



Climate Neutrality of the Economy and the Green Transition

Olena ZHYTKEVYCH,
Andriy MATVIYCHUK

DEVELOPING A HYBRID METHODOLOGICAL APPROACH TO FORECASTING A COUNTRY'S DECARBONIZATION POTENTIAL

Abstract

The high variability of data on decarbonization potential and the poor formalization of these processes require modern hybrid analysis methods. The study proposes an approach that combines Kohonen self-organizing maps (SOM) to identify patterns between countries, as well as linear regression (LR) and a multilayer perceptron (MLP) neural network to forecast the average CO₂ emission within clusters. The analysis covers 45 countries over a ten-year period and is based on 12 explanatory indicators. Thus, the scientific methodology based on the integration of clustering and neural network forecasting to model different national decarbonization trajectories. Kohonen maps allowed authors to identify seven groups of countries, which confirms the heterogeneity of the impact of factors and patterns of decarbonization across countries. The research results showed that linear models have limited explanatory power (R^2 of 0.30 on test data before clustering and 0.61 in average after clustering), while the optimized neural network provides higher forecast accuracy (average R^2 for all clusters of 0.75 on test data) supporting cluster-based modeling and ability of neural networks to capture nonlinear relationships. The outcomes of the research can be used by policymakers to design and tailor national climate policy and plan energy transitions.

© Olena Zhytkevych, Andriy Matviychuk, 2026.

Zhytkevych Olena, PhD (Economics), Associate Professor, College of Business, American University of Kurdistan, Duhok, Iraq. ORCID: 0000-0003-2042-8795 Email: olena.zhytkevych@auk.edu.krd
Matviychuk Andriy, DSc (Economics), Professor, Department of Artificial Intelligence, Modeling and Statistics, Kyiv National Economic University named after Vadym Hetman, Kyiv, Ukraine. ORCID: 0000-0002-8911-5677 Email: matviychuk@kneu.edu.ua

Key Words:

CO₂ emissions, data science, green transformation, Kohonen map, neural networks, nonlinear modeling.

JEL: Q50, C12, C38, C53, C55.

2 formulas, 2 tables, 4 figures, 8 appendices, 36 references.

Problem Statement

For the European economy, the key challenges include implementing climate neutrality policies, transitioning to a low-carbon development model and aligning national climate strategies with long-term decarbonization goals.

Effective implementation of low-carbon policies depends on strategic forecasting of decarbonization potential, which includes the selection, assessment and interpretation of environmental, energy, socio-economic and other data to substantiate management decisions. However, European countries differ significantly in terms of energy consumption structure, level of economic development, energy intensity, electrification, etc. This structural heterogeneity leads to differences in current levels of carbon intensity and potential decarbonization trajectories. Therefore, applying a single forecasting model to all countries may not sufficiently take into account the specific patterns of individual groups of countries. Therefore, continuous development of the theoretical, methodological and practical foundations of decarbonization forecasting is critical for the proper development of the green economy.

Literature Review

Accurately projecting the level of decarbonization at the national level is important both for individual countries and from a global perspective. However, this is a challenging task due to the nonlinear behavior of time series of average emission levels and unpredictable events. Current research uses a wide range of methods, including statistical models, such as spatio-temporal visualization approach, stationary time series (Encalada-Malca et al., 2021; Solci et al., 2023), machine learning algorithms (Zhu et al., 2018), artificial neural networks (Feng et al., 2015; Cordova et al., 2021; Fu et al., 2015; Kourentzes, et al., 2014; Wu et al., 2011; Paliwal & Kumar 2009), deep learning-based method (Li et al., 2020) and other hybrid approaches to solve such problems.

Multilayer perceptrons allow modeling complex relationships and have been used to estimate the impact of renewable energy sources on CO₂ emissions (Mussa & Khalifa, 2025), but they are resource-intensive and prone to overfitting (Haykin, 2009). To account for long-term dependencies, LSTMs are used, in particular, to

predict CO₂ emissions in China and other countries (Wen et al., 2023; Kumari & Singh, 2023), CNN–LSTMs have been applied to complex time series in ecological systems (Ding et al., 2021; Han et al., 2023). But they require significantly more computing resources compared to MLP.

Zheng et al. (2013) inferred the real-time and fine-grained air quality information throughout a city, based on the (historical and real-time) air quality data reported by existing monitor stations and a variety of data sources they observed in the city by using artificial neural network. Although these methods require significant computational resources and large amounts of data. Ensemble learning methods, such as Random Forest, are used to predict greenhouse gas emissions in energy transition scenarios (Singh et al., 2026), but are also resource-intensive method (Breiman, 2001; Bielinskyi et al., 2026). Hybrid models, such as ARIMA combined with MLP or XGBoost, allow for simultaneous modeling of trends and nonlinearities, which were used to forecast CO₂ emissions from megacities in China or countries with economies in transition (Zhang et al., 2024; Zhang, 2003), but require complex model construction and optimization. The Prophet method has shown effectiveness for predicting nonlinear dependencies in urban emission models and energy scenarios (Hasnain et al., 2022), but is sensitive to hyperparameter tuning (Taylor & Letham, 2017).

Our literature review shows that forecasting methods for decarbonization have different suitability and limitations: high computational costs, complex setup and risk of overfitting (given the short time series for individual countries). Hybrid models combining statistical analysis and machine learning are most promising for countries with transformational economies and a transitional energy landscape. The accuracy and stability vary across scales and data types. Resource intensity, the need for big data and low interpretability complicate adaptation, requiring the development of more universal and effective approaches to forecasting decarbonization potential.

The aim of the article is to develop and empirically test a hybrid approach to forecasting decarbonization potential, which involves the combined use of country clustering (Kohonen maps) and intra-cluster modeling using a multilayer perceptron and linear regression.

To achieve the goal, it is necessary to solve the following tasks:

1. data preprocessing and hypothesis testing on key indicators – perform pre-processing of data (logarithmization of the target variable, multicollinearity analysis) and test the hypothesis regarding the significance of key factors;
2. developing of a hybrid approach integrating clustering and nonlinear modeling of dynamic indicators: (2.1) clustering of countries by structural characteristics using SOM; (2.2) comparative evaluation of regression and MLP models (taking into account cluster structure).

Methodology

The study applies a hybrid quantitative methodology combining data preprocessing, statistical diagnostics, clustering, and predictive modeling. A panel dataset of 45 countries for 2014–2023 includes 12 explanatory indicators and the average CO₂ emission factor as the target variable. Data were cleaned, normalized

to [0–1], and the target variable was logarithmically transformed. Multicollinearity was assessed using Pearson correlation and VIF. Country profiles were clustered using Kohonen self-organizing maps. Linear regression and multilayer perceptron models were then estimated for the full sample and within clusters using a 70/30 train–test split and evaluated by R^2 , MAE, and MSE.

Research Results

Task 1. Data preprocessing and hypothesis testing on key indicators.

Forecasting is a form of scientific prediction with two aspects: descriptive, which assesses possible future scenarios, and prescriptive, which supports management decisions. It reduces uncertainty and identifies the consequences of actions or inactions, providing a cause-and-effect relationship between the current state and development prospects. An effective decarbonization policy of a country requires tools to assess current potential and predict its dynamics, taking into account internal and external factors.

The study examines the decarbonization potential of 45 countries over the period 2014–2023. The data collection includes 13 annual indicators from the socio-economic, energy, environmental, globalization, and geopolitical sectors, which together reflect a country's decarbonization potential. The indicators were selected from EnerData (2026), World Bank (n.d.), and Geopolitical Risk Index (n.d.) based on multicollinearity and significance analysis in previous studies (Zhytkevych et al., 2022, 2023). The sample covers OECD, G7, BRICS, EU, CIS, and regions of North and Latin America, Asia, Australia, Africa, and the Middle East. Limiting the selection to 45 countries ensures global representativeness and includes participants with varying levels of engagement in the decarbonization initiative, all of whom are characterized by sufficient availability of relevant statistical data.

Research of Banna et al. (2023) shows that geopolitical tensions, in particular Russia's invasion of Ukraine, destabilize energy systems, financial markets and investments in green technologies, shifting priorities to security and defense. Geopolitical risks are a critical variable for assessing global decarbonization. The Geopolitical Risk Index (GRI), described by Caldara & Iacoviello (2022), based on the analysis of media coverage of conflicts, reflects their impact on the global geopolitical landscape and economic dynamics. In this study, GRI is considered a key factor in predicting decarbonization potential, as it directly affects the energy system and the ability of a country to reduce CO₂ emissions. It was decided to retain one country missing one indicator (GRI): Uzbekistan. To maintain completeness, the GRI value for Uzbekistan was replaced with the average value for the Central Asia region (Kazakhstan, Kyrgyzstan, Tajikistan) over 10 years, which minimizes error. Each country's data is presented as a multivariate time series in our dataset (D):

$$D = \{X^k = x_{i,j}^k \mid k = 1, \dots, K; i = 1, \dots, n; j = 1, \dots, m\} \quad (1)$$

where $k = 1, \dots, K$ is the country index ($K = 45$); $i = 1, \dots, n$ is the time index for annual observations ($n = 10$); $j = 1, \dots, m$ is the indicator index ($m = 13$).

The dataset was split into training and testing (validation) subsets according to a 70/30 ratio in all our further modeling experiments.

