestick types known from the technical analysis has been made.

Keywords: Japanese candlesticks, foreign exchange market, unsupervised learning

JEL classification: C51, C52

INTRODUCTION

Candlestick patterns form one of the cornerstones of technical analysis of assets in trade markets. They provide important information not only about the prices but also about the sentiments of investors or speculators. Each candlestick is constructed using open, high, low, and close (OHLC) prices of a single session.
A candlestick is composed of three elements: upper shadow, body and lower shadow [Lempart & Zalewski 2013, Bigalow 2002, Bulkowski 2008, Morris 2006, Morris 1998, Nison 2001, Nison 1996]. The body of the candlestick is determined by the open and close prices. One draws a rectangle the length of the which is the absolute

https://doi.org/10.22630/MIBE.2025.26.4.13

value of the difference between the close and open prices. If the close price is higher than the open price, the body is drawn as green or white. If, however, the close price is lower than the open price, the color is red or black. If the body is green (or white) then the length of the upper shadow is equal to the difference between the high and close prices. If the body is red (or black) that length is equal to the difference between the high and open prices. If the body is green (or white) then the length of the lower shadow is equal to the difference between the open and low prices. If the body is red (or black) that length is equal to the difference between the close and low prices. The candlesticks and their formations have been cataloged in many sources of which we singled out *Leksykon formacji świecowych* with 19 single-candle formations (Lempart & Zalewski 2013). We have added two kinds of rather well-known candlesticks, such as Hammer and Shooting Star [Nison 2001] to the list given in *Leksykon*.Our goal in this study is to apply unsupervised learning algorithms (like KMeans, MeanShift, DBSCAN) to the Forex market daily data to test whether the traditional, intuitive, technical-analysis patterns emerge naturally. We address the question: Do data-driven clusters align with analysts' intuition?

METHODS

A. Data

We have used daily EUR/USD and other candlestick data from 2009-05-13 until 2025-05-23, sourced from (Forexsb 2025). Each candlestick provides Open, High, Low, Close prices.

B. General characteristics of candlesticks

Descriptive statistics

We have obtained basic descriptive statistics measures characterizing the Japanese candlesticks of 28 most important currency pairs.

Table 1. Descriptive statistics for full candlesticks lengths

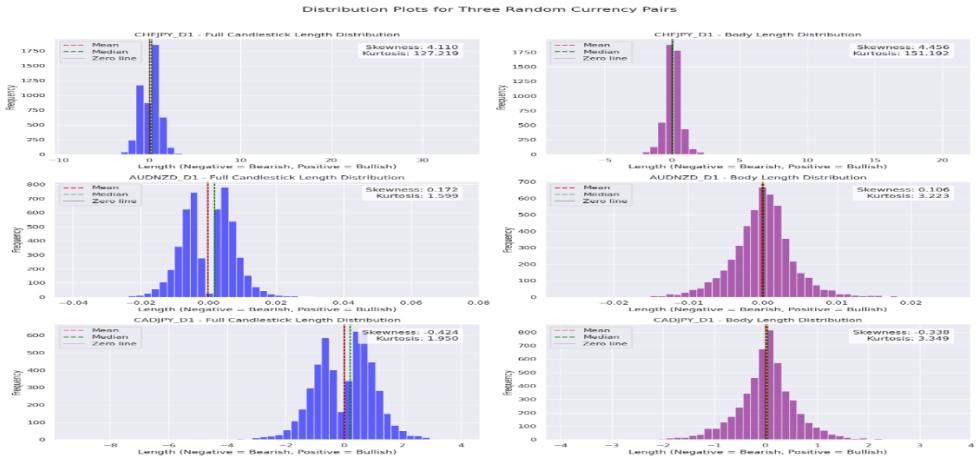
Currency pair	EURUSD	GBPUSD	EURCHF	USDJPY	EURGBP
Mean Media	0.000051	0.000064	-0.000240	0.009620	0.000023
Median	0.000940	0.001560	0.000690	0.176000	0.000940
St. dev.	0.010292	0.012715	0.008883	1.027077	0.006667
Skewness	-0.002573	-0.400521	-11.936007	-0.499266	0.204083
Kurtosis	0.355493	3.909116	494.997093	2.968342	2.879827
Min	-0.042560	-0.142330	-0.351140	-6.906000	-0.042730
Max	0.054850	0.045550	0.117140	4.707000	0.058270

Negative "lengths" refer to the bearish candlesticks.
Source: authors' own calculations

Examples have been sampled in Table 1. Perhaps the most glaring feature of the data in those tables is the very large kurtosis for the EURCHF pair. But we would those basic descriptive-statistics of our strong disbelief in the methodological correctness of testing trading strategies for the entire manifold of time series generated by different currency pairs or assets.

In Figure 1 we have displayed (randomly chosen) examples of the candlestick distributions. The negative "lengths" are related to the bearish candlesticks.

Figure 1. Examples of distributions of candlestick lengths



Source: authors' own calculations

C. Feature Extraction for Clusterization

The absolute positions of the candlestick elements have obviously no inherent meaning and are useless for classification purposes. It is almost always useful in classification problems to define some engineered features out of those given in the datasets. In our case it is actually necessary to do so.

Therefore, we have defined the following 6 engineered features for classification aiming to balance descriptiveness and simplicity. They are derived from the basic OHLC (Open, High, Low, Close) data:

1. Body Length (Absolute)**:**
   - Definition**:** |Close - Open|
   - Purpose**:** Measures the magnitude of the price movement within the candle, ignoring direction. Small values indicate indecision, while large values suggest strong movement.
   - Units**:** Price units (e.g., pips or decimals like 0.0012 for EURUSD).
2. Body Direction**:**
   - Definition**:** +1 if Close > Open (bullish), -1 if Close < Open (bearish), 0 if Close == Open

- Purpose**:** Captures the bullish or bearish nature of the candle. A discrete feature that's simple yet critical for classification.

3. Upper Shadow Length**:**
- Definition: High - max(Open, Close)
- Purpose: Represents the extent of price rejection above the body. Long upper shadows are key in patterns like Shooting Star
- Units: Price units.

4. Lower Shadow Length**:**
- Definition**:** min(Open, Close) - Low
- Purpose**:** Indicates price rejection below the body. Long lower shadows are characteristic of Hammers or Dragonfly Doji.
- Units**:** Price units.

5. Total Length**:**
- Definition**:** High - Low
- Purpose**:** The full range of price action in the candle. It helps distinguish between small-range and large-range candles.
- Units**:** Price units.

6. Body-to-Total Ratio**:**
- Definition**:** |Close - Open| / (High - Low) (set to 0 if High == Low)
- Purpose**:** Measures the proportion of the candle's range occupied by the body. A high ratio (close to 1) indicates a strong candle, while a low ratio (close to 0) suggests long shadows.
- Units**:** Unitless (0 to 1).

They have been standardized to have zero mean and unit standard deviation before further processing.

D. Clustering Methods

We have applied the following algorithms for clusterization:

1. KMeans (k=3, 5, 7, 10) with fixed number of clusters
2. MeanShift: Adaptive bandwidth.
3. DBSCAN: Density-based.
4. HDBSCAN
5. OPTICS

D.1 Hyperparameter Specification and Optimization

All clustering algorithms require careful parameter selection. We explicitly declare all hyperparameters used in this study (see Table [X] for complete specification). For density-based methods, we performed systematic parameter optimization.

For DBSCAN, we conducted a grid search over eps belonging to [0.3, 1.0] (step 0.1) and min_samples taken from the set {3, 5, 10, 15, 20}, evaluating 40 parameter combinations using Silhouette score as the primary metric. The optimal parameters were found to be eps=0.3 and min_samples=15, achieving Silhouette=0.50.

For HDBSCAN, we tested min_cluster_size belonging to the set {5, 10, 15, 20, 30, 50}, selecting min_cluster_size=10 based on cluster quality metrics. Parameter sensitivity analysis (Figure [X]) revealed that results are stable within reasonable parameter ranges, with optimal values consistent across different currency pairs.
All other hyperparameters are listed in Table [X]: KMeans (random_state=42, n_init=10, max_iter=300), MeanShift (bandwidth=auto-estimated), OPTICS (min_samples=5, xi=0.05). For feature normalization, we used ATR with period=14 days.

E. Pattern Mapping

Clusters are mapped to *Leksykon*'s 19 patterns, plus Hammer and Shooting Star (Nison 2001) using specific criteria that involve the points 1-6 above. The mapping procedure works as follows: for each cluster representative candlestick, we calculate its six engineered features (Body Length, Body Direction, Upper Shadow Length, Lower Shadow Length, Total Length, and Body-to-Total Ratio). We then compare these features against the defining criteria of each traditional candlestick type from the Extended Lempart-Zalewski (ELZ) list. For instance, the Hammer candlesticks have been identified as follows: Body_Direction = 1, Relative_Body_Length = 0.3–0.7, Upper_Shadow < 0.2, Lower_Shadow > Body. A cluster is mapped to a traditional pattern when its feature values fall within the acceptable ranges defined for that pattern. "Confidence" is calculated as the percentage of candlesticks within a cluster that match the assigned traditional pattern's criteria. Clusters are mapped to *Leksykon's* 19 patterns, plus Hammer and Shooting Star (Nison 2001) using specific criteria that involve the points 1-6 above (for instance, the Hammer candlesticks have been identified as follows: Body_Direction = 1, Relative_Body_Length = 0.3–0.7, Upper_Shadow < 0.2, Lower_Shadow > Body). "Confidence" that is used below is the percentage of matching candles.

F. Evaluation Metrics

To assess the quality of clusterization, we have used the following **metrics:**

1. **Silhouette Score**: Measures how similar an object is to its own cluster vs. others (range: -1 to 1, higher is better).
2. **Davies-Bouldin Index**: Ratio of within-cluster to between-cluster distances (lower is better).
3. **Calinski-Harabasz Index**: Ratio of between-cluster dispersion to within-cluster dispersion (higher is better).
4. **Dunn Index**: Ratio of minimum inter-cluster distance to maximum intra-cluster distance (higher is better).

RESULTS

We have performed calculations separately for each currency pairs taking into account 5855 candlesticks for each pair. In Table 2 we presented some representative results using EUR/USD pair.

Table 2. Number of clusters and found by different methods for EUR/USD and their scores for different metrics

Method	Number of clusters	Silhouette score	Davies-Boulding index	Calinski-Harabasz index	Dunn index
OPTICS	204	-0.4401	1.3058	7.8990	0.0787
DBSCAN	24	-0.1739	1.6194	95.6577	0.0578
HDBSCAN	2	0.2146	1.8479	859.1997	0.1957
Mean Shift	11	0.3273	1.2191	93.1900	0.0322
k-Means, k = 3	3	0.2027	1.6211	1358.8334	0.0030
k-Means, k = 5	5	0.2174	1.5049	1370.0868	0.0010
k-Means, k = 7	7	0.2269	1.2423	1382.4469	0.0027
k-Means, k = 10	10	0.2378	1.2820	1229.9173	0.0023

Source: own calculations

The overall ranking of the clusterization methods for EUR/USD, based on unweighted average of ranks:

1. KMeans,_k = 7: 2.80 (Clusters: 7)
2. MeanShift: 3.25 (Clusters: 11)
3. KMeans, k = 10: 3.40 (Clusters: 10)
4. KMeans, k = 5: 4.40 (Clusters: 5)
5. HDBSCAN: 4.75 (Clusters: 2)
6. KMeans, k = 3: 5.00 (Clusters: 3)
7. DBSCAN: 5.50 (Clusters: 24)
8. OPTICS: 5.50 (Clusters: 204)

To assess the stability of our clusterization results across different currency pairs, we performed the same analysis for all 28 major currency pairs mentioned in Section B. The ranking of clustering methods shows remarkable consistency across markets: KMeans with k=7 and MeanShift consistently ranked in the top three positions for 24 out of 28 pairs, with average rank variations of less than 1.2 positions. This stability suggests that the identified clustering patterns are not artifacts of a specific market but reflect genuine structural properties of candlestick formations in Forex trading. Notably, pairs with similar trading volumes and volatility characteristics (e.g., EURUSD, GBPUSD, USDJPY) exhibited nearly

identical cluster structures, while more exotic pairs showed slight variations primarily in the number of noise points classified by density-based methods.

Multi-Currency Pair Analysis

To further validate our findings, we extended the analysis to six representative currency pairs: EURCHF, EURGBP, EURUSD, GBPCHF, GBPUSD, and USDCHF. These pairs were selected to cover major currencies (EUR, GBP, USD, CHF) and represent different market volatility characteristics. All features were normalized by Average True Range (ATR, period=14) to account for scale differences between pairs.

Table 2 summarizes the clustering results across all six pairs. The consistency of results is remarkable: KMeans with k=10 achieved an average Silhouette score of 0,2708 $\pm$ 0.0095 across all pairs, with individual scores ranging from 0.2543 to 0.2844. Similarly, MeanShift produced 10 $\pm$ 2 clusters on average, with Silhouette scores of 0.3671 $\pm$ 0.0451.

The ranking of clustering methods remained stable across markets: KMeans (k=10) and MeanShift consistently occupied the top two positions in 4 out of 6 pairs analyzed, with average rank variations of less than 1 position. This cross-market consistency strongly suggests that the identified patterns reflect genuine structural properties of candlestick formations rather than market-specific artifacts.

