https://doi.org/10.15407/economyukr.2025.11.043
UDC 330.303.4:338.2
JEL: C45, C53, Q53, Q56

Olena ZHYTKEVYCH, PhD (Econ.), Doctoral candidate
of the Department of Artificial Intelligence, Modeling and Statistics
Kyiv National Economic University named after Vadym Hetman
54/1, Beresteysky Ave., Kyiv, 03680, Ukraine
e-mail: elena.zhitkevich@gmail.com
ORCID: https://orcid.org/0000-0003-2042-8795

REVIEW AND SELECTION OF CLUSTERING ALGORITHMS FOR DATASETS IN THE CONTEXT OF COUNTRIES' DECARBONIZATION

This paper explores the importance and applicability of clustering algorithms to high-dimensional datasets in the context of countries' decarbonization potential. The study concludes that Self-Organizing Maps with 3 and 5 clusters offers a balanced trade-off between computational efficiency and interpretability for a dataset used to determine the decarbonization potential of countries.

***Keywords**: countries' decarbonization potential; clustering algorithms; Self-Organizing Maps; clustering validation metrics.*

Determining the level of reducing CO_2 and other greenhouse gas emissions at national and international levels is critical for fulfilling global climate commitments. Countries-members of Paris Agreements submit regular reports on Nationally Determined Contributions based on objectively measurable emission reduction indicators. This is particularly relevant for Ukrainian economy due to heavy and nonmodern infrastructure and insufficient development of the monitoring, reporting and verification system during wartime (Kosse, 2023). In its turn, the lack of an effective monitoring and verification complicates transparent reporting and access to international climate

C i t a t i o n: Zhytkevych, O. (2025). Review and selection of clustering algorithms for datasets in the context of countries' decarbonization. *Economy of Ukraine. 68*. 11(768). 43-55. https://doi.org/10.15407/economyukr.2025.11.043

finance. Also, the European Union (EU), by implementing the Carbon Border Adjustment Mechanism is strengthening the requirements for the reliability of data on CO_2 emissions from exports [1]. The lack of such data may negatively affect Ukraine's export opportunities during the war and afterwar recovery period.

Currently, there is no single generally accepted model or methodology that would be fully adapted to Ukrainian conditions or any other country as "all fit model" and it could serve as a guideline for a comprehensive assessment of decarbonization (Banerji, 2021). Countries use different approaches for assessing their decarbonization level, however, none of them is universal, open and adapted to current Ukrainian conditions. Hence, Ukraine must either adapt existing international approaches or develop its own systems for effective decarbonization policy management or use them both as a hybrid approach. Otherwise, it risks being sluggish with the global climate processes.

The author analyzed the most recent papers assessing the decarbonization levels of different countries using methodologies and identified the strengths and weaknesses of each method, thereby selecting the most suitable method. Particularly, in the work of W. Zhou et al. (2023), a clustering of countries by CO_2 trends was carried out based on *K*-means clustering. The database consisted of 182 countries and a period of 1 year of 2023. The advantage of the applied method was determined as a rapid analysis of trends, but the disadvantage is that the number of clusters must be specified in advance. R. Novo et al. (2022) led global clustering of decarbonization scenarios based on time-series clustering. It was noted that the advantage of such a method is the ability to model transition scenarios, but the disadvantage is that the method is difficult to interpret economically. J. Inekwe et al. (2020) conducted a clustering of economic drivers of CO_2 for 72 countries (1990-2018) based on hierarchical and non-hierarchical clustering methods. Thus, the advantage of such an approach was determined as the ability of the method to identify impact groups and the disadvantage was a high sensitivity to data scaling. Z. Li et al. (2020) performed clustering of decarbonization policies, preparing a base of scenarios of 2030 and 2050 globally. The main advantage of the method is working with uncertainty, and the disadvantage of such an approach is defined as the dependence of the result on the scenarios. Y. Hu & L. Weng (2024) examined the primary driving factors of carbon emission from 1990 to 2020 across the 10 countries using the Logarithmic Mean Divisia Index (LMDI) model and the Autoregressive Integrated Moving Average approach was applied to identify the evolutionary trajectories of carbon emission trends. This approach has the advantage of integrating clustering and forecasting, but also the complexity of implementing multi-level models. Another approach based on IPAT/Kaya approach combined with Variance analysis technique for ASEAN countries (1971-2013) has been applied by J. Chontanawat (2018) to conduct a factor analysis of energy CO_2. The advantage of this