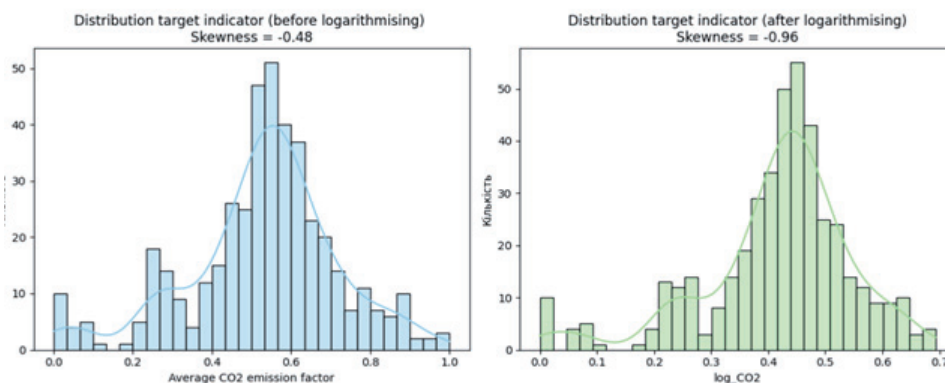
Among the 13 selected indicators, the target variable is the average CO₂ emission factor. The country's decarbonization potential is interpreted as its estimated ability to reduce CO₂ emissions, which is formed under the influence of factors such as the level of technological development, renewable energy capacity, energy efficiency, as well as economic and demographic growth. Thus, the specified indicator can be considered as an integrated indicator that reflects the cumulative impact of other determinants. The remaining 12 indicators are interpreted as explanatory variables that affect the decarbonization potential with varying intensity. It is hypothesized that some of these factors may have an ambiguous impact on the decarbonization potential of countries, but the vast majority of them are expected to have a direct impact. Taken together, they represent the key areas that shape the country's decarbonization potential and serve as the basis for formulating research hypotheses (see Appendix A).

Let Y_i^k denote the decarbonization potential of country k in year i , where Y_i^k – Average CO₂ emission factor ($x_{i,13}^k$) and $X_i^k = (x_{i,1}^k, \dots, x_{i,12}^k)$ is the vector of 12 explanatory indicators listed in Appendix A. Thus, general hypothesis formulated as null hypothesis $H_0: \beta_1 = \beta_2 = \dots = \beta_{12} = 0$ – none of the indicators has a statistically significant effect on Y_i^k and alternative hypothesis $H_1: \exists j \in \{1, \dots, 12\}, \beta_j \neq 0$ – at least one indicator has a statistically significant effect on Y_i^k .

After data preparation (cleaning, formatting, filling gaps, normalizing to [0–1]) and determining the target index, descriptive statistics for 13 indicators are calculated and presented in Appendix B. The mean CO₂ emission factor (x_{13}) is 0.531, close to the median, a standard deviation of 0.186, showing the difference in emission intensity between countries. Analysis of the distribution and skewness revealed the need to transform the target change: CO₂ was logarithmized with a constant added to avoid zero values, which shortened the distribution and reduced the impact of extreme values. The transformed measure is denoted as $\log_CO_2$ (Figure 1).

Figure 1

Distribution of the Average CO₂ emission factor before and after logarithmization



Source: calculated by the authors.

To assess the significance and direction of influence of the selected explanatory variables, we employed a multiple linear regression model. First, we specified the theoretical regression model, which takes the general form: $\log_CO_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{12} x_{12} + \varepsilon$, where β are the regression coefficients and ε is the error term.

Using the training dataset and implementing multiple linear regression in Power BI via Python scripting employing the *Linear Regression()* method from the *scikit-learn* package. The model parameters were estimated using the Ordinary Least Squares method, which minimizes the sum of the squared deviations of the predicted values from the actual ones. The following estimated regression equation was obtained:

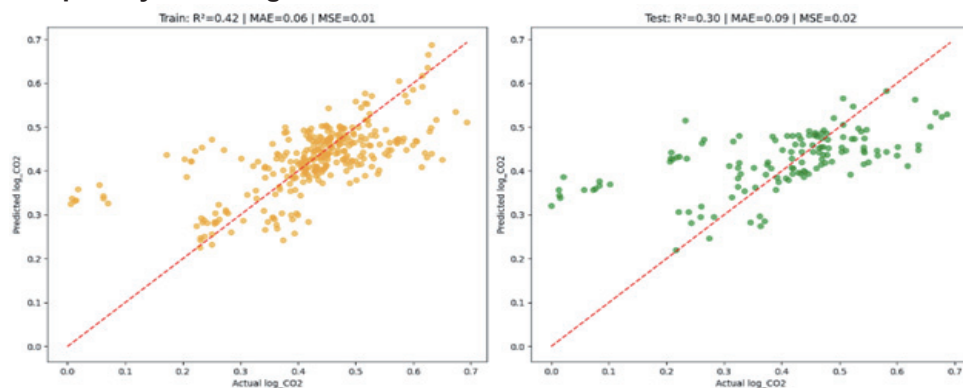
$$\hat{y} = 0.258 - 0.019x_1 + 0.048x_2 - 0.846x_3 - 0.305x_4 + 0.137x_5 - 0.002x_6 + 0.149x_7 - 0.786x_8 + 0.200x_9 - 0.051x_{10} + 0.162x_{11} - 0.020x_{12} \quad (2)$$

The list of indicators can be seen below in Table 1. The equation (2) quantifies the marginal effect of each explanatory variable on the logarithm of the Average CO₂ emission factor, holding all other 12 indicators constant.

The model's performance (see Figure 2) was evaluated on the training dataset $R^2 = 0.418$, $MAE = 0.060$, $MSE = 0.008$ and on the test dataset $R^2 = 0.299$, $MAE = 0.090$, $MSE = 0.016$, indicating moderate explanatory power for the training subset and even lower performance for the test subset. The overall significance of the model was confirmed by the general F-test ($p \approx 3.6 \times 10^{-31} < 0.001$), which strongly rejected the null hypothesis that all regression coefficients are equal to zero. Hypothesis testing based on the sign of β (see Appendix A) and the level of significance (p-value < 0.1) was conducted on the full dataset to describe the current empirical relationship, hence, the model was estimated on all observations to maximize information usage and obtain more accurate coefficient estimates (see Table 1).

Figure 2

The ratio of actual and predicted values of log_CO₂ in the training and test samples by linear regression model



Source: calculated by the authors.

Table 1

Testing hypothesis on full dataset

Indicator	β	p-value	Expected sign (H1)	Conclusions
Energy intensity of GDP at consumption	0.03	0.47062	>0	Not significant, H1 rejected
Oil products domestic consumption	0.19	0.34053	>0	Not significant, H1 rejected
Natural gas domestic consumption	0.42	0.00192	>0	Significant, H1 accepted
Electricity domestic consumption	-1.56	0.00038	Any	Significant, H1 accepted
Share of renewables in electrification	-0.49	1×10^{-5}	<0	Significant, H1 accepted
Share of wind and solar in electrification	0.30	1×10^{-5}	<0	Significant, but opposite sign
Urban population growth (annual %)	0.05	0.55406	Any	Not significant, H1 rejected
GDP per capita growth (annual %)	0.14	0.06451	Any	Significant, H1 accepted
Coal and lignite domestic consumption	1.36	1×10^{-5}	>0	Significant, H1 accepted
Electrification	0.36	1×10^{-5}	Any	Significant, H1 accepted
Geopolitical Risk Index	-0.06	0.29281	Any	Not significant, H1 rejected
Globalization Index	0.28	0.00008	Any	Significant, H1 accepted

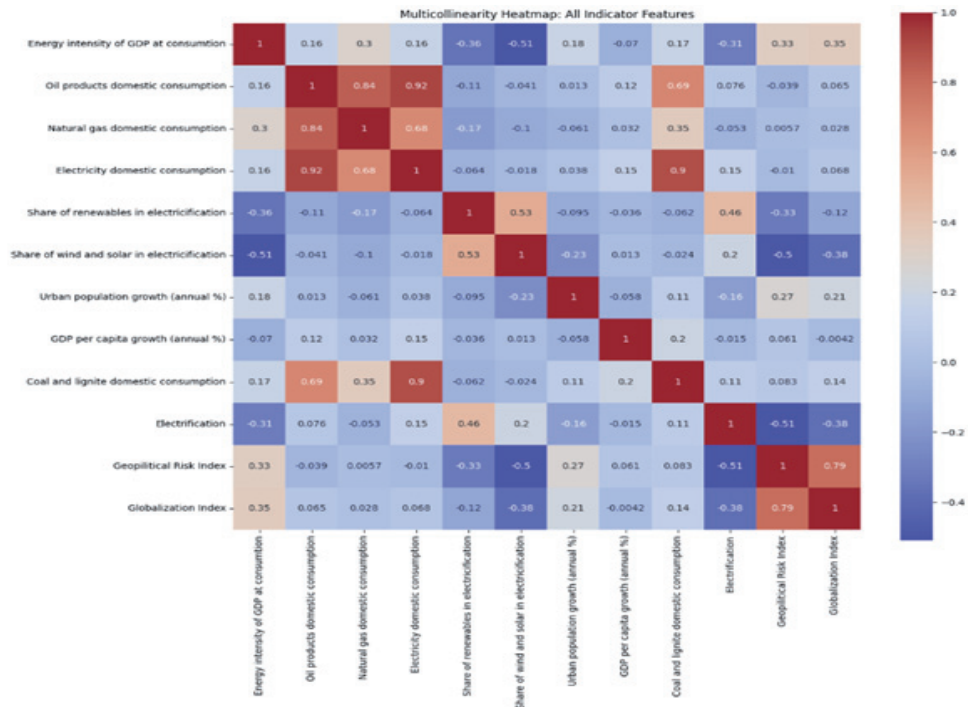
Source: calculated by the authors.

Table 1 presents the regression results, where seven indicators are significant and confirmed H1 (highlighted in bold), four indicators are not statistically significant ($p > 0.1$), even if the sign of β corresponds to the hypothesis, and one (Share of wind and solar in electrification) is significant, but the sign contradicts the hypothesis. The general hypothesis test is extremely significant and we rejected H0 and conclude that at least one of the 12 indicators has a statistically significant relationship with $\log_CO_2$.