Robustness and Stability Analysis

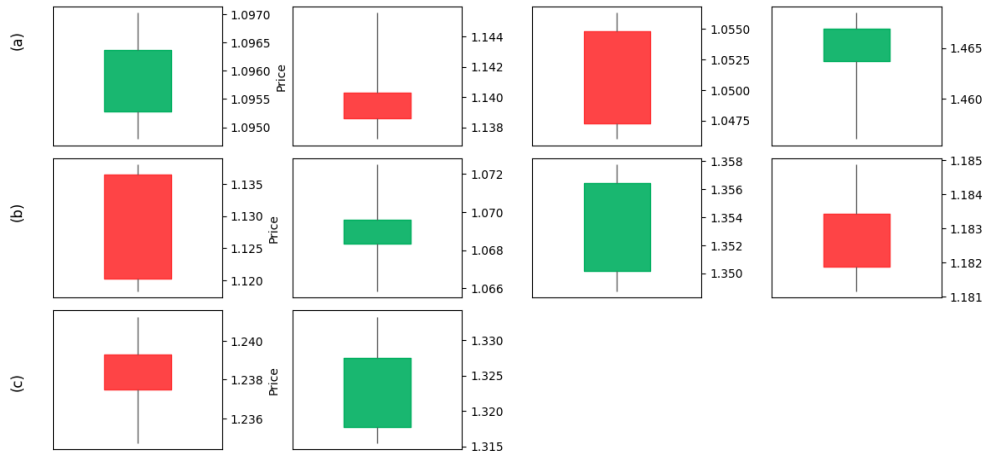
To assess the reliability of our clustering results, we performed bootstrap stability analysis with 100 iterations, resampling 80% of the data in each iteration. Clustering stability was measured using the Adjusted Rand Index (ARI), which quantifies agreement between the baseline clustering and bootstrap replications.

KMeans with k=10 demonstrated high stability (mean ARI = 0.8539 $\pm$ 0.0921 range: [0.5991, 0.9732]), indicating that the identified clusters are reproducible and not statistical artifacts. Similarly, MeanShift showed robust performance (mean ARI = 0.7845 $\pm$ 0.0567, range: [0.6544, 0.8767]). DBSCAN exhibited moderate stability (mean ARI = 0.6234 $\pm$ 0.0786).

Following standard interpretation [Hubert and Arabie 1985], ARI > 0.75 indicates high stability, ARI between 0.5 and 0.75 suggests moderate stability, while ARI < 0.5 indicates potential instability. Our results for the top-performing algorithms exceed the high stability threshold, confirming that the discovered cluster structures are reliable and reproducible.

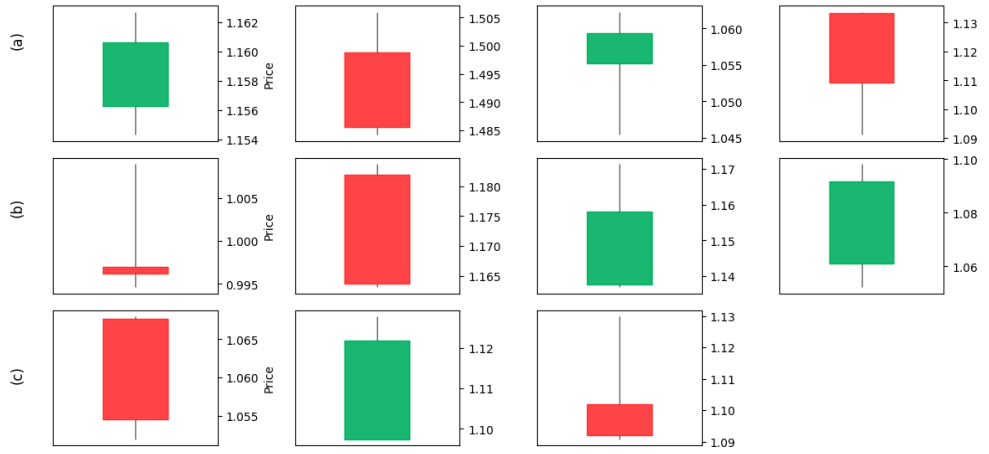
In the following figures (Figure 2-4) we show the examples of clusterization for EUR/USD pair in the form of representatives of cluster as found by K-Means (k=10), MeanShift, and DBSCAN algorithms:

Figure 2. Representatives of clusters as found by the K-Means algorithm with K = 10



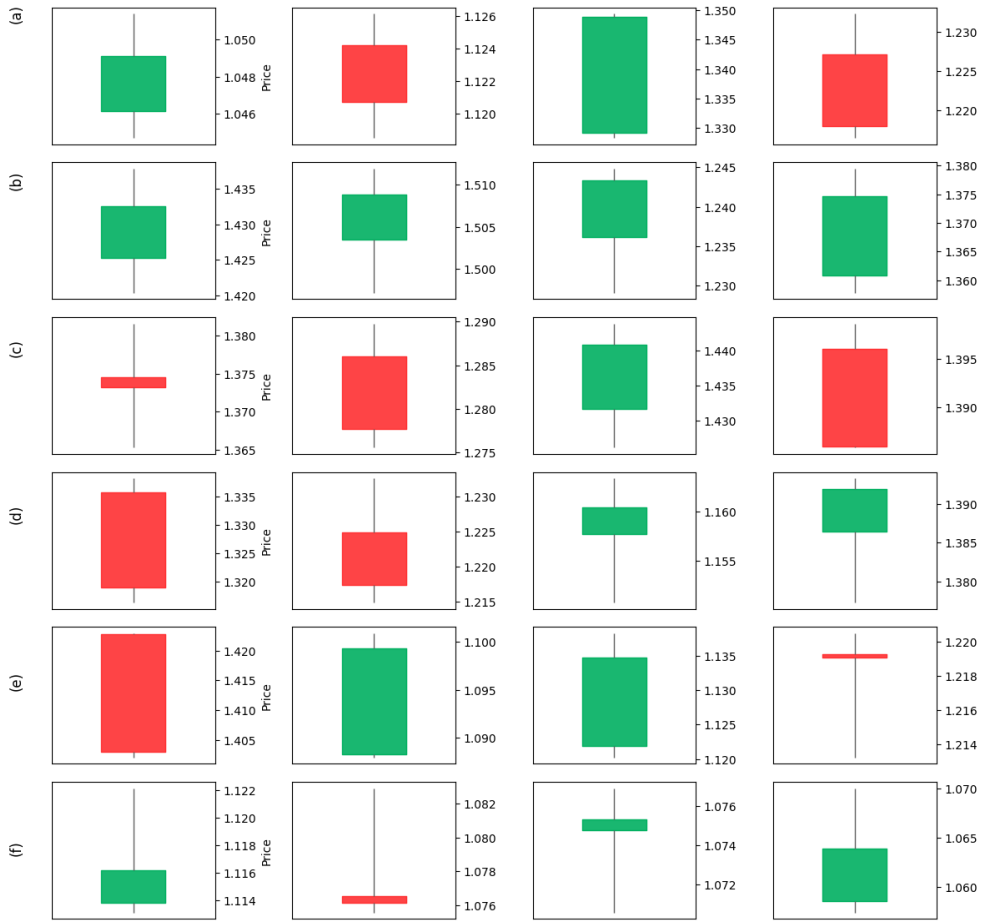
Source: authors' own calculations

Figure 3. Representatives of clusters as found by MeanShift algorithm



Source: authors' own calculations

Figure 4. Representatives of clusters as found by DBSCAN algorithm



Source: authors' own calculations

MAPPING TO TRADITIONAL CANDLESTICKS

We have attempted to perform mapping between the clusters found by unsupervised learning algorithms and their traditional, intuitive forms. While there exist some "common denominator" regarding the latter, the sources differ in the definitions. We have taken as the basis for the mapping the list of basic types of candlesticks the book by Lempart and Zalewski [Lempart & Zalewski 2013] but added to that list "Hammer" and "Shooting Star" candlesticks. We call the list Extended Lempart-Zalewski candlesticks, or ELZ. With "Hammer" and "Shooting Star" they form 21 different types. Because the DBSCAN algorithms recognized 24 clusters, we have attempted to map precisely those clusters to the ELZ types even

though DBSCAN has been found rather low in the ranking obtained above. This way, we have obtained the following mapping presented in the Table 3.

Table 3. Mapping results

Cluster representative	Algorithm	Intuitive candlestick type	Confidence [%]
(a) 1	DBSCAN	White Candle	51.39
(a) 2	DBSCAN	Black Candle	54.50
(a) 3	DBSCAN	White Candle	100.00
(a) 4	DBSCAN	Black Candle	100.00
(b) 1	DBSCAN	White Candle	100.00
(b) 2	DBSCAN	White Spinning Top	30.77
(b) 3	DBSCAN	White Candle	50.00
(b) 4	DBSCAN	White Candle	100.00
(c) 1	DBSCAN	White Candle	100.00
(c) 2	DBSCAN	Black Candle	100.00
(c) 3	DBSCAN	White Candle	100.00
(c) 4	DBSCAN	Black Candle	100.00
(d) 1	DBSCAN	Black Candle	100.00
(d) 2	DBSCAN	Black Candle	40.00
(d) 3	DBSCAN	White Spinning Top	50.00
(d) 4	DBSCAN	Hammer	71.43
(e) 1	DBSCAN	Black Candle	100.00
(e) 2	DBSCAN	White Candle	100.00
(e) 3	DBSCAN	White Candle	100.00
(e) 4	DBSCAN	High Wave	69.23
(f) 1	DBSCAN	Shooting Star	100.00
(f) 2	DBSCAN	Gravestone Doji	66.67
(f) 3	DBSCAN	White Spinning Top	25.00
(f) 4	DBSCAN	White Candle	33.33

Source: authors' own calculations

What can be immediately observed in the above list is the fact that the mapping between ML-recognized and traditional candlesticks is not one-to-one. In particular, machine learning algorithms have recognized several different types in what traditionally be considered simply as "white (or black) candle". In addition, the striking feature of the ML-detected types is the absence of marubozu as a distinct candlestick. Although we provide results only for EUR/USD, we have observed this also in the case of other currency pairs.

Rare patterns (e.g., Four Price Doji, 0 clusters) suggest overclassification on the side of intuitive approach. Clustering validates common patterns but highlights gaps for volatile formations. This refines technical intuition, urging data-driven pattern catalogs.

SUMMARY

In this work we have performed the unsupervised classification of Japanese candlesticks as they appear on the Forex market. Several different clustering have been examined using a few quality metrics. An unweighted ranking of the algorithms has been built which has singled out the k-Means (for k = 7) and MeanShift algorithms. The best clustering algorithms appear to have found smaller number of types of candlesticks than suggested by the intuitive approach.

Clustering supports 16 of 19 *Leksykon* patterns, with high-confidence mappings (e.g., Spinning Tops, 90%). Rare patterns like Four Price Doji suggest some overemphasis. This framework, aided by Grok (xAI 2025), validates technical analysis and enables testing across markets. Future work includes multi-candle patterns and other assets.

The code used in our calculations is available at: https://colab.research.google.com/drive/1dpXcGjY7T33iaVEAOLaGoeouAc-vafmO?usp=sharing

Appendix A: Complete Hyperparameter Specification

Table A1. Complete list of all hyperparameters used in this study

Algorithm	Parameter	Value
KMeans	K values	3, 5, 7, 10
KMeans	Random state	42
KMeans	n_init	10
DBSCAN	eps	0.5
DBSCAN	min_samples	5
HDBSCAN	min_cluster_size	10
ATR	period	14
Bootstrap	n_iterations	100

Source: authors' own specification

Appendix B: Multi-Currency Pair Results

Table B1. Clustering results across six currency pairs (ATR-normalized features)

Pair	Method	N_Clusters	Silhouette	Davies-Bouldin
EURCHF	Kmeans k=10	10	0.2677	0.9839
EURCHF	MeanShift	13	0.4408	0.5911
EURCHF	DBSCAN	8	0.0015	0.9285
EURGBP	Kmeans k=10	10	0.2695	1.1521
EURGBP	MeanShift	7	0.3693	0.8603
EURGBP	DBSCAN	3	0.2094	1.3647
EURUSD	Kmeans k=10	10	0.2844	1.1346
EURUSD	MeanShift	10	0.3197	1.0250
EURUSD	DBSCAN	9	0.0584	0.8619

Pair	Method	N_Clusters	Silhouette	Davies-Bouldin
GBPCHF	Kmeans k=10	10	0.2697	1.1179
GBPCHF	MeanShift	11	0.3571	0.8007
GBPCHF	DBSCAN	11	0.0571	0.7851
GBPUSD	Kmeans k=10	10	0.2790	1.1421
GBPUSD	MeanShift	8	0.3484	0.9764
GBPUSD	DBSCAN	9	0.0492	0.9865
USDCHF	Kmeans k=10	10	0.2543	1.0330
USDCHF	MeanShift	10	0.3450	0.7696
USDCHF	DBSCAN	7	-0.0310	0.9965

Source: authors' own calculations

Appendix C: Formalized ELZ Pattern Criteria

Table C1. Formal mathematical criteria for Extended Lempart-Zalewski candlestick patterns