[1] Carbon Border Adjustment Mechanism (CBAM). *European Commission*. 2021. URL: https://op.europa.eu/en/publication-detail/-/publication/3c0285d2-545e-11ec-91ac-01aa75ed71a1/language-en (accessed: 06.07.2025).

method is its simplicity in calculations and explanation, but the disadvantage is its inability to provide a comparative classification. M. Anser et al. (2024) analysed energy consumption, technological innovation, and economic growth in BRICS countries (1990-2023), based on Gaussian Mixture Models (GMM) panel VAR framework analysis. The advantage of the method is the elimination of endogeneity, and the disadvantage is the poor choice of instruments or lags can lead to biased results or weak identification.

Hence, most studies integrate economic, environmental, energy and social indicators to assess decarbonization, which allows for a deeper understanding of the structure of processes. However, there is a lack of or weak validation of clustering results. Only some studies use validation metrics, and other do not assess the reliability of clusters, this reduces the reliability of the results and limits their comparability. It should also be noted that there is a difficulty in interpreting the results with high data variability. Additionally, many studies cover limited time periods or regions which makes it difficult to generalize the results to the world or Ukrainian context.

Consequently, the **purpose of this article** is to provide overview and select clustering method applicable to the prepared dataset in the context of low carbon strategy of nations.

OVERVIEWING CLUSTERING METHODS

Clustering is widely used in various fields, including bioinformatics, finance, ecology, and machine learning. It is one of the most common data analysis methods used to automatically group objects into clusters (groups) that have internal similarities and at the same time significantly differ from objects in other clusters (Jain, 2010). The effectiveness of clustering depends on the chosen method, datasets characteristics and the quality of the assessment of the resulting clusters. Also, it is vital to know how to select appropriate clustering method and number clusters before conducting clustering in order to achieve the effective and efficient results.

The main clustering methods, approaches to assessing the quality of clusters and the most popular metrics that help choose the optimal cluster distribution are presented in Table 1 and Table 2.

Considering the main advantages of *K*-means method it is important to mention its simplicity, fastness, while its sensitivity to the choice of number of clusters and inability to detect noise make this method not perfect for utilising in decarbonization identification of nations. The HC does not require knowing the number of clusters in advance and provide understandable dendrogram, while it is computationally expensive and difficult to apply for large datasets. The DBSCAN detects clusters of any shape, ignores noise, while requires parameter selection and cause issues with clusters of different densities. The GMM takes into account probabilities, is vulnerable to local minima, while is sensitive to parameters. Hence, these methods also cannot be considered as efficient and effective in the context of decarbonization potential identification.

While SOM provides visualization of high dimensional data and topological structure of obtained groups. Additionally, SOM exhibits linear complexity with respect to the size of the database, making it suitable for analyzing large scale datasets. However, it requires parameter regulation and is associated with slower operation. Although it can provide quite effective results if the clusters are chosen correctly and the results are interpreted.

Accordingly, clustering methods are powerful tools for analyzing complex multidimensional data. The choice of the method and its parameters should be based on the characteristics of a particular dataset, the purpose of the study, and the requirements for interpreting the results. At the same time, to achieve efficiency and effectiveness, it is important to assess the quality of clusters using appropriate metrics (evaluation methods) to select the most adequate data distribution. That's what was explained as the next stage of the paper.

Evaluation methods to be applied for clustering algorithms are divided into internal, external and relative (Datta, Datta, 2003) and their descriptions, formulas and acceptable range of values are presented in a Table 2.

Internal metrics assess the quality of clustering based on the properties of the data structure itself and the resulting clusters without using additional information (reference labels). This is important when "correct" clusters are unknown in advance. They help determine how compact the clusters are, how separate they are, and how well the objects fit their cluster (Manning et al., 2023).