Although the overall model shows explanatory power, the non-significance of the individual-level predictor residuals may be due to multicollinearity or specification issues that require variable transformation. In this study, significant indicators were retained by a combination of correlation analysis and Variance Inflation Factor (*VIF*) filtering, in line with current recommendations for preventing multicollinearity in emission forecasting and energy modeling models (Fuinhas et al., 2025). To assess multicollinearity among the 12 decarbonization indicators, a correlation heat map was constructed and the *VIF* was calculated (Belsley et al., 1980). The Correlation Heatmap (multicollinearity diagnosis) was constructed in Microsoft Power BI using Python 3.10 script with the libraries *pandas*, *matplotlib* and *seaborn*. It visualizes Pearson correlation coefficients between all pairs of indicators (see Figure 3).

Figure 3

Multicollinearity heatmap with 12 indicators



Source: calculated by the authors.

Based on the calculated correlation coefficients for 12 indicators (Figure 3), variables with high multicollinearity were identified. Next, VIF was calculated in Microsoft Power BI via Python (statsmodels), which measures the overestimation of the variance of the coefficients due to collinearity ($VIF < 5$ – low, between 5 and 10 – moderate, and greater 10 – high multicollinearity). Initial values of VIF for domestic electricity consumption (149.62) and GDP growth per capita (114.61) indicated strong collinearity.

Feature selection was carried out by stepwise removal of variables with the highest VIF until the VIF for all remaining indicators was ≤ 10 , which reduced redundancy and preserved the diversity of information. As a result, 9 indicators were excluded: Domestic electricity consumption, GDP per capita growth, Urban population growth, Domestic petroleum consumption, Globalization index, Domestic coal and lignite consumption, Geopolitical risk index, Electrification and Domestic gas consumption.

Removing variables solely on the basis of statistical multicollinearity requires expert control, since the real relationships between decarbonization factors may be nonlinear or context-dependent.

Task 2. Developing of a hybrid approach integrating clustering and nonlinear modeling of dynamic indicators.

2.1. Clustering of countries by structural characteristics using SOM.

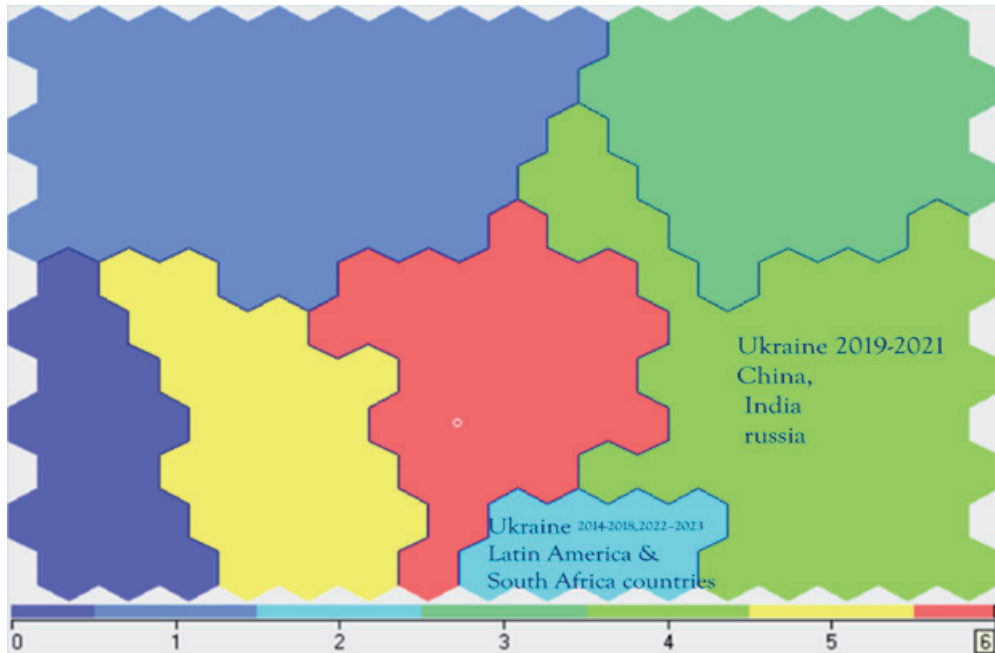
The next stage of the study is the analysis of the nonlinear relationships between indicators that cannot be fully determined by linear diagnostic methods. Here it is important to consider significant differences in the energy and economic development of various countries. For example, the differences in the development dynamics of China, the Netherlands, and Uzbekistan (not to mention many other countries not included in the research sample) seem quite obvious. Therefore, it makes sense to identify groups of similar countries and build predictive models within these groups. Therefore, we will cluster countries from the overall sample and construct forecasting models (linear regressions and MLP) for each cluster, as well as single models (also LR and MLP) for all countries, and compare their performance to justify the feasibility of this approach.

A series of experiments showed that the optimal number of indicators is 12. Clustering based on all 12 variables provided more stable clusters, so the complete set of indicators was retained in subsequent analyses. This is also consistent with the properties of SOMs, which work effectively with high-dimensional and nonlinearly related data, warming up the topological and metric structure of the input space (Peeters & Dassargues, 2006; Milone et al., 2013; Agrafiotis & Xu, 2002).

Comparing the linear and MLP models at the cluster level has several advantages. First, it allows us to check whether the nonlinear MLP model takes into account additional information that the linear model cannot capture, especially in groups of countries with nonstandard or extreme indicator profiles. Second, analysis within clusters helps to identify possible local heterogeneity in the data and assess how adaptive the models are to different contexts. Finally, such clustering allows integrating an intuitive spatial representation of the data (via SOM) with quantitative modeling results, which increases the confidence in the conclusions and their applied relevance in decarbonization policies and energy planning.

Thus, the proposed approach ensures an objective, effective and transparent decision-making process in the field of decarbonization. In our study, countries were clustered based on a set of 12 indicators of decarbonization potential (Appendix A for the indicator set; Appendix C for the clustering outcomes), with the clustering methodology building on approaches developed and successfully tested in the work of Zhytkevych & Brochado (2022). The database for analysis was updated, revised and expanded in the field of decarbonization by including indicators that take into account risk and globalization aspects (reasons mentioned above in the text) as well as expanding the sample to 45 countries, which increased the accuracy of the assessment of national potential for reducing CO₂ emissions.

Figure 4

Self-organizing map of seven clusters for 45 countries based on 12 indicators

Source: calculated by the authors.

It is noticeable that the results of clustering showed significant heterogeneity of European countries in terms of energy, economic and other characteristics that determine the potential for decarbonization. The identified clusters are characterized by both stable membership of individual countries and transitions between groups during 2014–2023. For example, cluster 0 is characterized by stable membership of Norway and Sweden throughout the decade. Cluster 1 is composed by the Czech Republic and France in 2014–2023, while Germany, Belgium, the Netherlands, Poland and the United Kingdom were in it during separate periods. Cluster 5 unites a number of European countries, in particular Denmark and Portugal throughout the period, as well as Spain, Germany, the Netherlands, Belgium, Italy and the United Kingdom in separate years. Cluster 6 includes Romania during 2014–2023, as well as Italy, Germany, Spain and Poland in separate periods (Appendix C).

The ability to observe country transitions between clusters using the SOM model illustrates its versatility and sensitivity to changes in energy structures and economic development. For example, Ukraine stayed in cluster 2 (in the lower center of the SOM, the number of which is identified by the color-coded ruler below the map) in 2014–2018 together with Latin America and South Africa countries, then moved to cluster 4 (lower right corner of the map) in 2019–2021, where China, India and Russia were based, and returned back to cluster 2 in 2022–2023. Similar transitions

were observed for Mexico, India and other countries, reflecting the evolution of their energy systems. Thus, SOM helps analysis (decision-makers) take into account both general trends and changes specific to each country.

2.2. Comparative evaluation of regression and MLP models (taking into account cluster structure).

Following clustering, the study proceeds to construction of forecasting models at two levels: for the full dataset and for each identified cluster separately. Cluster-specific models are estimated based on observations of countries belonging to a given cluster over analyzed period, which makes it possible to assess the feasibility of clustering as a tool for detecting heterogeneity between countries.

The neural network was then implemented and evaluated and visualized in Power BI with Python 3.10 script using libraries: scikit-learn, numpy, pandas and matplotlib. The set of explanatory factors consisted of 12 indicators, while the average CO₂ emission factor served as the dependent variable (Appendix A), and the dataset was randomly divided into a training subset and a testing subset for each identified cluster.

Thus, the optimization of network parameters was performed by conducting experiments with systematic changes in the architecture (number of hidden layers and neurons per layer) and the number of training iterations. Comparative experiments outcomes showed that the most stable performance in terms of R², MAE and MSE was achieved with an architecture with 4 hidden layers. The fragment of the Python script represents this function in Appendix D. In particular, the architecture of [12, 5, 6, 1] provided the most consistent approximation of nonlinear dependencies and model convergence on the train and test samples of the full dataset.

For comparison, the linear model showed significantly lower results than MLP (see Table 2). This confirms that traditional linear methods do not allow to fully reflect the complexity of the studied processes. Thus, even despite partial limitations, the MLP indicators turned out to be significantly better, which emphasizes the advantage of neural networks in modeling complex relationships.

Table 2

Performance comparison of LR and MLP models on full dataset

Model	Train data			Test data		
	R ²	MAE	MSE	R ²	MAE	MSE
LR (12 indicators)	0.418	0.060	0.008	0.299	0.090	0.016
MLP (12 indicators)	0.851	0.041	0.003	0.744	0.043	0.003

Source: calculated by the authors.

The results of a series of experiments conducted (based on the methods described above) showed that the quality of the model depends significantly not only on the number of indicators but also on its architecture. Therefore, experimental studies allowed us to more accurately assess the impact of configurations on the effectiveness of the model and determine that the optimal MLP architecture for 12 indicators on the full dataset.