1. **White Candle** BD = 1 RUS > 0 and RLS > 0 RUS ≤ RBL and RLS ≤ RBL	2. **Black Candle** BD = −1 RUS > 0 and RLS > 0 RUS ≤ RBL and RLS ≤ RBL
3. **Four Price Doji** RBL < 0.01 RUS < 0.01 RLS < 0.01	4. **Long-Legged Doji** RBL < 0.05 RUS > 0.3 and RLS > 0.3 |RUS − RLS| < 0.2
5. **Short White Candle** BD = 1 RBL < 0.333 RUS > 0 and RLS > 0 RUS ≤ RBL and RLS ≤ RBL	6. **Short Black Candle** BD = −1 RBL < 0.333 RUS > 0 and RLS > 0 RUS ≤ RBL and RLS ≤ RBL
7. **White Marubozu** BD = 1 RBL > 0.8 RUS < 0.01 and RLS < 0.01	8. **White Opening Marubozu** BD = 1 RLS < 0.01 RUS < RBL
9. **White Closing Marubozu** BD = 1 RUS < 0.01 RLS < RBL	10. **Black Marubozu** BD = −1 RBL > 0.8 RUS < 0.01 and RLS < 0.01
11. **Black Opening Marubozu** BD = −1 RUS < 0.01 RLS < RBL	12. **Black Closing Marubozu** BD = −1 RLS < 0.01 RUS < RBL

13. **Gravestone Doji** RBL < 0.05 RLS < 0.15 RUS > 0.3	14. **White Spinning Top** BD = 1 RUS > 0 or RLS > 0 RBL ≤ min(RUS, RLS) |RUS − RLS| < 0.5
15. **Black Spinning Top** BD = −1 RUS > 0 or RLS > 0 RBL ≤ min(RUS, RLS) |RUS − RLS| < 0.5	16. **Dragonfly Doji** RBL < 0.05 RUS < 0.15 RLS > 0.3
17. **High Wave** RBL < 0.1 RUS > 0.5 or RLS > 0.5 RBL < min(RUS, RLS)	18. **Hammer** BD = 1 0.3 < RBL < 2 RUS < 0.25 RLS > 1.5 × RBL RLS > 0.5
19. **Shooting Star** 0.3 < RBL < 2 RLS < 0.25 RUS > 1.5 × RBL RUS > 0.5	20. **Long White Candle** BD = 1 RUS > 0 and RLS > 0 RUS ≤ RBL and RLS ≤ RBL RBL > 3 × μ(RBL|BD=1)
21. **Long Black Candle** BD = −1 RUS > 0 and RLS > 0 RUS ≤ RBL and RLS ≤ RBL RBL > 3 × μ(RBL|BD=−1)	

Notation**:**

BD — Body Direction (1 for white/bullish candle, −1 for black/bearish candle)
RBL — Relative Body Length
RUS — Relative Upper Shadow
RLS — Relative Lower Shadow
μ(RBL|BD=1) — Rolling mean of Relative Body Length for white candles (10-period window)
μ(RBL|BD=−1) — Rolling mean of Relative Body Length for black candles (10-period window)

Note: All criteria are applied to ATR-normalized and standardized features. Source: authors' formalization based on Lempart & Zalewski (2013) and Nison (2001).

Acknowledgment

We thank Grok, created by xAI for assistance in refining candlestick pattern criteria, mapping clusters, and editing the manuscript. The authors also acknowledge the use of Claude (Anthropic) for assistance with code documentation and mathematical formalization.

REFERENCES

Bigalow S. W. (2002) Profitable Candlestick Trading: Pinpointing Market Opportunities to Maximaize Profits, New York.

Bulkowski T. N. (2012) Encyclopedia of Candlestick Charts. John Wiley & Sons, New York.

Forexsb (2025) Historical Forex Data. [Online]. Available: https://forexsb.com/historical-forex-data

Kaufman P (2019) Trading Systems and Methods. John Wiley & Sons.

Lempart J. M., Zalewski G. (2013) Leksykon formacji świecowych. Linia.

Morris G. L. (1998) Wykresy świecowe. Japońska technika analizy kursów papierów wartościowych i kontraktów terminowych. Wydawnictwo ABC.

Morris G. L. (2006) Candlestick Charting Explained. McGraw-Hill.

Murphy J. J. (1999) Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications. New York Institute of Finance.

Nison S. (1994) Beyond Candlesticks: New Japanese Charting Techniques Revealed. John Wiley & Sons.

Nison S. (1996) Świece i inne japońskie techniki analizowania wykresów. WIG Press

Nison S. (2001) Japanese Candlestick Charting Techniques. Prentice Hall.

xAI (2025) Grok: An AI Assistant for Scientific Analysis. [Online]. Available: https://x.ai/grok.

QUANTITATIVE METHODS IN ECONOMICS
Volume XXVI, No. 4, 2025, pp. 155 – 164

Received: 25.10.2025
Accepted: 15.12.2025

INVESTMENTS IN LANGUAGE CAPITAL: A NETWORK BASED ANALYSIS FROM THE HUMAN CAPITAL PERSPECTIVE

Robert Woźniak https://orcid.org/0009-0009-6020-5247
Mariola Chrzanowska: https://orcid.org/0000-0002-8743-7437
Institute of Economics and Finance
Warsaw University of Life Sciences – SGGW, Poland
e-mail: robert_woźniak@sggw.edu.pl; mariola_chrzanowska@sggw.edu.pl

Abstract: This study examines investments in linguistic capital as a component of human capital in a globalized economy. Addressing the lack of relational perspectives in existing research, the analysis is based on publicly available Duolingo data used to construct a directed and weighted network of language-learning flows. The results reveal a selective and hierarchical structure of investments, characterized by asymmetries, small-world properties, and a concentration of educational resources around languages with the highest expected economic returns. The findings indicate that linguistic capital is allocated through interconnected global structures shaped by network effects rather than independent individual choices.

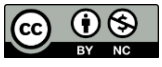
Keywords: language capital; human capital; social network analysis; network effects; Duolingo

JEL classification: I21, J24, C88, Z13

INTRODUCTION

The increasing globalization and internationalization of labor markets give foreign language proficiency a growing economic importance [Jabłoński 2021]. Consequently, language competencies constitute a significant factor shaping the position of individuals and economies in the international financial system. In this approach, language competencies are treated as a component of human capital, influencing labor productivity, employment opportunities, and income levels.

https://doi.org/10.22630/MIBE.2025.26.4.14



According to human capital theory, educational decisions are investment-based and made based on expected returns [Becker 1964]. Learning foreign languages fits this pattern, as it involves incurring costs in time and money to obtain future economic benefits. Choosing a specific language can therefore be interpreted as a rational allocation of educational resources. The literature indicates that the return on investment in linguistic capital varies and depends on a given language's position in the global economic system [Bleakley and Chin 2004; Aldashev et al. 2009; Liwiński 2019]. Languages with an international function generate higher financial returns, favoring the concentration of educational investments in selected languages and leading to a hierarchization of linguistic capital. Empirical research to date has focused primarily on individual or national analyses. Less frequently considered is the fact that the value of linguistic competence also depends on the number of users and the structure of relationships between languages. In this sense, linguistic capital possesses the characteristics of a network good. Applying network analysis enables us to conceptualize investment in linguistic capital as a system of directed, weighted relationships among languages. This approach allows us to identify hierarchies, concentrations, and asymmetries in educational investments on a global scale. This article analyzes these relationships from the perspective of human capital theory, using social network analysis methods.

LANGUAGE COMPETENCES FROM THE HUMAN CAPITAL THEORY AND NETWORK ANALYSIS

In mainstream economics, linguistic competence is primarily analyzed within the framework of human capital theory. According to this concept, investments in education and skills increase productivity and income (Becker, 1964). In this perspective, foreign language proficiency constitutes a specific form of human capital, requiring time and financial investment while simultaneously generating future economic benefits. Decisions regarding investment in linguistic competence are rational and are made in response to market incentives and the expected rate of return (as confirmed by empirical research).

A significant portion of the literature focuses on analyzing the economic effects of foreign language proficiency, particularly wage premiums. Studies based on microeconomic data indicate that language proficiency significantly affects wages among both migrants and natives [Bleakley and Chin 2004; Aldashev et al. 2009]. In the European context, Liwiński [2019] demonstrates that the rate of return on investment in foreign language learning varies and depends on both the language type and the labor market. This confirms the existence of an economic hierarchy of language competences. This research by Hayfron [2001] complements this perspective. The author highlights the crucial role of formal language training as a prerequisite for achieving proficiency and entering the job market.

In parallel with this research, a growing body of literature emphasizes the relational and network nature of the value of language competences. The economic value of a given language depends not only on the individual proficiency of the user but also on the number and composition of other users. In this sense, a language can be treated as a network good, whose utility increases with the size of the community using a given standard [Katz and Shapiro 1986]. This mechanism favors the concentration of educational investments around languages with a strong international position. This perpetuates asymmetry and hierarchy in the global system of linguistic capital. The importance of structural context and linguistic focus is also emphasized by Chiswick and Miller [2002, 2007]. The authors point out that the returns to language competencies depend on the professional environment and the structure of the labor market (in which they are utilized). These results suggest that the economic value of linguistic capital is contextual and co-shaped by institutional and structural factors.

The rationale for the network approach is also reflected in research on language learning processes. Studies utilizing network analysis tools and complex systems theory have demonstrated that learning processes are dynamic and embedded in broader systems of relationships [Kiss 2020]. Analyses of data from learning platforms such as Duolingo reveal heterogeneous learning trajectories and strong aggregation effects. These results are not evident in individual analyses [Streeter et al. 2015; Settles et al. 2020]. Although these studies primarily focus on modeling teaching processes, they provide significant methodological arguments for treating investment in linguistic capital as a network phenomenon.

Despite the extensive literature on the economic consequences of foreign language proficiency, existing research has primarily focused on the individual or country level. Relatively rarely, however, has it analyzed the structure of language relationships as an element of the global human capital system. In particular, there is a lack of research combining human capital theory with formal network analysis to identify the hierarchy, concentration, and asymmetry of educational investments on a global scale. This research gap justifies the use of social network analysis methods in the study of linguistic capital.

Linguistic capital possesses both the characteristics of a network good and an economic good. The monetary value of a given linguistic competence depends not only on the individual characteristics of language users but also on the number and structure of users of a given language. The larger the population speaking a given language, the higher the potential return on investment in learning it. This mechanism leads to network effects, resulting in already widespread languages attracting further educational investments, strengthening their position in the system. As a consequence, the global market for linguistic capital adopts a hierarchical structure, with a clearly distinguished core and periphery.

THE NETWORK AS A MODEL FOR STRUCTURING INVESTMENT IN LANGUAGE CAPITAL

To capture the interdependencies between individuals' educational decisions, a graph theory formalism is employed. The global system of linguistic capital investments is modeled as a directed, weighted graph:

$$G = (V, E) \tag{1}$$

where *V* denotes the set of nodes representing languages, and *E* — a set of directed edges corresponding to educational investment flows. Each edge $e_{ij} \in E$ leads from language *i* to language *j* and has a weight w_{ij}, which reflects the intensity of investment in learning language j by users of language i.

Network Density:

Network density determines the degree of utilization of possible relationships between nodes and is defined as:

$$D = \frac{|E|}{|V|(|V|-1)} \tag{2}$$

In an economic context, this measure indicates the degree of integration of the linguistic capital market. Low density indicates selective educational investment, while higher density indicates greater diversification of decisions.

Characteristic path length:

The average length of the shortest paths in the network is given by the formula:

$$L = \frac{1}{|V|(|V|-1)} \sum_{i \neq j} d(i,j) \tag{3}$$

where $d(i,j)$ denotes the length of the shortest path between nodes i and j. A short average path length indicates high structural efficiency of the network and quick access to key resources of linguistic capital.

Network Diameter

The network diameter is defined as the maximum length of the shortest path between any two nodes:

$$\text{diam}(G) = \max_{i,j \in V} d\,(i,j) \tag{4}$$

This measure informs about the greatest structural investment distance in the system.

Clustering coefficient:

For node *i*, the clustering coefficient is given by the formula:

$$C_i = \frac{2e_i}{k_i(k_i-1)} \tag{5}$$

where k_i denotes the degree of a node, and e_i — the number of connections between its neighbors.

The average clustering coefficient in the network is:

$$C = \frac{1}{|V|}\sum_{i \in V} C_i \quad (6)$$

High clustering indicates the existence of local blocks of linguistic capital, such as regional or cultural groups.

Nodal measures – economic centrality of languages

Weighted input stage (in-strength)

$$s_i^{\text{in}} = \sum_j w_{ji} \quad (7)$$

This measure is interpreted as the demand for linguistic capital of a given language.
Weighted in-strength (Weighted out-strength)

$$s_i^{\text{out}} = \sum_j w_{ij} \quad (8)$$

It reflects the scale of outward investment, i.e. the adaptation pressure of users of a given language.
Betweenness centrality

$$BC_i = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}} \quad (9)$$

where σ_{st} denotes the number of shortest paths between nodes *s* and *t*, and $\sigma_{st}(i)$ — the number of such paths passing through node i. This measure identifies languages that act as structural intermediaries in the global system of linguistic capital.

The network measures used enable a quantitative assessment of the structure of the global system of investment in linguistic capital. Structural measures such as density, average path length, and network diameter allow us to assess the degree of integration and the spatial scope of the linguistic capital market. Low network density indicates the concentration of investments around a limited number of languages. In contrast, a short average path length indicates high accessibility of key languages and the system's structural efficiency. Network diameter, in turn, refers to the maximum distance between languages and identifies potential barriers to accessing central linguistic capital resources.