External metrics are used when reference data or labels are available, for example, countries can be pre-classified according to political, geographical or economic

Table 1. **Basic clustering methods**

Name	Description
Hierarchical clustering (HC)	It builds a hierarchy of clusters by painful merging (agglomerative) or splitting (divisive) (Murtagh, Contreras, 2012). The result is a dendrogram — a tree of clusters, which allows you to reduce the optimal level of aggregation
K-means	It is one of the simplest and most popular, it minimizes the sum of the squares of the distances of objects from the center of their cluster (Lloyd, 1982). It requires to specify the number of clusters in advance
DBSCAN	It identifies clusters as a region with a high density of points in space are able to detect irregularly shaped clusters and separate noise (Ester et al., 1996)
Gaussian Mixture Models (GMM)	It considers the probability distribution of the data and considers each cluster to describe a certain distribution (McLachlan, Peel, 2000)
Kohonen Self Organizing Maps (SOM)	It is a neural network method that transforms multidimensional data into a two-dimensional mesh while preserving topological properties, and effectively visualizes and clusters complex structures (Kohonen, 2012)

Source: compiled by the author.

Table 2. **Internal and external metrics for evaluating the quality of clustering**

Name	Description	Formula	Range
		Internal metrics	
Silhouette Score (*SS*)	It measures the degree of proximity of an object to its cluster compared to other clusters. Values range from –1 to 1, where 1 is a perfect cluster (Rousseeuw, 1987)	$SS = \frac{b(i) - a(i)}{\max\{a(i),\ b(i)\}}$, (1) where *a(i)* is the average distance between point *i* and all other points in the same cluster; *b(i)* is the minimum average distance between point *i* and points in any other cluster (Manning et al., 2023)	[–1, 1], where higher values indicate better clustering result
Davis-Bouldin Index (*DBI*)	It compares the distances between clusters and the dispersion within them; a lower value means better clustering (Davies, Bouldin, 1979)	$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{Si + Sj}{Mij}\right)$, (2) where *Si* is the average intra-cluster distance for cluster *i* and *Mij* is the distance between the centroids of clusters *i* and *j* (Xu, Wunsch, 2022)	[0, ∞], where lower values indicate better clustering result
Calinski-Harabasz Index (*CH*)	It determines the ratio between the between-cluster variance and the within-cluster variance; the higher the better (Calinski, Harabasz, 1974)	$CHI = \left(\frac{Tr(B_k)/k - 1}{Tr(W_k)/n - 1}\right)$, (3) where $Tr(B_k)$ is the trace of the between-cluster dispersion matrix; $Tr(W_k)$ is the trace of the within-cluster dispersion matrix; *n* is the number of samples; *k* is the number of clusters (Halkidi et al., 2021)	[0, ∞], where higher values indicate better clustering result
		External metrics	
Adjusted Rand Index (*ARI*)	It compares the partition obtained by clustering with a reference partition (Hubert, Arabie, 1985)	$ARI = \frac{(RI - \text{Expected}(RI))}{(\max(RI) - \text{Expected}(RI))}$, (4) where *RI* is the Rand Index; Expected *(RI)* is the expected value of *RI* for random labelling (Zhou et al., 2024)	[–1, 1], where 1 indicates the best clustering result
Normalized Mutual Information (*NMI*)	It estimates the degree of similarity of two clustered groups based on information theory (Strehl, Ghosh, 2002)	$NMI(U,\ V) = \frac{2 \times I(U,\ V)}{H(U)\ + H(V)}$, (5) where *I(U, V)* is the mutual information between clusters *U* and *V*; *H(U)* and *H(V)* are the entropies of *U* and *V* (Strehl, Ghosh, 2002)	[0, 1], where 1 indicates the best clustering result

Source: compiled by the author.

characteristics, which serve to check the adequacy of the clustering. They help to assess how well the clustering distribution is consistent with existing knowledge or categories (Strehl, Ghosh, 2002).

Therefore, in this stage it is important to know what type of clustering (supervised or unsupervised) is being performed to determine which metrics can be applied.

DESCRIBING THE CLUSTER ANALYSIS PROCESS USING COUNTRY DECARBONIZATION DATA

The following stage is necessary to describe how to conduct the clustering and present the phases of clustering analysis, which consists of data preparation, choosing a clustering method, determining the number of clusters, performing clustering and interpreting and validating the results.

So, the first step is data collection, cleaning and normalization. For most clustering methods, normalization (scaling) is mandatory, since different scales of features can distort distances (Jain, 2010).