The next step was building separate MLP and linear models for each cluster of countries, which allows us to improve prediction accuracy and take into account the specifics of intra-cluster dependencies. For the 7 clusters obtained above using SOM, LR and MLP models were calculated based on Python code using libraries: scikit-learn, numpy, pandas and matplotlib. In this case, it was conducted a searching for the optimal architecture of the predictive neural network model of each cluster, as a result of which the optimal structure was determined to be [12, 6, 4, 1] for clusters 0, 4 and 5, [12, 9, 6, 1] for clusters 1 and 3, [12, 8, 6, 1] for cluster 2 and [12, 4, 3, 1] for cluster 6, respectively. An assessment of the forecasting accuracy for seven clusters based on LR and MLP with the optimal architecture for each cluster is presented graphically in Appendix E and summarized in Table E.1.

The last column in Table E.1 contains the average values of the performance metrics for LR and MLP over all 7 clusters. Analysis of the results of regression and MLP modeling for seven clusters (0–6) demonstrated differences in forecasting accuracy depending on the type of model and cluster characteristics. In general, MLP models showed higher values of the coefficient of determination R^2 and lower values of MAE and MSE compared to linear regressions (in particular for test datasets), indicating a greater ability of neural networks to reproduce nonlinear relationships between energy and economic indicators.

This suggests that the relationships between indicators vary among countries' clusters, thereby supporting the use of cluster-based forecasting models.

Thus, after clustering, MLP shows advantage (in terms of test sample performance) for clusters 2, 3, 4 and 6, confirming its ability to capture complex nonlinear relationships more effectively than LR. In contrast, clusters 0, 1 and 5 show limited or absence of MLP advantage, indicating that simpler linear models remain sufficient for these groups, and also suggesting potential structural homogeneity or higher variability that limits prediction accuracy. The significant indicators identified for each cluster and the corresponding recommended modeling methods are summarized in Appendices F, G. The cluster-specific hypothesis-testing results, based on the estimated relationships between the explanatory indicators and the target variable, are presented in Appendix H.

Discussion

The analysis of regression equations, which are presented in Appendix F, shows the specific influence of individual indicators in each of 7 clusters, confirming the presence of pronounced heterogeneity in the structure of indicators across the overall dataset. Thus, table in Appendix G summarizing key indicators for each cluster and recommended modeling methods based on a comparative analysis of model

performance and indicator relevance. Indicators were selected based on forecasting performance (R^2 , MAE, MSE) and their consistency across the best models in each cluster. This analysis confirms that CO_2 emission factors vary significantly across clusters – from traditional fossil-fuel dependence to renewable energy adoption, techno-institutional transformation and socio-economic drivers, underscoring the need for cluster-specific modeling strategies.

By analogy with the testing of the hypotheses described in Appendix A for the regression model built on the full dataset (see Table 1), we will test these hypotheses for the regression models built for each of the 7 clusters. Similarly, both p-value and sign of β were considered when testing the hypotheses. A hypothesis was considered confirmed only if estimated effect was statistically significant ($p < 0.1$) and matched the expected sign (Appendix A), hence highlighted in bold in Appendix H. The results of the analysis of the calculations in Appendix H generally confirmed the conclusions obtained for the full data set (Table 1). Although of the 12 hypotheses put forward in Appendix A regarding the influence of the factors under study on the dependent variable, all 12 were partially confirmed in at least one of the clusters, only 7 hypotheses were realized in the majority of clusters. For example, the insignificant x_1 indicator “Energy intensity of GDP at consumption” in the model on the full data set (see Table 1) turned out to be significant for only one cluster out of seven (cluster 6, which can be seen in Appendix H), which also confirms its overall insignificance.

Hence, the obtained results confirm the feasibility of using the hybrid approach, which combines clustering of countries using Self-organizing Kohonen maps and intra-cluster modeling. The scientific novelty of the approach lies in the integration of clustering and scenario modeling into a single sequential procedure. SOM allows to identify relatively homogeneous groups of countries by a set of energy, economic and technological characteristics, while the use of MLP within the formed clusters makes it possible to take into account, respectively, nonlinear dependencies and ensure the interpretability of the relationships between indicators.

The outcomes are important and can be implemented by European countries, which are characterized by different energy structures, levels of economic development and rates of transition to a low-carbon economy.

Conclusions

Decarbonization is complex and heterogeneous process across countries due to their varying socio-economic, energy and technological potential. The availability of short time series of data and rapid changes in energy consumption for individual countries significantly complicate the accurate modeling of CO_2 emissions in dynamic. To eliminate these limitations, the hybrid model was proposed that combines clustering with nonlinear forecasting and provides more accurate assessment of trends in order to support the development of effective sustainable development strategies. A dataset of 45 countries for 2014–2023 was compiled using 12 indicators and the average CO_2 emission factor as a target variable, combining technological, energy, economic and demographic factors. Data preparation included standardization, gap filling, and logarithmization of the target change to reduce the impact of extreme values.

Multicollinearity analysis and expert assessment of the clustering results ensured the preservation of 12 significant indicators. Multiple linear regression showed moderate explanatory power ($R^2 = 0.42$ on training, 0.30 on test samples), the F-test confirmed the significance of the model. To detect complex nonlinear relationships, a multilayer perceptron with optimized architecture [12, 5, 6, 1] was used, which provides high enough accuracy ($R^2 = 0.85$ on training, 0.74 on test samples). After clustering, average metrics across 7 clusters further highlight the improvement in forecasting performance (the average R^2 for LR was 0.61, MAE = 0.06, MSE = 0.01, and for MLP $R^2 = 0.75$, MAE = 0.03, MSE = 1×10^{-4} on test data).

Clustering of countries based on the dataset revealed 7 clusters with different energy, economic and social structures. The impact of individual indicators differs between clusters, justifying a differentiated approach to modeling.

The results show that all 12 hypotheses were only partly confirmed for some clusters, hence not all factors have consistently significant impact (their direction and strength vary substantially between clusters), indicating heterogeneous relationships between indicators and CO2 emissions and supporting cluster modeling.

The practical application of the results obtained can serve as a basis for developing national decarbonization strategies, identifying priority areas for the development of renewable energy and energy efficiency, as well as for monitoring the progress of countries in the context of global sustainable development goals.

Future research includes the use of alternative neural networks and ensemble learning to provide cluster-based interpretation of model results and validation of a hybrid forecasting approach, improve forecasting accuracy, model sensitivity analysis, regional decarbonization forecasting, and integration into decision support systems for energy policy and climate strategies.

Acknowledgement

This work is a part of the applied research “Data Science Analysis of the Green Transformation of Global Food Security and Implementation of International Practices in Ukraine”, which is funded by the Ministry of Education and Science of Ukraine (project No. 0126U001738).

References

- Agrafiotis, D. K., & Xu, H. (2002). A self-organizing principle for learning nonlinear manifolds. *Proceedings of the National Academy of Sciences*, 99(25), 15869–15872. <https://doi.org/10.1073/pnas.242424399>
- Banna, H., Alam, A., Chen, X. H., & Alam, A. W. (2023). Energy security and economic stability: The role of inflation and war. *Energy Economics*, 126, Article 106949. <https://doi.org/10.1016/j.eneco.2023.106949>
- Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley. <https://doi.org/10.1002/0471725153>

- Bielinskyi, A., Soloviev, V., Matviychuk, A., & Bezkorovainyi, V. (2026). Stock market movement forecasting using machine learning and complex network measures. *Artificial Intelligence and Applications*. <https://doi.org/10.47852/bonviewAIA62027537>
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Caldara, D., & Iacoviello, M. (2022). Measuring geopolitical risk. *American Economic Review*, 112(4), 1194–1225. <https://doi.org/10.1257/aer.20191823>
- Cordova, C. H., Portocarrero, M. N. L., Salas, R., Torres, R., Rodrigues, P. C., & López-Gonzales, J. L. (2021). Air quality assessment and pollution forecasting using artificial neural networks in Metropolitan Lima–Peru. *Scientific Reports*, 11, Article 24232. <https://doi.org/10.1038/s41598-021-03650-9>
- Ding, C., Wang, G., Zhang, X., Liu, Q., & Liu, X. (2021). A hybrid CNN-LSTM model for predicting PM_{2.5} in Beijing based on spatiotemporal correlation. *Environmental and Ecological Statistics*, 28, 503–522. <https://doi.org/10.1007/s10651-021-00501-8>
- Encalada-Malca, A. A., Cochachi-Bustamante, J. D., Rodrigues, P. C., Salas, R., & López-Gonzales, J. L. (2021). A spatio-temporal visualization approach of PM₁₀ concentration data in Metropolitan Lima. *Atmosphere*, 12(5), Article 609. <https://doi.org/10.3390/atmos12050609>
- EnerData. (2026). *Total energy consumption. World Energy & Climate Statistics – Yearbook 2026* [Data set]. <https://yearbook.enerdata.net/total-energy/world-consumption-statistics.html>
- Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., & Wang, J. (2015). Artificial neural networks forecasting of PM_{2.5} pollution using air mass trajectory based geographic model and wavelet transformation. *Atmospheric Environment*, 107, 118–128. <https://doi.org/10.1016/j.atmosenv.2015.02.030>
- Fu, M., Wang, W., Le, Z., & Khorram, M. S. (2015). Prediction of particular matter concentrations by developed feed-forward neural network with rolling mechanism and gray model. *Neural Computing and Applications*, 26, 1789–1797. <https://doi.org/10.1007/s00521-015-1853-8>
- Fuinhas, J. A., Belucio, M., Santiago, R., & Betencourt, M. (2025). Are the structure dynamics of capital stock impacting carbon intensity from energy consumption? European insights. *Energy, Ecology and Environment*, 10, 555–574. <https://doi.org/10.1007/s40974-025-00370-y>
- Geopolitical Risk Index. (n.d.). *Final Geopolitical Risk Index* [Data set]. <https://www.geopriskindex.com/results-final-risk-index/>
- Han, Z., Cui, B., Xu, L., Wang, J., & Guo, Z. (2023). Coupling LSTM and CNN neural networks for accurate carbon emission prediction in 30 Chinese provinces. *Sustainability*, 15(18), Article 13934. <https://doi.org/10.3390/su151813934>
- Hasnain, A., Sheng, Y., Hashmi, M. Z., Bhatti, U. A., Hussain, A., Hameed, M., Marjan, S., Bazai, S. U., Hossain, M. A., Sahabuddin, M., Wagan, R. A., & Zha, Y. (2022). Time series analysis and forecasting of air pollutants based on Prophet forecasting model in Jiangsu Province, China. *Frontiers in Environmental Science*, 10, Article 945628. <https://doi.org/10.3389/fenvs.2022.945628>