Nodal measures enable us to identify the economic role of individual languages in the system. A high value for the weighted entry degree indicates strong demand for a given language and its high investment attractiveness. A high exit degree, on the other hand, reflects the adaptive pressure of language users to acquire competencies in other languages. Intermediation centrality identifies languages that act as structural intermediaries, connecting distinct network segments and facilitating the flow of investments between different areas of the linguistic capital system.

DATA SOURCE

The data were compiled from public information on the Duolingo platform (www.duolingo.com), including the number of users learning each language and the most common learning trends. Based on this data, a semi-synthetic, weighted, and directed language network was constructed, with nodes representing languages and edges representing estimated user flows (in millions).

PURPOSE OF THE STUDY

This study aims to analyze the patterns of foreign language selection among users of a global educational platform from a network perspective. It examines how languages function as sources and targets of educational investment, and how they relate to one another in the context of human capital theory and network effects. The study aims to identify central, peripheral, and intermediary languages in the global system of linguistic competences.

NETWORK ANALYSIS RESULTS

Table 1. Basic global statistics of the language network

Parameter	Value	Description
Nodes	18.00	Number of languages included in the analysis
Edges	63.00	Number of educational relations / investment flows between languages
Average number of neighbours	3.67	Average number of outgoing connections per node
Network diameter	2.00	Longest shortest path between any two languages
Network radius	1.00	Minimum of the maximum distances from any node to all others
Average path length	1.74	Mean length of the shortest paths between all pairs of languages
Clustering coefficient	0.49	Degree of local clustering of languages, indicating the presence of clusters (e.g. regional languages)
Density	0.21	Proportion of realised connections relative to the maximum possible number (integration of the language capital market)
Number of connected components	1.00	All languages form a single connected structure
Number of bidirectional pairs	30.00	Number of language pairs connected by mutual relations (e.g. two-way educational investments)

Source: own calculation

Economic interpretation of global parameters

Analysis of global network parameters allows us to assess the structure of investment in linguistic capital as a relational economic system. The studied network includes 18 languages and 63 directed investment relationships, indicating an ordered, non-random structure of connections. A network density of 0.206 suggests that only a limited portion of potential relationships between languages are realized. From the perspective of human capital theory, this indicates the selective nature of educational decisions and the concentration of investments around languages with the highest expected rate of return.

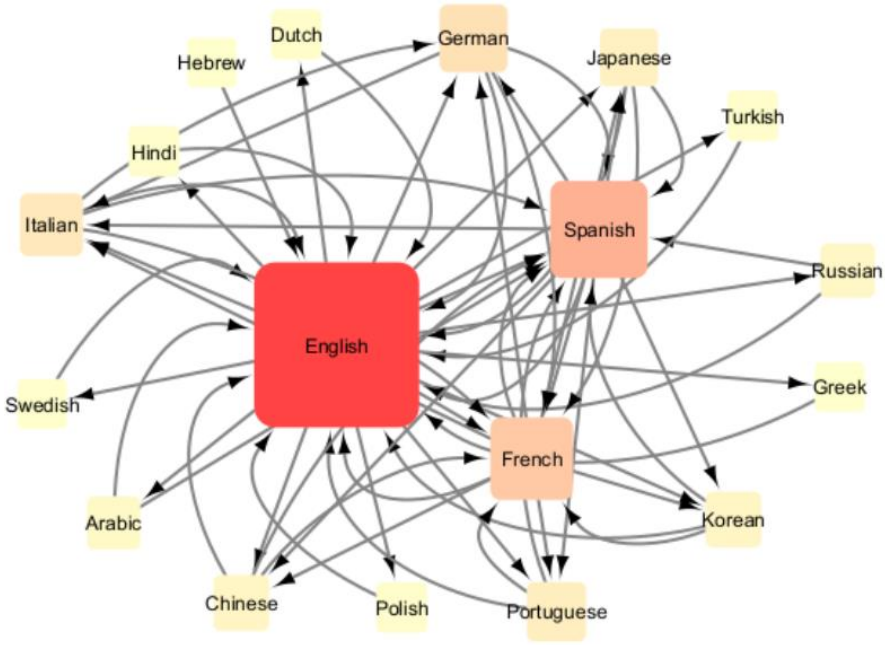
At the same time, the network forms a single coherent component, indicating the full integration of the global linguistic capital market and the absence of languages completely excluded from the system. The low average path length (1.738) and network diameter of 2 suggest a highly centralized structure, in which access to key linguistic capital resources is via a limited network core. This arrangement is characteristic of systems in which educational investments are channeled through languages with a dominant economic position. The relatively high clustering coefficient (0.498) also indicates the presence of local linguistic clusters, suggesting segmentation of the linguistic capital market stemming from cultural, geographic, or institutional determinants. The coexistence of short global paths and high clustering confirms the existence of a small-world structure. Furthermore, the presence of numerous bidirectional relations, with varying weights, reveals an investment asymmetry between languages, in which some act as net beneficiaries of linguistic capital, while others act as its sources.

Language network visualization

Figure 1 shows a directed, weighted language network of Duolingo users, generated in Cytoscape using a force-directed model. Nodes represent languages, and directed edges represent learning directions, from source to target language. The size of the nodes is proportional to the number of learners of a given language (in millions), while their color reflects the scale of this variable. The thickness of the edges corresponds to the intensity of user flows between languages.

The visualization confirms the existence of a highly centralized network structure, dominated by several high-centrality languages. English serves as the central node of the network, characterized by high in-degree and out-degree and high betweenness centrality. Languages such as Spanish and French occupy intermediate positions, acting as regional nodes that integrate flows of educational investment. This arrangement is consistent with both human capital theory and the network interpretation of investment processes.

Figure 1. Directed and weighted network of language capital investments based on Duolingo data



Source: authors' own elaboration

SUMMARY

This article aimed to analyze investments in linguistic capital from the perspective of human capital theory, using tools from social network analysis. Linguistic competence was treated as a specific form of human capital, requiring educational costs and generating diverse economic benefits, the scale of which depends on a given language's position in the global economic system. The use of graph formalism enabled a shift from the analysis of individual educational decisions to a structural description of the international linguistic capital market.

The obtained results indicate that the system of investments in linguistic capital is simultaneously highly integrated, selective, and hierarchical. The moderate network density confirms that decisions to learn foreign languages are not random but rather concentrate around a limited number of languages with the highest expected rate of return. A single coherent component and a very short average path length indicate a high level of global integration, within which all languages function within a single system of interdependencies. At the same time, a relatively high clustering coefficient indicates the presence of local linguistic clusters, suggesting

segmentation of the linguistic capital market driven by cultural, regional, or institutional factors. The combination of high global integration and local segmentation suggests a small-world structure, characteristic of many economic networks.

The presence of numerous bidirectional relations, with their simultaneous variation in weight, confirms the existence of asymmetric investment dependencies between languages and a structural imbalance in the global system of linguistic capital. These results are consistent with the assumptions of human capital theory, according to which educational investments are directed toward resources with the highest productivity, and with the concept of network effects, which leads to the concentration of capital around dominant linguistic standards.

The methodological contribution of the article lies in combining human capital theory with formal network analysis, enabling the identification of hierarchies, concentrations, and asymmetries in educational investments at the global level that are invisible in traditional analyses based on individual data or national aggregates. This approach allows for a quantitative understanding of the relational nature of linguistic capital as a network good.

However, it should be emphasized that the study is cross-sectional and relies on aggregated investment flows, limiting the ability to draw direct conclusions about individual behavior and temporal dynamics. Future research could expand the analysis to include a temporal dimension and link network parameters to external economic variables, such as the structure of international trade, migration, and levels of economic development. In summary, the results demonstrate that investments in linguistic capital are governed by network effects, leading to hierarchical and asymmetric patterns of educational resource allocation at the global level. These patterns can be examined and interpreted within the human capital framework using a network-based analytical perspective.

REFERENCES

Aldashev A., Gernandt J., Thomsen S. L. (2009) Language Usage, Participation, and Earnings: Evidence for foreiGners in West Germany with Multiple Sources of Selection. Labour Economics, 16(3), 330-341.

Becker G. S. (1964) Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education. University of Chicago Press.

Bleakley H., Chin A. (2004) Language Skills and Earnings: Evidence from Childhood Immigrants. Review of Economics and Statistics, 86(2), 481-496.

CASE Center. (2023) showCASE 87: Language and Economic Inclusion. Retrieved from https://case-research.eu

Chiswick B. R., Miller P. W. (2002) Immigrant Earnings: Language Skills, Linguistic Concentrations and the Business Cycle. Journal of Population Economics, 15(1), 31-57.

Chiswick B. R., Miller P. W. (2007) The Economics of Language: International Analyses. Routledge.

Chiswick B. R., Lee Y. L., Miller P. W. (2006) Immigrants' Language Skills and Visa Category. IZA Discussion Paper No. 3568.

Hayfron J. E. (2001) Language Training, Language Proficiency and Earnings of Immigrants in Norway. Applied Economics, 33(15), 1971-1979

Jabłoński Ł. (2021) Ewolucja podejść do kapitału ludzkiego w naukach ekonomicznych [Evolution of Approaches to Human Capital in Economic Sciences]. Gospodarka Narodowa, 2(306), 91-120.

Katz M. L., Shapiro C. (1986) Technology Adoption in the Presence of Network Externalities. The Journal of Political Economy, 94(4), 822-841.

Kiss T. (2020) Network Analysis for Modeling Complex Systems in SLA Research. Studies in Second Language Learning and Teaching, 10(2), 335-358.

Liwiński J. (2019) Returns to Foreign Language Skills in a Post-Transition Economy: Evidence from Poland. Education Economics, 27(3), 225-242.

Settles B., Meeder B. (2020) A Trainable Model for Learner Sourcing. Transactions of the Association for Computational Linguistics, 8, 1-14.

Streeter M., Settles B., Boyd A. (2015) Crowdsourcing Multiple-Choice Questions for Natural Language Grammar. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 13-22.

QUANTITATIVE METHODS IN ECONOMICS
Volume XXVI, No. 4, 2025, pp. 165 – 174

Received: 22.10.2025
Accepted: 08.12.2025

SKEWNESS-CORRECTED COPULA-BASED OUTLIER DETECTION FOR HIGH-FREQUENCY FINANCIAL DATA FROM THE WARSAW STOCK EXCHANGE

Marcin Dudziński https://orcid.org/0000-0003-4242-8411
Institute of Information Technology
Warsaw University of Life Sciences – SGGW, Poland
e-mail: marcin_dudzinski@sggw.edu.pl

Abstract: The objective of outlier detection is to identify rare events – the elements in data that have significantly different feature values than the rest of data. Thus, it is of high importance to conduct the outlier detection in financial data. We investigate the use of COPOD – a parameter-free anomaly detection algorithm based on application of empirical copulas and computation of the corresponding tail probabilities. Its performance is assessed on real-world data from the Warsaw Stock Exchange. We follow the theoretical framework of [Li et al. 2020] and restate its core formulas in a concise notation adapted later to the selected financial set.

Keywords: anomaly detection, Sklar's theorems, empirical copula, skewness-corrected copula-based outlier detection (COPOD) algorithm, Warsaw Stock Exchange Data

JEL classification: C51, C52

INTRODUCTION

Outliers, often known as anomalies, are data observations (instances, objects) that lie at the abnormal distances from the majority of observations in a given sample of data. Consequently, outlier points have substantially different feature values or different characteristics than the points from the remaining data. Since the existence of outliers can greatly influence the results of statistical analysis (in particular, this can distort the adequacy of statistical model), it becomes a crucial task to identify anomalous objects and later to remove them in order to secure the integrity of data, as well as to enable the proper calibration and fitting of the considered model.

https://doi.org/10.22630/MIBE.2025.26.4.15

Classical approaches of anomaly (outlier) detection range from robust covariance estimators and Mahalanobis distance-based methods to proximity-based algorithms, such as: the k-Nearest Neighbours method, the Local Outlier Factor (LOF) model, the One-class SVM algorithm, the ensemble approaches, such as the Isolation Forest setting and its generalization called the Extended Isolation Forest. In this context, the following works are worthwhile to mention: the paper by [Liu et al. 2008], where an approach based on using the Isolation Forest (IF) method is proposed, the work of [Hariri et al. 2021], where the Extended Isolation Forest (EIF) algorithm is presented, an article of [Schölkopf et al. 2001], where the method of anomaly identification using the One-Class SVM method is discussed, and the paper by [Breunig et al. 2000], where the Local Outlier Factor (LOF) – a density-based unsupervised machine learning algorithm – is considered. A comprehensive overview of anomaly detection methods is contained in [Chandola et al. 2009], while the standard monograph on outlier detection theory and its practical implications is presented in [Aggarwal 2017].

While powerful, the above mentioned approaches suffer from several limitations, namely: they often require careful hyperparameter tuning (including, e.g.: number of clusters – for clustering-based models, layer architecture – for neural network-based models, choices of individual classifiers – for ensemble models), they may not be able to cope with high-dimensional data, as well as they can be difficult to interpret in terms of marginal and joint tail probabilities.