- Haykin, S. S. (2009). *Neural Networks and Learning Machines* (3rd ed.). Pearson.
- Kourentzes, N., Barrow, D. K., & Crone, S. F. (2014). Neural network ensemble operators for time series forecasting. *Expert Systems with Applications*, 41(9), 4235–4244. <https://doi.org/10.1016/j.eswa.2013.12.011>
- Kumari, S., & Singh, S. K. (2023). Machine learning-based time series models for effective CO₂ emission prediction in India. *Environmental Science and Pollution Research*, 30(55), 116601–116616. <https://doi.org/10.1007/s11356-022-21723-8>
- Li, S., Xie, G., Ren, J., Guo, L., Yang, Y., & Xu, X. (2020). Urban PM_{2.5} concentration prediction via attention-based CNN–LSTM. *Applied Sciences*, 10(6), Article 1953. <https://doi.org/10.3390/app10061953>
- Milone, D. H., Stegmayer, G., Kamenetzky, L., López, P., & Carrari, F. (2013). Clustering biological data with SOMs: On topology preservation in non-linear dimensional reduction. *Expert Systems with Applications*, 40(9), 3841–3845. <https://doi.org/10.1016/j.eswa.2012.12.074>
- Mussa, A. A. Y., & Khalifa, W. M. S. (2025). MLP enhanced CO₂ emission prediction model with LWSSA nature inspired optimization. *Scientific Reports*, 15, Article 1891. <https://doi.org/10.1038/s41598-025-85709-5>
- Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications*, 36(1), 2–17. <https://doi.org/10.1016/j.eswa.2007.10.005>
- Peeters, L., & Dassargues, A. (2006). *Comparison of Kohonen's self-organizing map algorithm and principal component analysis in the exploratory data analysis of a groundwater quality dataset*. University of Liège. <https://orbi.uliege.be/bitstream/2268/3388/1/publi144-2007.pdf>
- Singh, S., Kumar, M., Verma, B. K., & Kumar, S. (2026). Optimizing air pollution prediction with random forest algorithm. *Aerosol Science and Engineering*, 10, 208–221. <https://doi.org/10.1007/s41810-025-00292-6>
- Solci, C. C., Reisen, V. A., & Rodrigues, P. C. (2023). Robust local bootstrap for weakly stationary time series in the presence of additive outliers. *Stochastic Environmental Research and Risk Assessment*, 37, 2977–2992. <https://doi.org/10.1007/s00477-023-02430-3>
- Taylor, S. J., & Letham, B. (2017). Forecasting at scale. *PeerJ Preprints*, 5, Article e3190v2. <https://doi.org/10.7287/peerj.preprints.3190v2>
- Wen, T., Liu, Y., Bai, Y. H., & Liu, H. (2023). Modeling and forecasting CO₂ emissions in China and its regions using a novel ARIMALSTM model. *Heliyon*, 9(11), Article e21241. <https://www.sciencedirect.com/science/article/pii/S2405844023084499>
- World Bank. (n.d.). *World Development Indicators* [Data set]. <https://datacatalog.worldbank.org/search/dataset/0037712>
- Wu, S., Feng, Q., Du, Y., & Li, X. (2011). Artificial neural network models for daily PM₁₀ air pollution index prediction in the urban area of Wuhan, China. *Environmental Engineering Science*, 28(5), 357–363. <https://doi.org/10.1089/ees.2010.0219>

- Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159–175. [https://doi.org/10.1016/S0925-2312\(01\)00702-0](https://doi.org/10.1016/S0925-2312(01)00702-0)
- Zhang, Y., Wang, J., Chen, R., Zhang, S., Liao, X., & Xie, J. (2024). Complementary ARIMA and XGBoost carbon emission timing forecasting methods for different terminal energy consumption industries. In *Fifth International Conference on Green Energy, Environment, and Sustainable Development (GEESD 2024)* (Vol. 13279, Article 1327946). SPIE Digital Library. <https://doi.org/10.1117/12.3044690>
- Zheng, Y., Liu, F., & Hsieh, H. P. (2013). U-Air: When urban air quality inference meets big data. In R. Ghani, T. E. Senator, P. Bradley, R. Parekh, & J. He, (Eds.), *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1436–1444). Association for Computing Machinery. <https://doi.org/10.1145/2487575.2488188>
- Zhu, S., Lian, X., Wei, L., Che, J., Shen, X., Yang, L., Qiu, X., Liu, X., Gao, W., Ren, X., & Li, J. (2018). PM_{2.5} forecasting using SVR with PSOGSA algorithm based on CEEMD, GRNN and GCA considering meteorological factors. *Atmospheric Environment*, 183, 20–32. <https://doi.org/10.1016/j.atmosenv.2018.04.004>
- Zhytkevych, O., & Brochado, A. (2022). Modeling national decarbonization capabilities using Kohonen maps. *Neuro-Fuzzy Modeling Techniques in Economics*, 11, 3–24. <http://doi.org/10.33111/nfmte.2022.003>
- Zhytkevych, O., Matviychuk, A., & Kmytiuk, T. (2023). Modeling Nations' Decarbonisation Potential. In F. Ortiz-Rodríguez, S. Tiwari, P. Usoro Usip, & R. Palma (Eds.), *Electronic Governance with Emerging Technologies. EGETC 2023. Communications in Computer and Information Science* (Vol. 1888, pp. 60–77). Springer, Cham. https://doi.org/10.1007/978-3-031-43940-7_6

Appendix A

Table A1

The list of indicators in the decarbonization context and a set of associated research hypotheses

Category	Indicator for country	Source	Expected effect	Research hypothesis
Energy	Energy intensity of GDP at consumption (x1)	EnerData, 2024	Positive	H0a: $\beta_1 < 0$ H1a: $\beta_1 > 0$
	Oil products domestic consumption (x2)		Positive	H0b: $\beta_2 < 0$ H1b: $\beta_2 > 0$
	Natural gas domestic consumption (x3)		Positive	H0c: $\beta_3 < 0$ H1c: $\beta_3 > 0$
	Electricity domestic consumption (x4)		Mixed: depends on energy mix, for example, increase in clean energy share (positive) but in fossil-heavy (negative), vice versa is true	H0d: $\beta_4 = 0$ H1d: $\beta_4 \neq 0$
	Share of renewables in electrification (x5)		Negative	H0e: $\beta_5 > 0$ H1e: $\beta_5 < 0$
	Share of wind and solar in electrification (x6)		Negative	H0f: $\beta_6 > 0$ H1f: $\beta_6 < 0$
	Electrification (x7)		Mixed: depends on energy mix, for example, increase in clean energy share (positive) but in fossil-heavy (negative), vice versa is true	H0g: $\beta_7 = 0$ H1g: $\beta_7 \neq 0$
	Coal and lignite domestic consumption (x8)		Positive	H0h: $\beta_8 < 0$ H1h: $\beta_8 > 0$
Demography	Urban population growth (annual %) (x9)	World Bank, 2024	Mixed: can improve via efficiency (positive), worsen via demand growth (negative)	H0i: $\beta_9 = 0$ H1i: $\beta_9 \neq 0$
Economy	GDP per capita growth (annual %) (x10)		Mixed: higher income can finance clean tech (positive) but may raise consumption (negative)	H0j: $\beta_{10} = 0$ H1j: $\beta_{10} \neq 0$
Risk	Geopolitical Risk Index (GPR) (x11)	Geopolitical Risk Index, 2025	Mixed: rising geopolitical risk could lead to reduced energy consumption (positive), although spending on energy efficiency would also decrease, potentially resulting in higher emissions (negative)	H0k: $\beta_{11} = 0$ H1k: $\beta_{11} \neq 0$
Globalization	Globalization Index (x12)		Mixed: can enable clean tech transfer (positive) but also increase carbon footprint (negative)	H0l: $\beta_{12} = 0$ H1l: $\beta_{12} \neq 0$
Environment	Average CO ₂ emission factor (x13)	EnerData, 2024	Target variable	N/A

Source: developed by the authors.

Appendix B

Table B1

The descriptive statistics for normalized indicators

Indicator	Mean	SD	Median
x1	0.3423	0.2340	0.2964
x2	0.0936	0.1799	0.0328
x3	0.0830	0.1572	0.0438
x4	0.0534	0.1316	0.0166
x5	0.2954	0.2517	0.2010
x6	0.1405	0.1650	0.0851
x7	0.5313	0.1861	0.5489
x8	0.6806	0.0952	0.6808
x9	0.6415	0.0961	0.6456
x10	0.0342	0.1270	0.0036
x11	0.4029	0.1472	0.3902
x12	0.4230	0.2534	0.3868
x13	0.5202	0.1836	0.4973

Source: calculated by the authors.