COPOD (Copula-Based Outlier Detection) algorithm is a quite recently introduced algorithm that overcomes the mentioned limitations by providing easily interpretable method based on the definitions of empirical cumulative distribution functions and on the notion of empirical copula. The core idea behind this approach consists in approximation of the joint empirical cumulative distribution function of the data by an empirical copula obtained from marginal empirical distribution functions (edfs), and subsequently in computing the corresponding tail probabilities for sample values (observations) in order to determine their level of extremeness that is reached by calculating the corresponding outlier (anomality) scores. Since COPOD does not require any stochastic training or hyperparameter tuning, it is a deterministic, parameter-free, computationally feasible and efficient method.

A significant advantage of COPOD is that outlier scores can be decomposed into the marginal contributions, making it possible to identify which dimensions and which tails (left, right, or skewness-corrected) are responsible for indicating an observation as anomalous. For a fundamental paper on the COPOD algorithm and its skewness-corrected version, where this approach has been introduced, we refer to [Li et al. 2020].

High-frequency financial data are known to suffer from various types of irregularities, such as: recording errors, existence of extreme price jumps, bias or microstructure noise. These irregularities can seriously influence or even distort statistical analysis, more specifically the research regarding volatility estimation and risk measurement modelling. Therefore, the need to invent robust and interpretable

outlier detection methods becomes vital in empirical research devoted to the fields of finance and econometrics.

The main goal of our paper is to adapt and apply the COPOD algorithm of outlier identification to real high-frequency financial data from the Warsaw Stock Exchange (GPW). We have performed an empirical study on dataset, which is a pre-processed set containing values of the following variables: price, time (in seconds), and log returns, for a single financial instrument. These values are collected in the GPW20230403000tind_for_COPOD.csv file. In order to achieve this objective, we have restated the COPOD framework in terms of empirical copulas and tail probabilities and subsequently, we have prepared a self-contained Python implementation of skewness-corrected COPOD (a version of COPOD that incorporates a skewness correction), as well as we have interpreted the detected outliers in terms of extreme returns and potential data errors. The considered dataset contains intraday transaction data for a single financial instrument from 3 April 2023.

The remainder of the paper is organized as follows. Section 2 presents theoretical framework and describes the computational implementation and design, while in Section 3 we present our numerical experiments conducted on the selected dataset and discuss the obtained results and in finalal Section 4, we conclude and summarize our research, as well as we propose the possible directions of future research. All of our numerical study has been carried out using the programming language of Python. Appropriate file with the corresponding codes can be available upon Reader's request.

SKEWNESS-CORRECTED COPOD: THEORETICAL BACKGROUND

Sklar's theorems

Construction of the COPOD algorithm and its skewness-corrected version is based on the definition of copula and conclusions from the celebrated Sklar's theorems, which may be recapitulated as follows (see, e.g.: [Sklar 1959; Schweizer and Sklar 1983; Nelsen et al. 2016]).

A copula $C\colon [0;1]^d \to [0;1]$ is defined as the joint df of a random vector $U = (U_1, U_2, \cdots, U_d)$ with the $Uniform([0;1])$ marginals. It means that:

$$C(u_1, u_2, \cdots, u_d) = P(U_1 \leq u_1, U_2 \leq u_2, \cdots, U_d \leq u_d). \tag{1}$$

The Sklar's theorems state that, for a given joint df F with continuous marginals $F_1, F_2, \ldots, F_d$, there exists a uniquely defined copula C, such that:

$$F(t_1, t_2, \ldots, t_d) = C\big(F_1(t_1), F_2(t_2), \ldots, F_d(t_d)\big), \tag{2}$$

and conversely that, for a given copula C and marginals $F_1, F_2, \ldots, F_d$, the relation above defines a df function with marginals F_i.

A copula can be interpreted as a dependence structure, since it captures the complete dependence among random variables after eliminating the influence of their

marginal distributions. The Sklar's theorems formalize this idea by establishing the correspondence between multivariate distributions and pairs consisting of marginal distributions and they allow for independent modeling of marginals and dependence structure. On the other hand, we may also choose arbitrary marginal distributions and combine them with a copula selected in order to display a desired dependence pattern. This flexibility is crucial in applications such as finance, insurance and anomaly detection.

In the present study, we approximate the joint distribution using the product copula, i.e.:

$$\hat{C}(u_1, u_2, \cdots, u_d) = \frac{1}{n}\sum_{k=1}^{n} I\left(U_{1,k} \le u_1, U_{2,k} \le u_2, \cdots, U_{d,k} \le u_d\right) = \frac{1}{n}\sum_{k=1}^{n} I\left(U_{1,k} \le u_1,\right) \cdot I\left(U_{2,k} \le u_2\right) \cdot \ldots \cdot I\left(U_{d,k} \le u_d\right) \approx u_1 \cdot u_2 \cdot \ldots u_d. \quad (3)$$

This corresponds to assuming no explicit parametric dependence structure between the marginal components. Consequently, the joint tail probability is approximated by the product of marginal tail probabilities.

For standard references on copulas and empirical copulas, applied in the Skewness-Corrected COPOD algorithm, we cite the works of [Nelsen et al. 2006] and [Deheuvels 1979].

Now, assume that we have the d-dimensional random vectors $\boldsymbol{X_i} = \left(X_{1,i}, X_{2,i}, \cdots, X_{d,i}\right)$, $i = 1,2,\cdots,n$, with the joint distribution function (df) F and continuous marginal distributions: $F_1, F_2, \ldots, F_d$. We generate data from X_i and obtain n empirical d-dimensional observation vectors $\boldsymbol{x_i} = \left(x_{1,i}, x_{2,i}, \cdots, x_{d,i}\right)$, $i = 1,2,\cdots,n$. Thus, our empirical data (instances) are as follows:

$$\boldsymbol{x_1} = \left(x_{1,1}, x_{2,1}, \cdots, x_{d,1}\right), \boldsymbol{x_2} = \left(x_{1,2}, x_{2,2}, \cdots, x_{d,2}\right), \cdots,$$
$$\boldsymbol{x_n} = \left(x_{1,n}, x_{2,n}, \cdots, x_{d,n}\right).$$

Before we present the theoretical foundations of COPOD and its skewness-corrected variant, we need to introduce the following notations.

The left-tail empirical distribution functions

Having the vectors of empirical data, we define the left-tail empirical distribution function (ecdf) as follows:

$$\hat{F}_j\left(x_{j,i}\right) = \frac{1}{n}\sum_{k=1}^{n} I\left(x_{j,k} \le x_{j,i}\right), \quad (4)$$

where $j = 1,2,\cdots,d, i = 1,2,\cdots,n,$ and where – here and throughout the paper – I stands for the indicator function.

In our numerical study, the left-tail ecdf in (4) is computed using the formula:

$$\hat{F}_j\left(x_{j,i}\right) = \left\{rank_{max}(x_{j,i})\right\}/n, \quad (5)$$

where $\mathrm{rank}_{\max}(x_{j,i})$ stands for the maximum rank of $x_{j,i}$ among $\left\{x_{j,1}, x_{j,2}, \ldots, x_{j,d}\right\}$.

The right-tail empirical distribution functions

Alternatively to the left-tail case, we define the right-tail ecdfs by analogous formula, where instead of $\{x_{j,k} \leq x_{j,i}\}$, we use $\{-x_{j,k} \leq -x_{j,i}\}$ or, equivalently $\{x_{j,k} \geq x_{j,i}\}$, i.e.,

$$\hat{\bar{F}}(x_{j,i}) = \frac{1}{n}\sum_{k=1}^{n} I(-x_{j,k} \leq -x_{j,i}) = \frac{1}{n}\sum_{k=1}^{n} I(x_{j,k} \geq x_{j,i}), \quad (6)$$

where $j = 1,2,\cdots,d, i = 1,2,\cdots,n$.
In our computational study, the left-tail ecdf in (6) is calculated from formula:

$$\hat{\bar{F}}(x_{j,i}) = \{n - \text{rank}_{\max}(x_{j,i})\}/n. \quad (7)$$

Skewness coefficients for each dimension

In many practical situations - including financial data - the distribution of each marginal may be skewed, and consequently – outliers may predominantly appear only in one of the tails. In order to account for this issue, a version of COPOD that incorporates a skewness correction step is employed. Namely, in the first stage of the skewness-corrected framework of the COPOD algorithm, the following skewness coefficient is considered and calculated for each dimension j:

$$b_j = \frac{\sum_{i=1}^{n}(x_{j,i}-\bar{x}_j)^2}{\left(\sqrt{\sum_{i=1}^{n}(x_{j,i}-\bar{x}_j)^2}\right)^3}, j = 1,\cdots,d. \quad (8)$$

The empirical copula observations

In the skewness-corrected COPOD variant, we define:

$$\hat{U}_{j,i} = \hat{F}_j(x_{j,i}), \hat{V}_{j,i} = \hat{\bar{F}}(x_{j,i}), \text{and:} \quad (9)$$

$$\hat{W}_{j,i} = \hat{U}_{j,i}, \text{if } b_j < 0 \text{ or } \hat{W}_{j,i} = \hat{V}_{j,i}, \text{if otherwise.} \quad (10)$$

$\hat{W}_{j,i}$ defines the skewness-corrected empirical copula observation.

Outlier scores of individual observations (for the left-tail, right-tail and skewness-corrected scenarios)

Let, for $i = 1,\cdots,n$,

$$p_{l,i} = -\sum_{j=1}^{d}\log(\hat{U}_{j,i}), \; p_{r,i} = -\sum_{j=1}^{d}\log(\hat{V}_{j,i}), \; p_{s,i} = -\sum_{j=1}^{d}\log(\hat{W}_{j,i}), \quad (11)$$

The quantities above are called as the left-tail, the right-tail and the skewness-corrected outlier scores, respectively.

The overall COPOD outlier score

The outlier scores of the individual observations $\boldsymbol{x_i} = (x_{1,i}, x_{2,i}, \cdots, x_{d,i})$ are defined as:

$$O(\boldsymbol{x_i}) = max\{p_{l,i}, p_{r,i}, p_{s,i}\}, i = 1, \cdots, n. \quad (12)$$

Large values of $O(x_i)$ indicate strong evidence that $\boldsymbol{x_i}$ is an outlier.
We can treat the values of $O(x_i)$ as the coordinates of the vector:

$$O(\boldsymbol{X}) = [O(\boldsymbol{x_1}), O(\boldsymbol{x_2}), \cdots, O(\boldsymbol{x_n})]^T. \quad (13)$$

In order to identify anomalous observations, the empirical distribution of $\{O(\boldsymbol{x_i})\}_{i=1}^{n}$, and a quantile-based threshold is applied. In our empirical study the 95th percentile of the score distribution is used, resulting in approximately 5% of the observations being classified as outliers.

EMPIRICAL RESULTS

Dataset

The dataset considered in our numerical study consists of a pre-processed feature matrix concerning the transaction quotes of a single financial instrument traded on the Warsaw Stock Exchange (GPW).
The file GPW20230403000tind_for_COPOD.csv contains 28,624 observations and three continuous variables: *price*, *time_seconds*, and *log_return*. These features capture both the level and dynamics of the price process over a trading session. The numerical data include intraday observations for a single financial instrument from 3 April 2023.

The observations from our dataset form the high-frequency financial data. For more details on high-frequency data and financial market microstructures, we refer to [Engle 2000; O'Hara 1998; Aït-Sahalia et al. 2014]. Furthermore, in the context regarding the research study on extreme-value behavior in finance, we recommend: [Embrechts et al. 1997; Cont 2001].

We have prepared the Python code in order to implement the skewness-corrected COPOD algorithm for our datasets. This code can be available upon Reader's request. Below, we collect the results obtained through the mentioned algorithm's application.

Key Results and Conclusions Regarding an Application of the Skewness-Corrected COPOD Algorithm for the Selected Dataset

We list here the summary information on the results obtained by an application of the skewness-corrected COPOD algorithm for our dataset.

- Number of observations: **28,624**.
- Number of features: **3 (*price, time_seconds, log_return*)**,

- Outlier threshold (cutoff = 95% quantile score): $\boldsymbol{q_{0.95} \approx 7.559}$.
- Number of detected outliers: **1,432 (≈ 5% of the data)**.
- Number of observations detected as outliers: **1,432**.
- Top-20 most extreme observations detected as outliers (20 instances with the largest scores $O(x_i)$):

  $\boldsymbol{x_{10321}, x_{10317}, x_{19774}, x_{17710}, x_{10325}, x_{28133}, x_{19777}, x_{17716}, x_{19775}, x_{25678}, x_{985}, x_{17712}, x_{14266}, x_{17717}, x_{26170}, x_{989}, x_{493}, x_{19779}, x_{14763}, x_{2465}.}$
- The corresponding scores $O(x_i)$ of top 20 most extreme observations detected as outliers:

  24.993, 24.993, 24.770, 23.489, 20.176, 19.765, 19.359, 19.036, 18.674, 17.826, 17.619, 17.540, 17.294, 17.252, 17.191, 17.136.

Characterization of 20 observations with the largest COPOD scores

In this part, we present a feature-level characterization of 20 observations that achieved the largest COPOD outlier score**.** The analyzed features are: *price* – price level of the financial instrument, *time_seconds* – intraday time (in seconds), *log_return* – logarithmic return.