Appendix C

Table C1

Clustering outcomes

Clusters	Countries
Cluster 0	Canada (2014–2023), New Zealand (2014–2023), Norway (2014–2023), Sweden (2014–2023)
Cluster 1	Australia (2014–2022), Belgium (2014–2021), Czech Republic (2014–2023), France (2014–2023), Germany (2014–2015), Japan (2014–2023), Malaysia (2016–2023), Netherlands (2014–2019), Poland (2014–2022), Singapore (2014–2023), South Korea (2014–2023), United Arab Emirates (2014–2023), United Kingdom (2014–2016), United States (2014–2023)
Cluster 2	Brazil (2014–2023), Colombia (2014–2023)
Cluster 3	China (2014–2023), Iran (2014–2023), Kazakhstan (2014–2023), Russia (2014–2023), Saudi Arabia (2014–2023), South Africa (2014–2023), Ukraine (2014–2018, 2022–2023), Uzbekistan (2014–2022)
Cluster 4	Algeria (2014–2023), Azerbaijan (2014–2023), Egypt (2014–2023), India (2014–2023), Indonesia (2014–2023), Malaysia (2014–2015), Mexico (2014–2019, 2021–2023), Nigeria (2014–2023), Ukraine (2019–2021), Uzbekistan (2023), Thailand (2014–2023), Turkey (2014–2023), Vietnam (2014–2020)
Cluster 5	Australia (2023), Belgium (2020, 2022–2023), Chile (2022–2023), Denmark (2014–2023), Germany (2019–2023), Italy (2023), Netherlands (2020–2023), Portugal (2014–2023), Spain (2014–2018, 2020–2023), United Kingdom (2017–2023)
Cluster 6	Argentina (2014–2023), Chile (2014–2021), Germany (2016–2018), Italy (2014–2022), Mexico (2020), Poland (2023), Romania (2014–2023), Spain (2019), Vietnam (2021–2023)

Source: developed by the authors.

Appendix D

The fragment of the script:

```

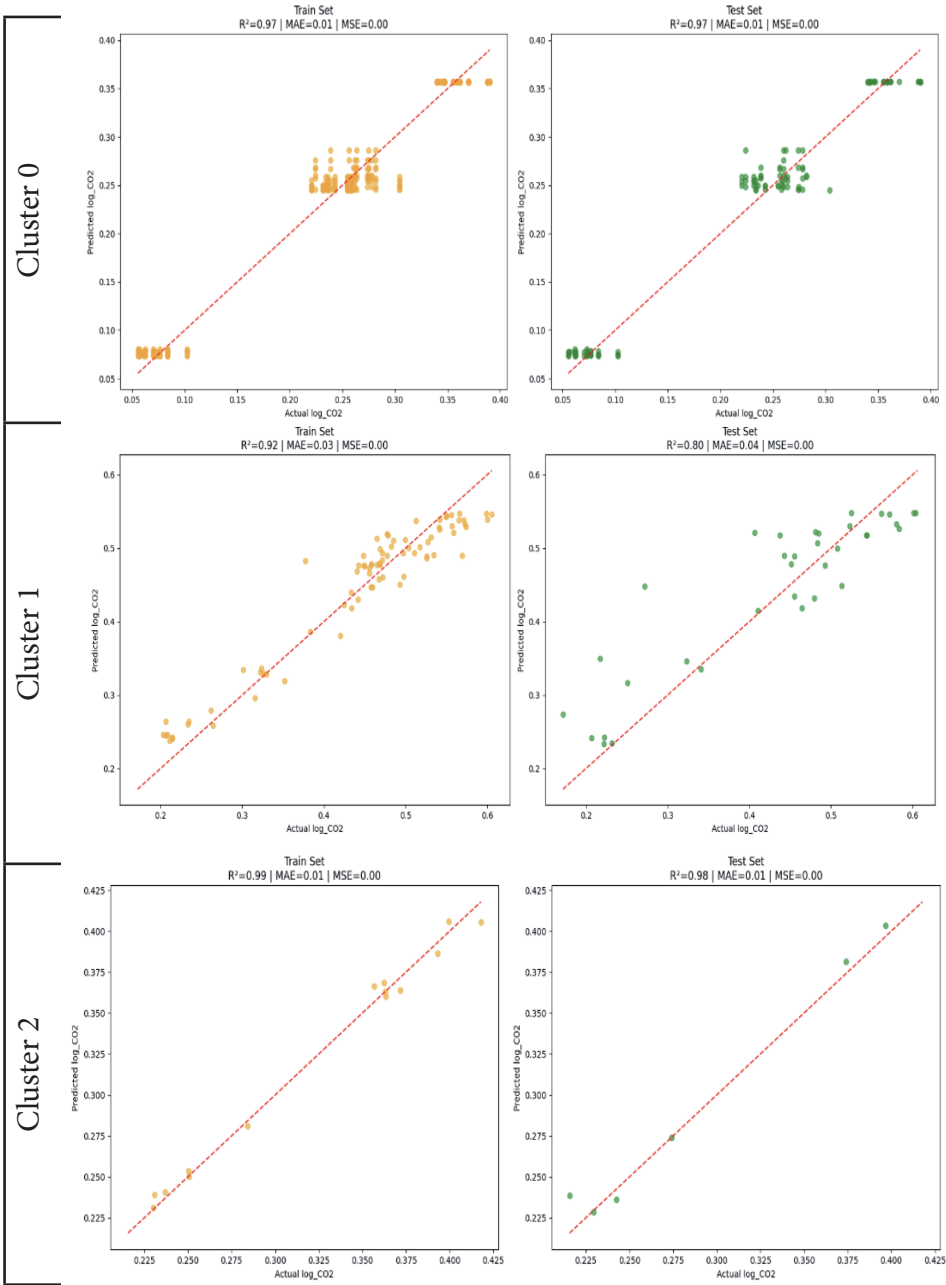
"# 3. Define search space for hidden layers (start from 12)
layer1 = 12 # start from 12 neurons
layer2_options = range(3, 10) # second layer: 3-9 neurons
layer3_options = range(2, 7) # third layer: 2-6 neurons
layer4_options = [1] # output layer
results = []
"#4. Train and evaluate MLPs for all layer combinations
for l2 in layer2_options:
    for l3 in layer3_options:
        for l4 in layer4_options:
            arch = (layer1, l2, l3, l4)
            mlp = Pipeline([
                ("scale", StandardScaler()),
                ("mlp", MLPRegressor(
                    hidden_layer_sizes=arch,
                    activation="logistic",
                    solver="adam",
                    max_iter=3000,
                    early_stopping=True,
                    n_iter_no_change=30,
                    random_state=42,
                    learning_rate="adaptive",
                    learning_rate_init=0.001
                ))
            ])
            mlp.fit(X_train, y_train)
            y_pred_test = mlp.predict(X_test)
            results.append({
                "Architecture": arch,
                "R2_Test": r2_score(y_test, y_pred_test),
                "MAE_Test": mean_absolute_error(y_test, y_pred_test),
                "MSE_Test": mean_squared_error(y_test, y_pred_test)
            })
# 5. Select the best architecture
results_df = pd.DataFrame(results)
best_idx = results_df["R2_Test"].idxmax()
best_model = results_df.iloc[best_idx]
best_arch = best_model["Architecture"]
print("Best Model")
print(best_model.round(4))
print(f"\n Optimal architecture (calculated): {best_arch}")

```

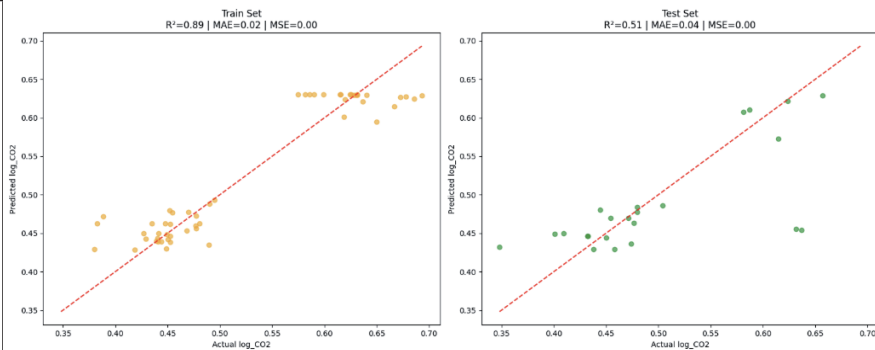
Appendix E

Figure E1

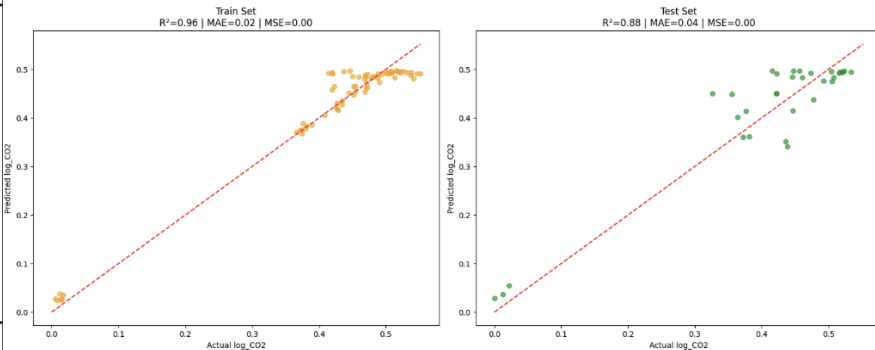
Estimation of forecast accuracy for 7 clusters based on MLP with optimal architecture for each cluster



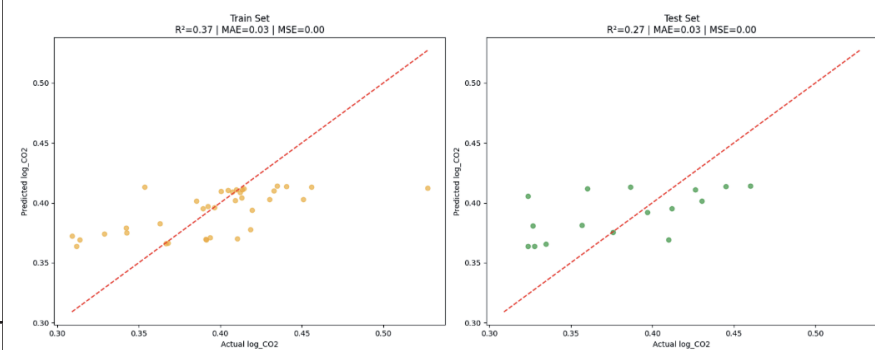
Cluster 3



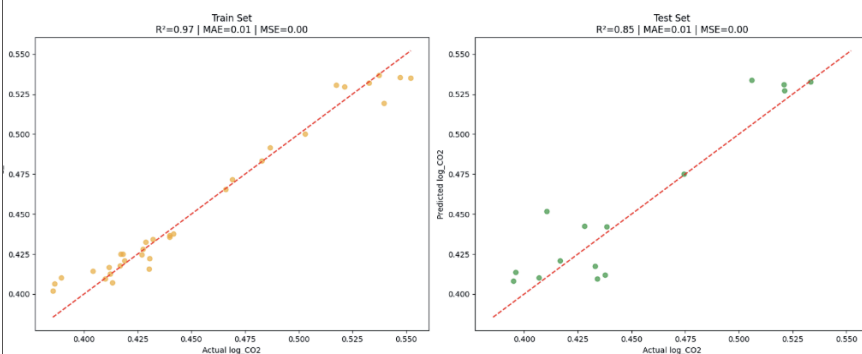
Cluster 4



Cluster 5



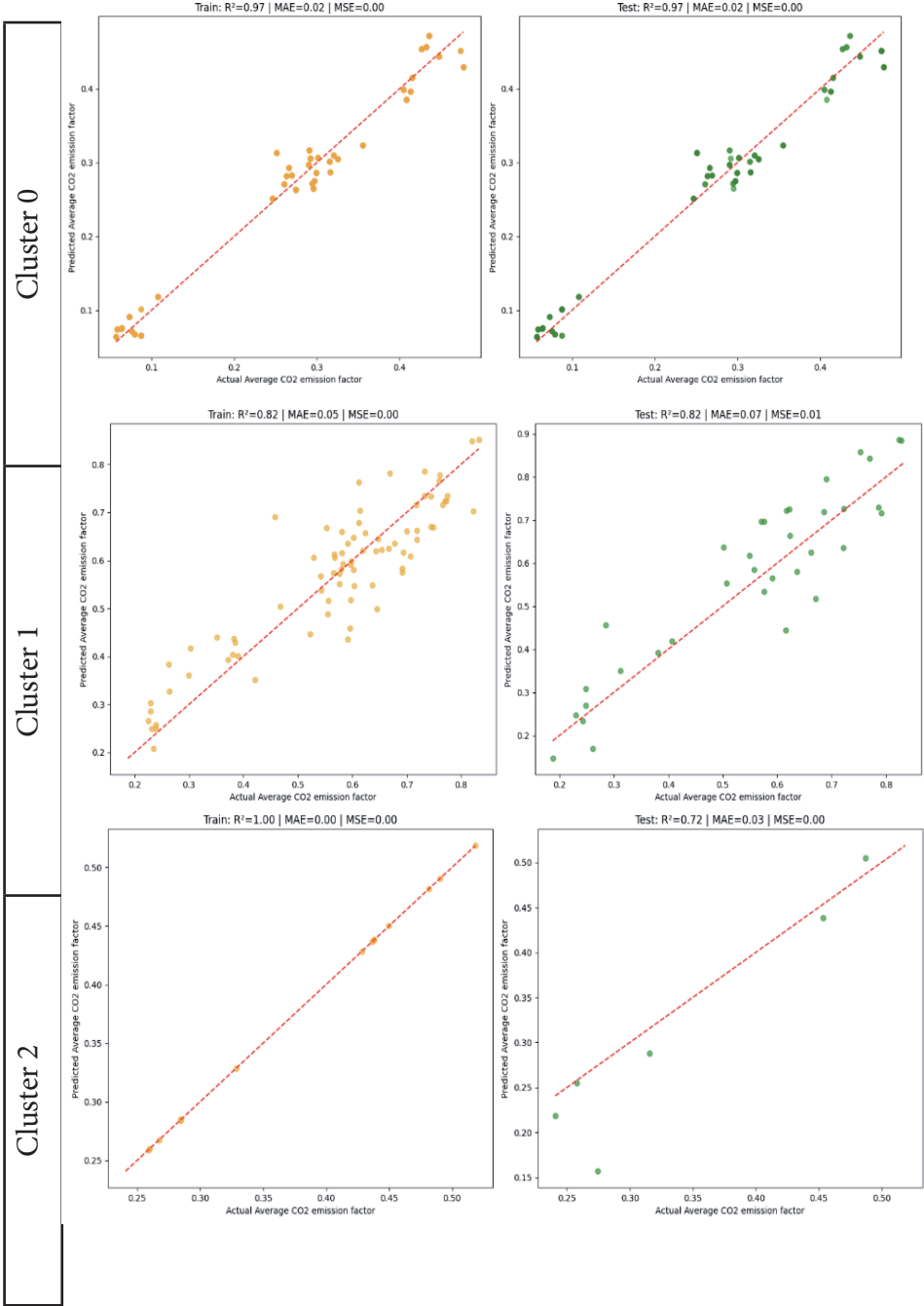
Cluster 6

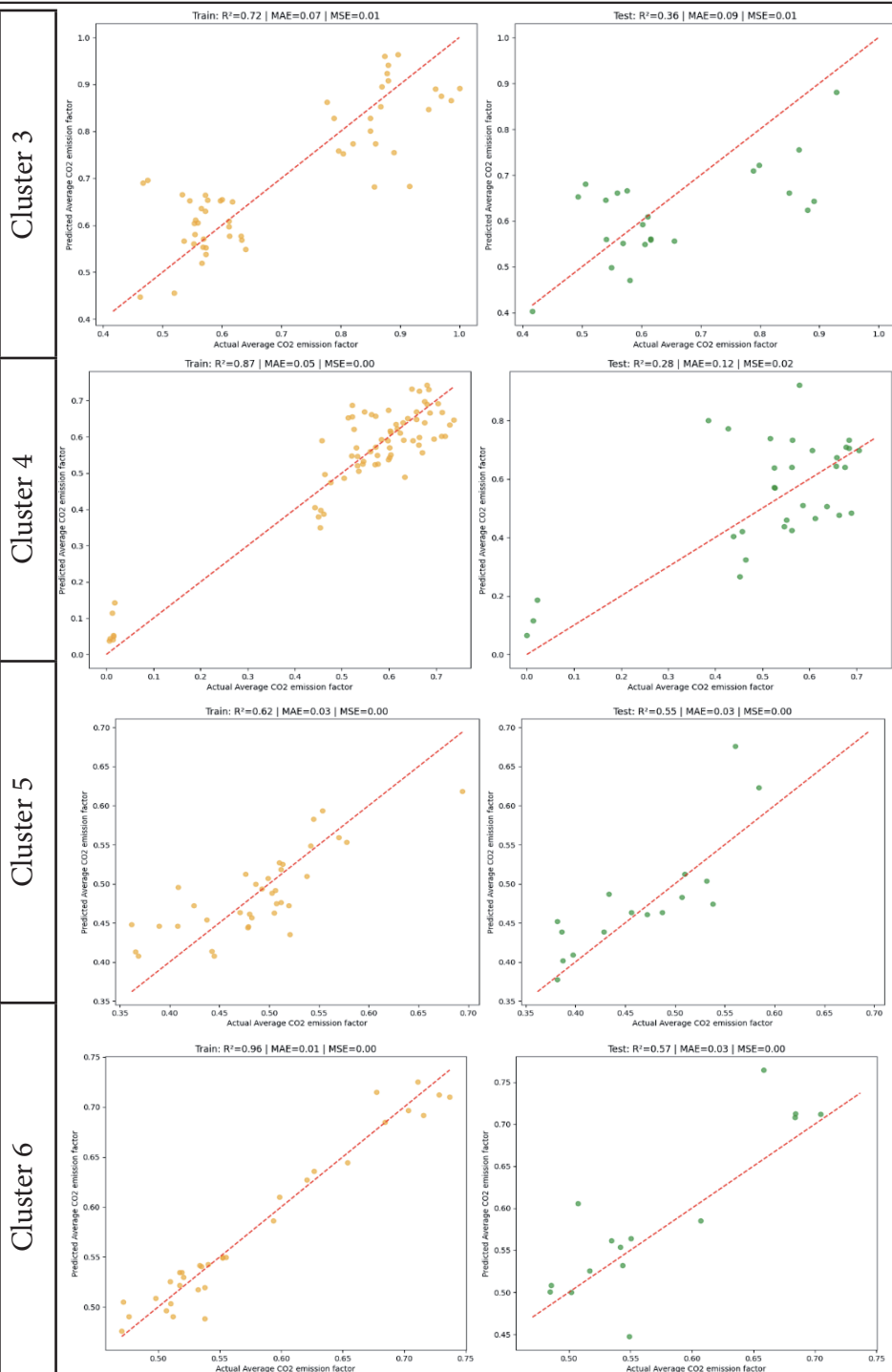


Source: developed by the authors.

Figure E2

Estimation of forecast accuracy for 7 clusters based on linear regression models





Source: developed by the authors.

Table E1

Performance Comparison of LR and MLP Models for 7 clusters

Model	Clusters							Average
LR	0	1	2	3	4	5	6	
R ² (Train)	0.97	0.82	1.00	0.72	0.87	0.62	0.96	0.85
MAE (Train)	0.02	0.05	1×10 ⁻²	0.07	0.05	0.03	0.01	0.03
MSE (Train)	1×10 ⁻²	1×10 ⁻²	1×10 ⁻²	0.01	1×10 ⁻²	1×10 ⁻²	1×10 ⁻²	1×10 ⁻²
R ² (Test)	0.97	0.82	0.72	0.36	0.28	0.55	0.57	0.61
MAE (Test)	0.02	0.07	0.03	0.09	0.12	0.03	0.03	0.06
MSE (Test)	1×10 ⁻²	0.01	1×10 ⁻²	0.01	0.02	1×10 ⁻²	1×10 ⁻²	0.01
MLP	0	1	2	3	4	5	6	
R ² (Train)	0.97	0.92	0.99	0.89	0.96	0.37	0.97	0.87
MAE (Train)	0.01	0.03	0.01	0.02	0.02	0.03	0.01	0.02
MSE (Train)	1×10 ⁻²	1×10 ⁻²	1×10 ⁻²	1×10 ⁻²	1×10 ⁻²	1×10 ⁻²	1×10 ⁻²	1×10 ⁻²
R ² (Test)	0.97	0.80	0.98	0.51	0.88	0.27	0.85	0.75
MAE (Test)	0.01	0.04	0.01	0.04	0.04	0.03	0.01	0.03
MSE (Test)	1×10 ⁻²	1×10 ⁻²	1×10 ⁻²	1×10 ⁻²	1×10 ⁻²	1×10 ⁻²	1×10 ⁻²	1×10 ⁻²

Source: calculated by the authors.