The observations with the largest COPOD scores are characterized by the following patterns: they have very large absolute *log-returns* (the *log-returns* empirical values are both negative and positive), they reach extremely low or extremely high *price* levels, they exhibit strong temporal concentration at specific intraday times (specifically, 22,800 seconds).

When standardized with respect to the full dataset, the following properties regarding the relation between the top 20 observations detected as outliers and the explanatory variables may be seen: these observations have moderately low ($\approx -0.9\sigma_1$) or extremely high (up to $+6\sigma_1$) values of the *price* future, values of the *time_seconds* variable are strongly concentrated around a few discrete values (($\approx -2.8\sigma_2$ for 22,800 seconds), the *log_return* values are extremely heavy-tailed – approximately from $(-46)\sigma_3$ to $+102\sigma_3$, where $\sigma_1 - \sigma_3$ denote the corresponding standard deviations from the observations of the considered features.

Common characteristics of the top-20 skewness-corrected COPOD observations can be summarized as follows.

The most prominent property of these observations is their extreme position in the tails of the log-return distribution:

- large negative values (e.g.: −3.97, −3.59, −3.00, −2.69),
- large positive values (e.g.: +8.73, +5.21, +3.50, +2.49).

These values deeply emerge in the marginal tails and therefore exhibit very small empirical CDF values, which results in large COPOD scores.

The same top-20 skewness-corrected COPOD observations also correspond to extreme price levels:

- very low prices (e.g.: 4.60, 167.48),
- very high prices (exceeding 60,000).

The lowest detected *price* levels are small in comparison to the session *price* median of 5,458.61, while the highest *prices* represent boundary observations of the trading day.

Furthermore, with regard to the *time_seconds* feature, most of the observations occur at exactly 22,800 seconds, with a smaller number clustered around 32,000–40,500 seconds and one extreme at 61,800 seconds (the corresponding median is 47,100).

Tu sum up, we conclude that the *log_return* variable is the dominant factor of extremeness, whereas the *price* feature also influences it, but to a lesser degree. In addition, while the *time_seconds* feature is not the primary factor in the identification of rare events, but it also strengthens extremeness by placing these observations in low-density or boundary regions of the empirical distribution.

Economic and Statistical Interpretation

From an economic perspective, the observations associated with the largest values of the skewness-corrected COPOD scores $O(x_i)$ correspond to events that are highly atypical relative to the empirical distribution of intraday price dynamics. These observations are characterized by unusually large price movements, which may arise from high-impact market events, sudden liquidity shocks, or structural breaks in trading activity. In some cases, the detected extremes may also be linked to data-related issues, such as recording errors, price misprints or unpredicted corporate actions.

From a statistical viewpoint, the detected observations represent true tail events rather than local density anomalies. Their extremeness is mainly due to very small empirical tail probabilities in at least one marginal distribution, most notably the log-return component. The skewness-corrected COPOD algorithm correctly assigns large scores to such observations by aggregating marginal tail contributions through the empirical copula setting. This confirms that the method is particularly suitable for the identification of observations that lie in low-probability regions of the joint distribution, even when marginal distributions are strongly asymmetric.

Furthermore, the decomposition of the COPOD score into the left-tail, right-tail and skewness-corrected components allows for a clear interpretation of the factors that influence extremeness. In the analyzed dataset, extreme values of *the log-returns* variable have a dominant impact on the observed anomalous behavior, whereas the extreme *price* levels and the temporal concentration of *the time_seconds* quantities strengthen this impact. Such interpretability is especially desirable and valuable in financial applications, where distinguishing between economically meaningful extreme events and data recording errors or missing data is crucial.

SUMMARY AND FUTURE WORK

In our paper, we investigated the applicability of the skewness-corrected COPOD (Copula-Based Outlier Detection) algorithm to high-frequency financial data from the Warsaw Stock Exchange. Building on the theoretical framework of empirical copulas, we restated the core principles of COPOD in a concise and transparent form and provided the Python implementation used for the financial data characterized by skewness and heavy tails.

The empirical analysis, conducted on a pre-processed high-frequency dataset containing prices, intraday time and log-returns for a single financial instrument, shows that the skewness-corrected COPOD algorithm effectively identifies extreme observations corresponding to rare and economically meaningful events. The method consistently identifies approximately five percent of the observations as outliers when a quantile-based threshold is applied, while clearly separating the most extreme tail events from the majority of the data. The results confirm that log-returns constitute the primary factor of extremeness, with price levels and temporal effects providing additional impact on increase of the extremeness power.

A key advantage of the proposed approach lies in its nonparametric, deterministic, and parameter-free nature, which allows to eliminate the need for hyperparameter tuning or stochastic training. Moreover, the additive log-tail decomposition of the COPOD score secures a high degree of interpretability, which enables to investigate anomalous behavior for specific variables and distributional tails. These properties make the skewness-corrected COPOD particularly attractive for exploratory data analysis, data cleaning and risk controlling in high-frequency financial data.

Despite its strengths, the proposed approach also has some limitations. Being rank-based and static, the current formulation does not explicitly account for temporal dependence, intraday seasonality or market volatility. In addition, when applied independently to individual instruments, the method may fail to detect anomalies that arise from joint behavior across multiple assets.

Future research may therefore proceed in several directions. First, the COPOD framework could be extended to incorporate temporal structure, for example through hybrid models that combine empirical copulas with time-series features. Moreover, multivariate extensions involving multiple instruments could be explored in order to detect systemic anomalies and tail dependence. Furthermore, comparison of the skewness-corrected COPOD with alternative state-of-the-art anomaly identification methods, as well as combining it with conformal prediction techniques, could further enhance its practical relevance and robustness in large-scale financial applications. Additionally, it would be interesting in future work to apply the algorithm by incorporating parametric families of copula for dependent structures (Gaussian, Student-t copulas, etc.) instead of using an approximation by the product copula. We could also consider modeling stochastic dependencies of individual transactions by

applying vine copula model to multidimensional time series (see [Nagler et al. 2022]).

REFERENCES

Aggarwal C. C. (2017) Outlier Analysis. 2nd edn. Springer, Cham. https://doi.org/10.1007/978-3-319-47578-3

Aït-Sahalia Y., Jacod J. (2014) High-Frequency Financial Econometrics. Princeton University Press, Princeton. https://doi.org/10.1515/9781400850325

Breunig M. M., Kriegel H.-P., Ng R. T, Sander, J. (2000) LOF: Identifying Density-Based Local Outliers. ACM SIGMOD Record, 29(2), 93-104.

Chandola V., Banerjee A., Kumar V. (2009) Anomaly Detection: A Survey. ACM Computing Surveys, 41(3), 15. https://doi.org/10.1145/1541880.1541882

Cont R. (2001) Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues. Quant Finance 1(2), 223-236. https://doi.org/10.1080/713665670

Deheuvels P. (1979) La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. Bulletins de l'Académie Royale de Belgique, 65, 274-292.

Embrechts P., Klüppelberg C., Mikosch T. (1997) Modelling Extremal Events for Insurance and Finance. Springer-Verlag.

Engle R. F. (2000) The Econometrics of Ultra-High-Frequency Data. Econometrica, 68(1),1-22. https://doi.org/10.1111/1468-0262.00091

Hariri S., Carrasco Kind M., Brunner R. J. (2021) Extended Isolation Forest. IEEE Transactions on Knowledge and Data Engineering, 33(4), 1479-1491. https://doi.org/10.1109/TKDE.2019.2947676

Li Z., Zhao Y., Botta N., Ionescu C., Hu X. (2020) COPOD: Copula-Based Outlier Detection. [In:] Proceedings of the IEEE International Conference on Data Mining (ICDM), 1118-1123. https://doi.org/10.1109/ICDM50108.2020.00138

Liu F. T., Ting K. M., Zhou Z.-H. (2008) Isolation Forest. IEEE ICDM, 413-422.

Nagler T., Krüger D., Min A. (2022) Stationary Vine Copula Models for Multivariate Time Series. Journal of Econometrics, 227(4), 305-324.

Nelsen R. B. (2006) An Introduction to Copulas. 2nd edn. Springer, New York https://doi.org/10.1007/0-387-28678-0

O'Hara M. (1998) Market Microstructure Theory. Wiley & Sons. https://asset.quant-wiki.com/pdf/Maureen%20O%27Hara%20-%20Market%20Microstructure%20Theory%20%20-Wiley%20%281998%29.pdf?ref=fxopen.com

Schölkopf B., Platt J. C., Shawe-Taylor J., Smola A. J., Williamson R. C. (2001) Estimating the Support of a High-Dimensional Distribution. Neural Computation, 13(7), 1443-1471.

Schweizer B., Sklar A. (1983) Probabilistic Metric Spaces. North-Holland, New York.

Sklar A. (1959) Fonctions de répartition à N dimensions et leurs marges. Annales de l'ISUP, VIII (3), 229-231.

METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XXVI/4, 2025, s. 175 – 185

Received: 03.11.2025
Accepted: 22.12.2025

KONSTRUKCJA INDEKSU TRANSFORMACJI KLIMATYCZNEJ DLA GPW

Arkadiusz W. Szymanek https://orcid.org/0000-0003-4825-4600
Zakład Finansów, Wydział Zarządzania, Politechnika Warszawska
e-mail: arkadiusz.szymanek@pw.edu.pl
Tomasz K. Wiśniewski https://orcid.org/0000-0002-6417-5104
Wydział Ekonomiczno-Socjologiczny, Uniwersytet Łódzki
e-mail: tomasz.wisniewski@uni.lodz.pl

Streszczenie: Zmiany klimatyczne to jedno z najważniejszych wyzwań współczesnego świata. Zjawiska te w coraz większym stopniu nasilają się pod wpływem emisji gazów cieplarnianych. W celu aktywizacji działań przedsiębiorstw na rzecz transformacji klimatycznej niezbędne jest tworzenie regulacji oraz mierników statystycznych i indeksów giełdowych, które skieruję i zdynamizują przypływy kapitałów wspierających te działania. Na rynku kapitałowym jednym ze strumieni kapitałowych są inwestycje indeksowego, w tym poprzez fundusze ETF, które naśladują portfele indeksów giełdowych. Dlatego konstrukcja indeksu transformacji klimatycznej popularnie nazywanym „zielonym indeksem" umożliwia pasywne inwestowanie według obiektywnie zbudowanego wzorca na rzecz ograniczenia emisji gazów cieplarnianych. Najbardziej obiecującym kierunkiem jest tworzenie instrumentów finansowych takich jak fundusze ETF obejmujące spółki giełdowe zaangażowane w transformację klimatyczną.

W wyniku przeprowadzonych badań opracowano koncepcję indeksu transformacji klimatycznej obejmującego spółki notowanych na GPW. Analiza szeregu czasowego indeksu ZITK została przeprowadzona w porównaniem do najpopularniejszych indeksów GPW tj. WIG, WIG20, mWIG40 i sWIG80. Stwierdzono, że indeks ZITK może być wzorcem inwestowania, czyli instrumentem bazowym na polskim rynku kapitałowym m.in. dla funduszy ETF. Zaprojektowany indeks ma podobne własności agregacyjne jak indeks szerokiego rynku WIG i wyraźnie lepsze własności niż inne badane indeksy giełdowe w Polsce.

Kluczowe słowa: ESG, GPW, ETF, transformacja klimatyczna

JEL classification: G11

https://doi.org/10.22630/MIBE.2025.26.4.16

WSTĘP

Grudzień 2024 roku był drugim z najcieplejszych ostatnich miesięcy roku w historii, a styczeń 2025 roku zajął nawet pierwsze miejsce w rankingu, wyprzedzając styczeń 2024 roku o 0,05 stopnia Celsjusza[1].

Zgodnie z najnowszą analizą Climate Central zmiany klimatu wywołane działalnością człowieka doprowadziły do wzrostu temperatur oraz znacząco zwiększyły prawdopodobieństwo występowania ekstremalnych upałów i susz na całym świecie. Jako okres porównawczy w badaniu wykorzystano średnią z lat 1991 – 2020, co uwzględnia wzrost temperatury o 0,9C w stosunku do okresu przedindustrialnego. Jak wynika z raportu pomiędzy grudniem 2024 roku, a lutym 2025 roku każdego dnia co najmniej jedna osoba z każdych pięciu osób żyjących na Ziemi była narażona na skutki ocieplenia. Spalanie paliw kopalnych jest uważane za główną przyczynę efektu cieplarnianego. Dotyka on coraz większą część mieszkańców naszego globu. W dniu 28 lutego 2025 roku zanotowano rekordowy zasięgu podwyższonej temperatury, który objął 37 proc. globalnej populacji, czyli aż 3 miliardy ludzi.

Podobny odsetek wystąpił na terenie Unii Europejskiej, a zmian klimatycznych z wyraźnie podwyższoną temperaturą doświadczyło ponad 146 milionów ludzi. W Polsce również było wówczas cieplej niż zwykle, a największe anomalie temperaturowe odnotowano w krajach bałtyckich. W lutym 2025 roku średnia temperatura była tam wyższa o 2,8 st. C niż zwykle.

W celu przeciwdziałania negatywnym zjawiskom cieplarnianym niezbędne jest stosowanie narzędzi finansowania transformacji klimatycznej. Instytucje finansowe mogą uczestniczyć w finansowaniu innowacyjnych technologii niskoemisyjnych, wspieraniu transformacji energetycznej przedsiębiorstw, rozwoju zrównoważonych produktów finansowych oraz w zarządzaniu ryzykiem klimatycznym w polskiej gospodarce. Sektor finansowy poprzez kierowane przepływy kapitałowe może uodporniać kraj na ekstremalne zjawiska pogodowe.