Appendix F

Table F1

Linear regression equations for 7 clusters

Cluster	Linear regression equations
Cluster 0	$y = 0.173 - 0.230 \cdot \text{Energy intensity of GDP at consumption} - 0.725 \cdot \text{Oil products domestic consumption} + 6.772 \cdot \text{Natural gas domestic consumption} - 15.248 \cdot \text{Electricity domestic consumption} + 0.378 \cdot \text{Share of renewables in electrification} - 0.382 \cdot \text{Share of wind and solar in electrification} + 0.160 \cdot \text{Urban population growth (annual \%)} - 0.043 \cdot \text{GDP per capita growth (annual \%)} + 43.494 \cdot \text{Coal and lignite domestic consumption} - 0.111 \cdot \text{Electrification} - 0.111 \cdot \text{Geopolitical Risk Index} + 0.007 \cdot \text{Globalization Index}$
Cluster 1	$y = -0.033 + 0.011 \cdot \text{Energy intensity of GDP at consumption} - 1.358 \cdot \text{Oil products domestic consumption} + 3.148 \cdot \text{Natural gas domestic consumption} - 5.318 \cdot \text{Electricity domestic consumption} - 1.502 \cdot \text{Share of renewables in electrification} + 1.331 \cdot \text{Share of wind and solar in electrification} + 0.171 \cdot \text{Urban population growth (annual \%)} - 0.168 \cdot \text{GDP per capita growth (annual \%)} + 7.994 \cdot \text{Coal and lignite domestic consumption} + 0.319 \cdot \text{Electrification} + 0.828 \cdot \text{Geopolitical Risk Index} + 1.000 \cdot \text{Globalization Index}$
Cluster 2	$y = 2.170 - 2.622 \cdot \text{Energy intensity of GDP at consumption} + 5.064 \cdot \text{Oil products domestic consumption} - 15.148 \cdot \text{Natural gas domestic consumption} - 4.122 \cdot \text{Electricity domestic consumption} - 0.575 \cdot \text{Share of renewables in electrification} + 0.622 \cdot \text{Share of wind and solar in electrification} - 0.922 \cdot \text{Urban population growth (annual \%)} - 0.073 \cdot \text{GDP per capita growth (annual \%)} + 50.395 \cdot \text{Coal and lignite domestic consumption} - 0.855 \cdot \text{Electrification} - 0.374 \cdot \text{Geopolitical Risk Index} + 0.377 \cdot \text{Globalization Index}$
Cluster 3	$y = -0.013 + 0.254 \cdot \text{Energy intensity of GDP at consumption} - 0.161 \cdot \text{Oil products domestic consumption} + 0.090 \cdot \text{Natural gas domestic consumption} - 1.340 \cdot \text{Electricity domestic consumption} + 0.127 \cdot \text{Share of renewables in electrification} + 1.290 \cdot \text{Share of wind and solar in electrification} + 0.418 \cdot \text{Urban population growth (annual \%)} + 0.134 \cdot \text{GDP per capita growth (annual \%)} + 1.147 \cdot \text{Coal and lignite domestic consumption} + 0.660 \cdot \text{Electrification} - 0.111 \cdot \text{Geopolitical Risk Index} - 0.065 \cdot \text{Globalization Index}$
Cluster 4	$y = -0.329 + 0.132 \cdot \text{Energy intensity of GDP at consumption} + 1.750 \cdot \text{Oil products domestic consumption} - 0.074 \cdot \text{Natural gas domestic consumption} - 20.538 \cdot \text{Electricity domestic consumption} - 0.903 \cdot \text{Share of renewables in electrification} + 2.914 \cdot \text{Share of wind and solar in electrification} + 0.741 \cdot \text{Urban population growth (annual \%)} - 0.254 \cdot \text{GDP per capita growth (annual \%)} + 9.946 \cdot \text{Coal and lignite domestic consumption} + 1.728 \cdot \text{Electrification} - 0.612 \cdot \text{Geopolitical Risk Index} + 0.751 \cdot \text{Globalization Index}$
Cluster 5	$y = 0.200 - 0.178 \cdot \text{Energy intensity of GDP at consumption} + 4.200 \cdot \text{Oil products domestic consumption} + 3.386 \cdot \text{Natural gas domestic consumption} - 17.846 \cdot \text{Electricity domestic consumption} - 0.546 \cdot \text{Share of renewables in electrification} + 0.303 \cdot \text{Share of wind and solar in electrification} + 0.054 \cdot \text{Urban population growth (annual \%)} - 0.051 \cdot \text{GDP per capita growth (annual \%)} + 10.619 \cdot \text{Coal and lignite domestic consumption} + 0.664 \cdot \text{Electrification} + 0.132 \cdot \text{Geopolitical Risk Index} + 0.329 \cdot \text{Globalization Index}$
Cluster 6	$y = 0.687 + 0.428 \cdot \text{Energy intensity of GDP at consumption} + 1.028 \cdot \text{Oil products domestic consumption} + 0.955 \cdot \text{Natural gas domestic consumption} - 6.813 \cdot \text{Electricity domestic consumption} - 0.234 \cdot \text{Share of renewables in electrification} - 0.199 \cdot \text{Share of wind and solar in electrification} - 0.070 \cdot \text{Urban population growth (annual \%)} - 0.004 \cdot \text{GDP per capita growth (annual \%)} + 4.097 \cdot \text{Coal and lignite domestic consumption} + 0.363 \cdot \text{Electrification} + 0.271 \cdot \text{Geopolitical Risk Index} - 0.473 \cdot \text{Globalization Index}$

Source: developed by the authors.

Appendix G

Table G1

Summarizing significant indicators per cluster and recommended modeling methods

Cluster	Significant indicators	Recommended modeling method
0	Energy intensity of GDP, Electricity consumption	Linear Regression (LR)
1	Electricity consumption, Coal and lignite consumption, Natural gas	Linear Regression (LR)
2	Share of renewables, Wind-solar generation	Nonlinear models (MLP)
3	Urban population growth, GDP per capita growth	Nonlinear models (MLP)
4	Electrification, Globalization index	Nonlinear models (MLP)
5	Geopolitical risk, GDP intensity	Linear Regression (LR)
6	Coal and lignite consumption, Electrification	Nonlinear models (MLP)

Source: developed by the authors.

Appendix H

Table H1

Testing hypotheses on datasets for each of 7 clusters

Indicator		x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
Expected sign (H1)		>0	>0	>0	any	<0	<0	any	>0	any	any	any	any
Cluster 0	β	-0.22	-0.61	6.74	-15.42	0.37	-0.37	0.16	-0.02	44.80	-0.09	-0.11	0.04
	p	1×10^{-4}	0.41	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	0.30	1×10^{-4}	0.12	1×10^{-4}	0.38
Cluster 1	β	0.08	-1.92	3.05	-4.10	-1.37	1.31	0.14	-0.07	8.48	0.37	0.74	0.82
	p	0.34	0.01	1×10^{-4}	0.01	1×10^{-4}	1×10^{-4}	0.17	0.53	1×10^{-4}	0.01	1×10^{-4}	1×10^{-4}
Cluster 2	β	-1.47	1.56	-1.05	-4.53	-0.20	0.33	-0.58	-0.07	34.82	-0.95	-0.10	0.12
	p	0.01	0.01	0.62	0.08	0.10	0.03	0.06	0.06	0.27	1×10^{-4}	0.51	0.20
Cluster 3	β	0.26	0.22	-0.07	-1.65	0.47	1.65	0.43	0.16	1.06	0.45	-0.15	-0.04
	p	0.17	0.74	0.81	0.01	0.37	0.01	0.00	0.20	0.08	0.02	0.37	0.80
Cluster 4	β	-0.06	0.36	1.94	-14.12	-0.55	1.43	0.54	-0.05	8.28	1.34	-0.37	0.44
	p	0.59	0.76	0.12	0.01	0.01	1×10^{-4}	0.05	0.73	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
Cluster 5	β	-0.05	1.39	3.18	-11.09	-0.58	0.34	-0.08	0.03	8.31	0.72	0.05	0.33
	p	0.82	0.69	0.03	0.23	1×10^{-4}	0.03	0.79	0.61	1×10^{-4}	0.05	0.71	0.10
Cluster 6	β	0.37	0.51	-0.82	0.19	-0.05	-0.11	0.16	0.03	1.23	-0.19	0.25	-0.27
	p	0.08	0.58	0.54	0.96	0.72	0.23	0.30	0.49	0.59	0.47	1×10^{-4}	0.01
Conclusions		Significant effect, H1 confirmed for cluster 6	Significant effect, H1 confirmed for cluster 2	Significant effect, H1 confirmed for clusters 0,1, 5	Significant effect, H1 confirmed for clusters 0-4,	Significant effect, H1 confirmed for clusters 1,2,4,5	Significant effect, H1 confirmed for cluster 0,	Significant effect, H1 confirmed for clusters 0,2,3,4	Significant effect, H1 confirmed for clusters 0,2	Significant effect, H1 confirmed for clusters 0,1,3,4,5	Significant effect, H1 confirmed for clusters 0,1,2,4,5	Significant effect, H1 confirmed for clusters 0,1,4,6	Significant effect, H1 confirmed for clusters 0,1,4,5,6,

Source: developed by the authors.

Received: April 5, 2026.
Reviewed: August 6, 2026.
Accepted: September 16, 2026