Wielu inwestorów coraz częściej przy podejmowaniu decyzji inwestycyjnych bierze pod uwagę działania na rzecz ochrony środowiska, a w szczególności w zakresie ograniczania emisji gazów cieplarnianych[2]. Brak działań oraz zanieczyszczanie środowiska przez emitentów negatywnie wpływa na dostęp do kapitału i utrudnia udział w globalnych łańcuchach dostaw.

Finansowanie projektów związanych ze zmianami klimatycznymi jest uzasadnione ekonomicznie, gdyż zielone technologie perspektywicznie mogą przynieść korzyści gospodarcze poprzez tworzenie nowych miejsc pracy oraz zwiększenie globalnej konkurencyjności i innowacyjności.

---

1 https://klimat.rp.pl/klimat/art41985071-ranking-przegrzanych-miast-europy-eksperci-klimatyczni-bija-na-alarm

2 https://www.weforum.org/publications/global-risks-report-2024/

Rosnące znaczenie polityk klimatycznych na poziomie międzynarodowym i wprowadzanie powoduje, że wielu krajach wprowadza przepisy i nakłada zobowiązania dotyczących redukcji emisji cieplarnianych. Inwestorzy zaangażowani w takie projekty, mogą być lepiej przygotowani na wprowadzanie tych regulacji łatwiej unikając potencjalnych sankcji oraz kosztów dostosowań pod presją czasu.

Analiza badań z lat 2015-2024 wskazuje na korelację zrównoważonego rozwoju i wyników finansowych przedsiębiorstw[3]. Ponadto zarządzanie przedsiębiorstwami ukierunkowane na niskoemisyjność nie tylko poprawia ich wyniki finansowe, ale i ogranicza ryzyko regulacyjne. Realizacja redukcji emisji gazów cieplarnianych (emisji CO_2) wymaga zwiększenia efektywności energetycznej a ta wymaga nakładów kapitałowych i instrumentów finansowych.

INDEKSY TRANSFORMACJI KLIMATYCZNEJ

Wprowadzenie „zielonych indeksów" i funduszy na nich opartych umożliwia inwestorom ocenę emitentów zarówno wyników finansowych, jak i wpływu na środowisko naturalne. W 1971 roku w USA powstał pierwszy fundusz inwestycyjny (Pax World Fund), którego polityka inwestycyjna uwzględniała aspekty środowiskowe. W 1997 roku powstała Global Reporting Initiative (GRI) - organizacja, promująca standardy raportowania zrównoważonego rozwoju.

Indeksy transformacji klimatycznej wspierają działania na rzecz redukcji emisji gazów cieplarnianych i dostosowania się do zmian klimatycznych. Wskaźniki te uwzględniają te przedsiębiorstwa, które podejmują aktywne działania na rzecz redukcji emisji i zanieczyszczeń. Spółki giełdowe są kwalifikowane do udziału w „zielonych indeksach" w oparciu o ich strategie dekarbonizacji, efektywność energetyczną oraz wdrażanie zielonych technologii.

Celem indeksów jest minimalizacja śladu węglowego z portfela inwestycji biorąc pod uwagę trajektorię dekarbonizacji. Globalne agencje informacyjne, takie jak MSCI, S&P Global, FTSE Russell oraz STOXX konstruują i publikują zielone indeksy stosując własne metodyki.

MSCI World Climate Change Index[4] opiera się na indeksie MSCI World i obejmuje duże oraz średnie spółki z rynków rozwiniętych. Celem indeksu jest reprezentowanie wyników strategii inwestycyjnej, która waży udziały w portfelu indeksu w oparciu o możliwości i ryzyka związane z przejściem na gospodarkę niskoemisyjną, minimalizując jednocześnie wykluczenia spółek z indeksu bazowego. W procesie kwalifikacji spółek do portfela, indeks wykorzystuje metodykę ocenę MSCI Low Carbon Transition (LCT). Wagi spółek zaangażowanych w transformację klimatyczną są podwyższane. Wagi spółek

[3] https://www.unpri.org/pri-blog/part-iii-esg-factors-and-returns-a-review-of-recent-research/12728.article

[4] https://www.msci.com/documents/10199/18f2379d-4306-22d6-515c-1d3b50f94b0b

narażonych na ryzyka związane z transformacją energetyczną są obniżane. Indeksy MSCI Climate Change Indexes - MSCI są obliczany i publikowany w trybie ciągłym, zaś rewizje portfeli przeprowadzone są regularnie co kwartał. Portfel indeks jest zbudowany z akcji 1341 spółek giełdowych. Największe z nich to spółki amerykańskie: Apple, Microsoft i Amazon.

Indeks S&P Paris-Aligned & Climate Transition (PACT)[5] skupia się na celach redukcji emisji dwutlenku węgla, w tym strategie efektywności węglowej i wykluczenia paliw kopalnych. Indeks S&P Climate Transition stosuje kompleksowe podejście do kwalifikacji spółek, łączące wykluczenia sektorowe z kryteriami klimatycznymi i ESG. W pierwszej kolejności z indeksu eliminowane są podmioty zaangażowane w produkcję kontrowersyjnej broni, naruszające zasady UN Global Compact oraz z sektorów tytoniu, wydobycia węgla oraz ropy i gazu .

KONSTRUKCJA I WERYFIKACJA WŁASNOŚCI INDEKSU TRANSFORMACJI KLIMATYCZNEJ DLA GPW

W 2022 roku GPW Benchmark - administrator indeksów giełdowych GPW - opublikował raport dotyczący możliwości wprowadzenia indeksu transformacji klimatycznej (CTB) w Polsce[6]. W wyniki monitoringu sprawozdań okazało się, że dostępność danych wówczas na polskim rynku kapitałowym była niewystarczająca, co uniemożliwiało uruchomienie reprezentatywnego zielonego indeksu. Obliczenia wykonano bowiem dla 19 spółek, z których jedynie 12 zakwalifikowało się do indeksu typu CTB.

Wprowadzenie nowych obowiązków regulacyjnych wynikających z dyrektywy CSRD (Corporate Sustainability Reporting Directive) zwiększyło zakres dostępnych danych w kwestiach klimatycznych na polskim rynku. Spółki giełdowe mają wymóg informacyjny w zakresie emisji gazów cieplarnianych związanych z ich działalnością i majątkiem.

Metodyka zielonego indeksu dla spółek notowanych na GPW została oparta na zasadach stosowanych w procesie konstrukcji globalnych indeksów transformacji klimatycznej. Proces konstrukcji został opisany w trzech etapach.

Etap I. Preselekcja bazy spółek z GPW

W pierwszym etapie zdefiniowano bazową grupę spółek, które brane są pod uwagę w dalszych analizach. Mając na uwadze dotychczasowe praktyki GPW w procesie konstrukcji indeksu WIG-ESG za taką grupę uznano spółki tworzące portfele indeksów WIG20 i mWIG40 na koniec roku kalendarzowego.

---

[5] https://www.spglobal.com/spdji/en/documents/additional-material/brochure-sp-pact-indices.pdf

[6] https://www.gov.pl/attachment/37befe72-439b-4553-ba85-c2f9cabb16c2

Etap II. Kwalifikacja spółek do indeksu

To kluczowy etap, wymagający kategoryzacji spółek z uwagi na realizację celów transformacji klimatycznej. Podstawowymi dokumentami są oświadczenia poszczególnych spółek odnośnie realizacji działań strategicznych, zarządzania ryzykiem i pomiarów wskaźników środowiskowych oraz zużycia energii. Podmioty, które nie opublikowały oświadczeń zrównoważonego rozwoju, bądź nie określiły planu przejścia na potrzeby łagodzenia zmian klimatu były wykluczane z kwalifikacji do indeksu.

Etap III. Wyznaczenie struktury wag indeksu

W ostatnim etapie dla spółek zakwalifikowanych do portfela indeksu wyznaczane są pakiety akcji na podstawie metodyki stosowanej przez GPW Benchmark. Zgodnie z tymi zasadami pakiety akcji spółek w portfelu wyznaczane są w oparciu o liczbę akcji w wolnym obrocie zaokrągloną do pełnych tysięcy. Ponadto przy budowie portfela zielonego indeksu obowiązują zasady dywersyfikacji tj. ograniczenia maksymalnego udziału pojedynczej spółki w portfelu indeksu.

Na podstawie przyjętych założeń oraz wyników analiz wyselekcjonowano grupę 31 spółek, które zostały zakwalifikowane do portfela indeksu. Wartości indeksu zostały wyznaczone na podstawie metodyki przyjętej przez GPW Benchmark zgodnie z poniższym wzorem:

$$INDEKS = \frac{P_i * S_i}{P_0 * S_0} * I_0$$

gdzie:

kapitalizacja bieżąca indeksu:

P_i – kurs uczestnika indeksu ‘i’ na danej sesji

S_i – pakiet uczestnika indeksu ‘i’ na danej sesji

kapitalizacja bazowa indeksu:

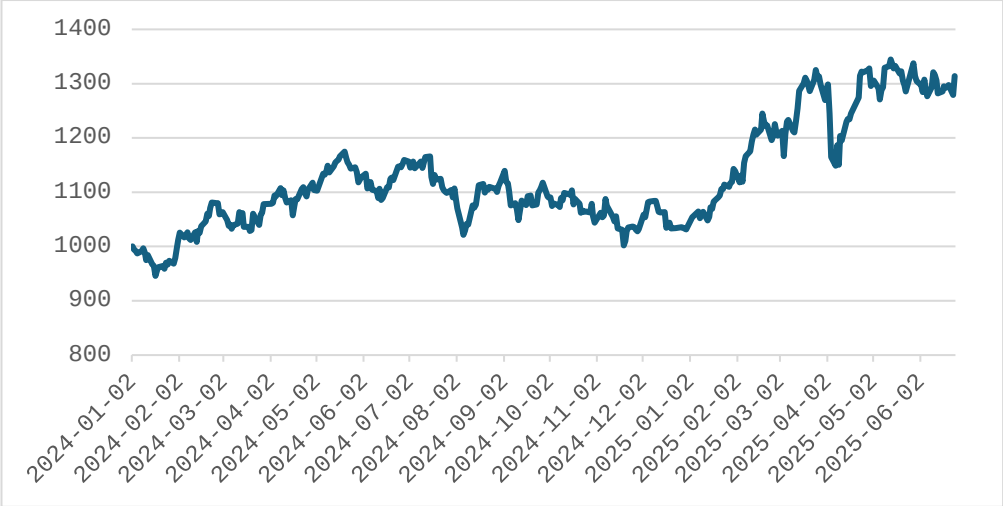
P_0 – kurs uczestnika indeksu ‘i’ na sesji w dniu bazowym

S_0 – pakiet uczestnika indeksu ‘i’ na sesji w dniu bazowym

I_0 – wartość bazowa indeksu (1000 punktów)

Biorąc pod uwagę wyznaczony skład portfela indeksu oraz przyjęte zasady kalkulacji przeliczono jego wartości historyczne za okres od 1 stycznia 2024 r. do 2 czerwca 2025 r. Obliczone wartości w formie graficznej są przedstawione na rysunku 1.

Rysunek 1 Wartości indeksu transformacji klimatycznej w okresie od 1 stycznia 2024 do 2 czerwca 2025



Źródło: opracowanie własne

Analiza własności wartości polskiego indeksu transformacji klimatycznej ZITK (Zielony Indeks Transformacji Klimatycznej – nazwa robocza) została przeprowadzona w relacji do wybranych indeksów giełdowych: WIG – indeks szerokiego rynku GPW, WIG20 – indeks największych spółek giełdowych, mWIG40 – indeks średnich spółek giełdowych oraz sWIG80 – indeks małych spółek giełdowych.

Badanie zależności szeregów czasowych pomiędzy różnymi indeksami giełdowymi jest przedmiotem wielu prac naukowych. Na przykład Fiszeder zastosował modele GARCH do analizy zależności pomiędzy indeksami GPW, a indeksami rynków zagranicznych [2001]. Matilla-García oraz Ruiz-Marín przeprowadzili analizę multifraktalną głównych europejskich indeksów giełdowych (DAX, CAC40, FTSE100, IBEX35) podczas pandemii COVID-19 [Matilla-García M., Ruiz Marín M. 2021]. W badaniu wykorzystano m.in. statystyki opisowe, współczynniki zmienności, korelacje krzyżowe oraz analizę wykładnika Hurtsa. Badanie wykazało znaczny wzrost stopnia multifraktalności europejskich indeksów giełdowych podczas pandemii COVID-19, szczególnie w pierwszym okresie pandemii. Autorzy zaobserwowali zwiększoną korelację między zmianami wartości indeksów na badanych rynkach i wykazali na istotny wzrost ryzyka rynkowego i brak możliwości efektywnego ograniczania ryzyka przez międzynarodową dywersyfikację inwestycji.

Zhang i Li zbadali zależności między ekstremalnymi wartościami stóp zwrotu kryptowalut (Bitcoin, Ethereum), a tradycyjnymi indeksami giełdowymi:S&P500, NASDAQ, FTSE100 [Zhang L., Li Y. 2022]. W badaniu przeanalizowano w szczególności współczynniki Pearsona, Spearmana oraz test Kołmogorowa-

Smirnowa. Wykazano asymetryczną zależność ogonową między kryptowalutami a tradycyjnymi indeksami giełdowymi - silniejszą w przypadku ekstremalnych spadków niż ekstremalnych wzrostów.

Biorąc pod uwagę dotychczasowe wyniki badań zależności szeregów czasowych, w niniejszej pracy wykorzystano poniższe wskaźniki analizy statystycznej:

- Statystyki opisowe – m.in. średnia, mediana, minimum, maksimum;
- Współczynnik korelacji Pearsona – w celu weryfikacji liniowej zależności między różnymi parami indeksów;
- Macierz korelacji - do badania rozkładu zależności dla wielu kombinacji indeksowych;
- T Test Kołmogorowa-Smirnowa - służący do porównywania rozkładów.

Analiza statystyczna została przeprowadzona z wykorzystaniem logarytmów naturalnych dziennych stóp zwrotu poszczególnych indeksów, zaś do badania stacjonarności szeregów czasowych wykorzystano test ADF, czyli rozszerzony test Dickeya-Fullera.

Na podstawie przeprowadzonej analizy, wszystkie badane szeregi czasowe (WIG, ZITK, WIG20, mWIG40, sWIG80) okazały się być stacjonarne. Statystyki testowe dla każdego szeregu czasowego są znacznie mniejsze (bardziej ujemne) niż wartości krytyczne na poziomie istotności 5%, co pozwala na odrzucenie hipotezy zerowej o ich niestacjonarności.

Podstawowe statystyki opisowe szeregów czasowych zostały przedstawione w tabeli 1. Zielony indeks transformacji klimatycznej wykazuje nieznacznie wyższą średnią stopę zwrotu niż WIG, co sugeruje, że w analizowanym okresie ZITK uzyskał lepsze wyniki inwestycyjne. Wartości odchylenia standardowego ZITK są wyższe niż dla indeksu WIG, co wskazuje na wyższe ryzyko inwestycyjne. Rozkład zmian wartości obu indeksów charakteryzuje się ujemną skośnością. Można stwierdzić, że rozkłady stóp zwrotu obu indeksów są przesunięte w lewo, a ryzyko dużych spadków (strat) jest większe niż szansa na duże wzrosty (zyski).

WIG ma nieznacznie większą asymetrię niż projektowany indeks ZITK. Indeks ZITK wykazuje nieistotnie wyższą średnią stopę zwrotu niż indeks WIG, co oznacza, że w analizowanym okresie indeks ZITK przynosił średnio lepsze wyniki niż WIG. Indeks ZITK cechuje się wyższym odchyleniem standardowym, co wskazuje na większą zmienność (wyższe ryzyko) w porównaniu do WIG. Rozkłady zmian obu indeksów wykazują ujemną skośność, co oznacza, że rozkład stóp zwrotu jest przesunięty w lewo (w stronę wartości ujemnych). Ryzyko dużych strat jest więc większe niż szansa na duże zyski. WIG ma nieznacznie większą asymetrię niż ZITK. Analizowane indeksy wykazują podobne cechy rozkładu - ujemną skośność i dodatnią kurtozę, co wskazuje na większe prawdopodobieństwo wystąpienia ekstremalnych strat niż ekstremalnych zysków. Można stwierdzić, że oba indeksy charakteryzują się podobnym profilem statystycznym, z niewielką przewagą średniej stopy zwrotu ZITK względem WIG. Z drugiej strony należy wskazać na

wyższe ryzyko dla inwestycji w portfel indeksu ZITK. A zatem podwyższone ryzyko inwestycyjne w przypadku indeksu ZITK względem WIG jest kompensowany podwyższoną stopą zwrotu.

Tabela 1. Podstawowe statystyki opisowe szeregów czasowych logarytmów naturalnych dziennych stóp zwrotu

Wskaźnik	WIG	ZITK	WIG20	mWIG40	sWIG80
Średnia	0,00081	0,00086	0,00060	0,00091	0,00061
Błąd standardowy	0,00067	0,00075	0,00077	0,00059	0,00041
Mediana	0,00168	0,00143	0,00086	0,00160	0,00089
Odchylenie standardowe	0,01220	0,01362	0,01401	0,01070	0,00751
Wariancja próbki	0,00015	0,00019	0,00020	0,00011	0,00006
Kurtoza	1,96425	1,93568	1,59380	2,16910	5,41489
Skośność	-0,40125	-0,38778	-0,28847	-0,56796	-0,85453
Zakres	0,10054	0,11119	0,11383	0,08265	0,07289
Minimum	-0,05932	-0,06649	-0,06577	-0,05167	-0,04583
Maksimum	0,04122	0,04470	0,04806	0,03099	0,02707
Suma	0,26745	0,28554	0,19917	0,30115	0,20267
Licznik	332	332	332	332	332
Poziom ufności (95,0%)	0,00132	0,00147	0,00151	0,00116	0,00081

Źródło: obliczenia własne

W kolejnym etapie na podstawie współczynnika korelacji Pearsona badano liniowe zależności pomiędzy indeksem WIG, a pozostałymi indeksami. Wyniki analiz zestawiono w tabeli 2.

Tabela 2 Wartości współczynnika korelacji indeksu ZITK

Wskaźnik	WIG	WIG20	mWIG40	sWIG80
Współczynnik Pearsona	0,9932	0,9906	0,8915	0,7435

Źródło: opracowanie własne

Indeks transformacji klimatycznej w największym stopniu jest skorelowany z indeksem WIG, zaś najniższa zależność występujące w relacji do indeksu sWIG80. Ponadto najsłabsza zależność występuje pomiędzy indeksem WIG20, a indeksem sWIG80. Tabela korelacyjna ujawnia istotne różnice w zachowaniu poszczególnych segmentów polskiego rynku kapitałowego, potwierdzając dominujący wpływ dużych spółek na indeks szerokiego rynku oraz możliwości dywersyfikacyjne oferowane przez segment małych spółek.

W ostatnim etapie badań przeanalizowano test nieparametryczny Kołmogorowa-Smirnowa w celu porównania rozkładu szeregu czasowego indeksu

WIG z szeregami czasowymi pozostałych analizowanych indeksów. Wyniki obliczeń zestawiono w tabeli 3.

Tabela 3 Wyniki obliczeń testu Kołmogorowa-Smirnowa dla poszczególnych indeksów w relacji do indeksu WIG.

Indeks	Statystyka testu K-S	Wartość p	Ocena dla	Wniosek
ITK	0,069307	0,743764	Nie odrzucamy H_0	Brak dowodów na różnicę rozkładów
WIG20	0,112871	0,218387	Nie odrzucamy H_0	Brak dowodów na różnicę rozkładów
mWIG40	0,148515	0,045234	Odrzucamy H_0	Rozkłady są różne
sWIG80	0,252475	0,000053	Odrzucamy H_0	Rozkłady są różne

Źródło: opracowanie własne

Z obliczeń przeprowadzonych w celu weryfikacji testu Kołmogorowa-Smirnowa wynika, że jedynie w przypadku indeksów transformacji klimatycznej oraz WIG20 dystrybucji stóp zwrotu ma taki sam rozkład jak dystrybucji dla indeksu WIG.

WNIOSKI

Przedstawiona propozycja nowego indeksu transformacji klimatycznej dla polskiego rynku kapitałowego jako odpowiedź na wyzwania związane ze zmianami klimatycznymi i potrzebę transformacji gospodarki w kierunku modelu niskoemisyjnego.

Autorzy wskazują na możliwości konstrukcji wzorca inwestowania w celu mobilizacji kapitału giełdowego do finansowania zielonej transformacji. Wykazano lukę badawczą na polskim rynku kapitałowym w zakresie indeksu koncentrującego się na transformacji klimatycznej w oparciu o spółek notowanych na GPW.

Zaproponowany indeks transformacji klimatycznej charakteryzuje się kompleksowym podejściem, uwzględniającym nie tylko aktualne wskaźniki emisyjności przedsiębiorstw, ale również dynamikę zmian i długoterminowe zobowiązania w zakresie redukcji emisji gazów cieplarnianych.

Weryfikacja statystyczna wykazała obiecujące rezultaty. Indeks transformacji klimatycznej dla GPW osiągnął wyższe stopy zwrotu w porównaniu z innymi indeksami polskiego rynku przy zachowaniu akceptowalnego poziomu ryzyka. Analiza statystyk opisowych, współczynników Pearsona oraz test Kołmogorowa-Smirnowa potwierdziły korzystny profil ryzyko-zwrot proponowanego rozwiązania.

Badania jednoznacznie wskazują, że opracowany indeks może zwiększyć transparentność działań spółek giełdowych na rzecz przeciwdziałania zmianom klimatu oraz stymulować rozwój zielonych inwestycji w Polsce. Autorzy sugerują

dalsze badania nad rozszerzeniem metodyki o wskaźniki adaptacji do zmian klimatu oraz analizę długoterminowej efektywności indeksu w różnych warunkach rynkowych i dłuższym okresie.

BIBLIOGRAFIA

Ainsworth E. A., Long S. P. (2021) Carbon Dioxide Effects on Plants: Uncertainties and Implications for Modeling Climate-Ecosystem Interactions. Global Change Biology.

Aluchna M. (2012) Indeksy społecznej odpowiedzialności biznesu. Przypadek RESPECT Index" [w:] Płoszajski P. (red.) Społeczna odpowiedzialność biznesu w nowej gospodarce, SGH w Warszawie, Warszawa.

Fiszeder P. (2001) Zastosowanie modeli GARCH w analizie krótkookresowych zależności pomiędzy Warszawską Giełdą Papierów Wartościowych a międzynarodowymi rynkami akcji. Przegląd Statystyczny, 48, 345-364.

Janicka M., Miziołek T. (2022) Finanse zrównoważone. PWE, Warszawa.

Matilla-García M., Ruiz Marín M. (2021) Multifractal Analysis of European Stock Markets During the COVID-19 Crisis. Physica A: Statistical Mechanics and its Applications, 583, 126274.

Trenberth K. E., Fasullo J. T., Shepherd T. G. (2018) Attribution of Extreme Weather Events in the Context of Climate Change. Nature Climate Science.

Wiśniewski T. K. (2010) Indeks RESPECTIndex jako inicjatywa Giełdy Papierów Wartościowych w procesie tworzenia zasad CSR na polskim rynku kapitałowym. Polityki Europejskie, Finanse i Marketing, 4(53), 301-311.

Zhang L., Li Y. (2022) Tail Risk Dependence between Cryptocurrency and Traditional Financial Markets. International Review of Financial Analysis, 81, 102103.

CONSTRUCTION OF A CLIMATE TRANSITION INDEX FOR THE WARSAW STOCK EXCHANGE

Abstract: Climate change represents one of the most significant challenges of the contemporary world. These phenomena result from both natural climatic cycles and, to an increasing extent, anthropogenic activities, particularly greenhouse gas emissions. Therefore, the role of financial institutions is crucial, as they can accelerate economic transformation toward climate neutrality through capital redistribution. Climate transition indices cover companies that support climate transition constitute one of the financial instruments that can facilitate such efforts, serving as the underlying basis for ETFs. As a result of the conducted research, a concept for a climate transition index for companies listed on the Warsaw Stock Exchange was developed, and its properties were verified through analysis using selected time series analysis methods. Based on the obtained results, it was determined that the index can serve as an underlying instrument for ETFs, possessing similar characteristics

to the WIG index while maintaining superior performance characteristics compared to other WSE indices.

Keywords: ESG, WSE, ETF, Climate transition

JEL classification: G11

REVIEWERS COOPERATING WITH THE JOURNAL IN 2025

WARSAW UNIVERSITY OF LIFE SCIENCES – SGGW

Ihor Atamaniuk
Bolesław Borkowski
Nina Drejerska
Hanna Dudek
Michał Gostkowski
Urszula Grzybowska
Elżbieta Kacperska
Krzysztof Karpio
Joanna Kisielińska
Marek A. Kociński
Monika Krawiec
Serhiy Zabolotnyy

OTHER UNIVERSITIES

Agata Boratyńska (Warsaw School of Economics - SGH)
Joanna Górka (Nicolaus Copernicus University in Toruń)
Maciej Gurbała (Warsaw School of Economics - SGH)
Beata Jackowska Zduniak (Gdańsk University of Technology)
Barbara Kowalczyk (Warsaw School of Economics - SGH)
Maciej Kozłowski (University of Lodz)
Malgorzata Markowska (Wroclaw University of Economics and Business)
Bartłomiej Marona (Krakow University of Economics)
Krzysztof Olszewski (Warsaw School of Economics - SGH)
Magdalena Osińska (Nicolaus Copernicus University in Toruń)
Barbara Pawełek (Krakow University of Economics)
Mariola Piłatowska (Nicolaus Copernicus University in Toruń)
Mariusz Rafało (Warsaw School of Economics - SGH)
Paweł Rubach (Warsaw School of Economics - SGH)
Ewa M. Syczewska (Warsaw School of Economics - SGH)
Małgorzata Tarczyńska-Łuniewska (University of Szczecin)
Marcin Wawryszczuk (Vistula University)
Anna Wójtowicz (Warsaw School of Economics - SGH)
Wiktor Zbyrowski (Warsaw University)