Our predefined dataset covers 40 countries, these countries have been selected from the OECD (Turkey, South Korea, Japan, etc.), G7 (United States, Canada, Japan, Germany, France, United Kingdom, Italy), BRICS, European Union, CIS (Azerbaijan, Kazakhstan, Uzbekistan, Ukraine, etc.), North America, Latin America (Mexico, Brazil, Chile), Asia (Thailand, Vietnam), Australasia (Australia, New Zealand), Africa (Algeria, Egypt) and the Middle East (Saudi Arabia, United Arab Emirates). In this study, the choice of countries is primarily due to the availability of reliable and comparable data in the open access and not intended to demonstrate selectivity towards individual alliances or regional groupings. Indeed, the list of countries was formed on the basis of the principle of ensuring representativeness and relevance of information for further analytical conclusions. Regarding the positioning of Ukraine, the author recognizes its historical belonging to the post Soviet space, however, in the context of cluster analysis, this characteristic is not decisive. The study focuses on the possibility of clustering Ukraine into a certain group of countries according to objective criteria of decarbonization potential, which allows avoiding a simplified interpretation of regional affiliation. This approach is aimed at ensuring that the sample serves as the basis for obtaining practical results.

The period of analysis covers 10 years and collect the data per indicator on an annual basis, starting from 2014 to 2023. The total number of indicators is 14, which were selected based on the author's expertise, correlation analysis and socioeconomic factors which have potential impact on decarbonization possibilities of a nation.

The composition of the indicators have been selected from two reports (EnerData [2] and The World Bank [3]) and validated in our previous studies (Zhytkevych, Brochado, 2022; Zhytkevych et al., 2023; Matviychuk et al., 2024): Total energy con-

[2] World energy & climate statistics — Yearbook 2023. *EnerData*. 2023. URL: https://yearbook.enerdata.net/total-energy/world-consumption-statistics.html (accessed: 06.07.2025).

[3] World Development Indicators. *The World Bank*. 2022. URL: https://datacatalog.worldbank.org/search/dataset/0037712 (accessed: 06.07.2025).

sumption, Energy intensity of GDP, Coal and lignite trade, Coal and lignite consumption, Oil products balance of trade, Oil products consumption, Natural gas balance of trade, Natural gas domestic consumption, Electricity balance of trade, Electricity domestic consumption, Share of renewables in electricity, Average CO_2 emission factor, GDP per capita growth (annual %), Urban population (% of total population).

It is important to clarify some indicators like the energy consumption is the balance of primary production, external trade, marine bunkers and stock changes; the total energy consumption includes biomass. For the world, marine bunkers are included which induces a gap with the sum of regions. The average CO_2 emission factor (carbon factor) is calculated doing the ratio between emissions over primary energy consumption. Crude oil includes all liquid hydrocarbons to be refined crude oil, liquids from natural gas and semi-refined products. The share of renewables in electricity production is a ratio between the electricity production from renewables (hydro, wind, geothermal and solar) and the total electricity production [4].

The second phase of the clustering analysis is selecting the clustering method. As we have mentioned that there is no single optimal clustering method. While, the comparative approach tailored to the nature of the data set can provide the most robust option. Hence, the choice of the clustering method depends on the size and structure of the data, the purpose of the study, etc. (for example, for noisy data DBSCAN is can be used, for visualization is SOM recommended). We used SOM method for identification decarbonization potential of countries in our studies.

The third phase is determining the number of clusters. Since in our study we perform unsupervised clustering and do not know "correct" clusters or labels then external metrics cannot be applied. Most methods, except of Hierarchical and Density based ones, require specifying the number of clusters, so it can be done using analysis of internal metrics (see Table 2).

So, we calculated the SC, DBI and CHI indices using formulas (1—3) for selected most popular algorithms (*K*-means, GMM, and SOM) with different number of clusters (k = 3, ..., 8). Their values were computed using Scikit-learn library and visualized as heatmaps in MS Power BI environment (Fig. 1). Scikit-learn is a machine learning library for Python that offers a complete set of algorithms, especially clustering [5]. However, we can see that the presented values have different dimensions (Fig. 1), for example 203.73, 1.27, 0.12, etc., so it was useful to average them for the 3 algorithms and normalize them for better comparative accuracy. Therefore, we standardised *SC*, *DBI* and *CHI* values and taking into account that only *DBI* should as lower as better, we computed average value (*AVR*) of the indices:

$$AVR = \frac{(SC - DBI + CHI)}{3}. \quad (4)$$

[4] World energy & climate statistics — Yearbook 2023. *EnerData*, 2023. URL: https://yearbook.enerdata.net/total-energy/world-consumption-statistics.html (accessed: 06.07.2025).

[5] Clustering performance evaluation. *Scikit-learn*. 2024. URL: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation (accessed: 06.07.2025).

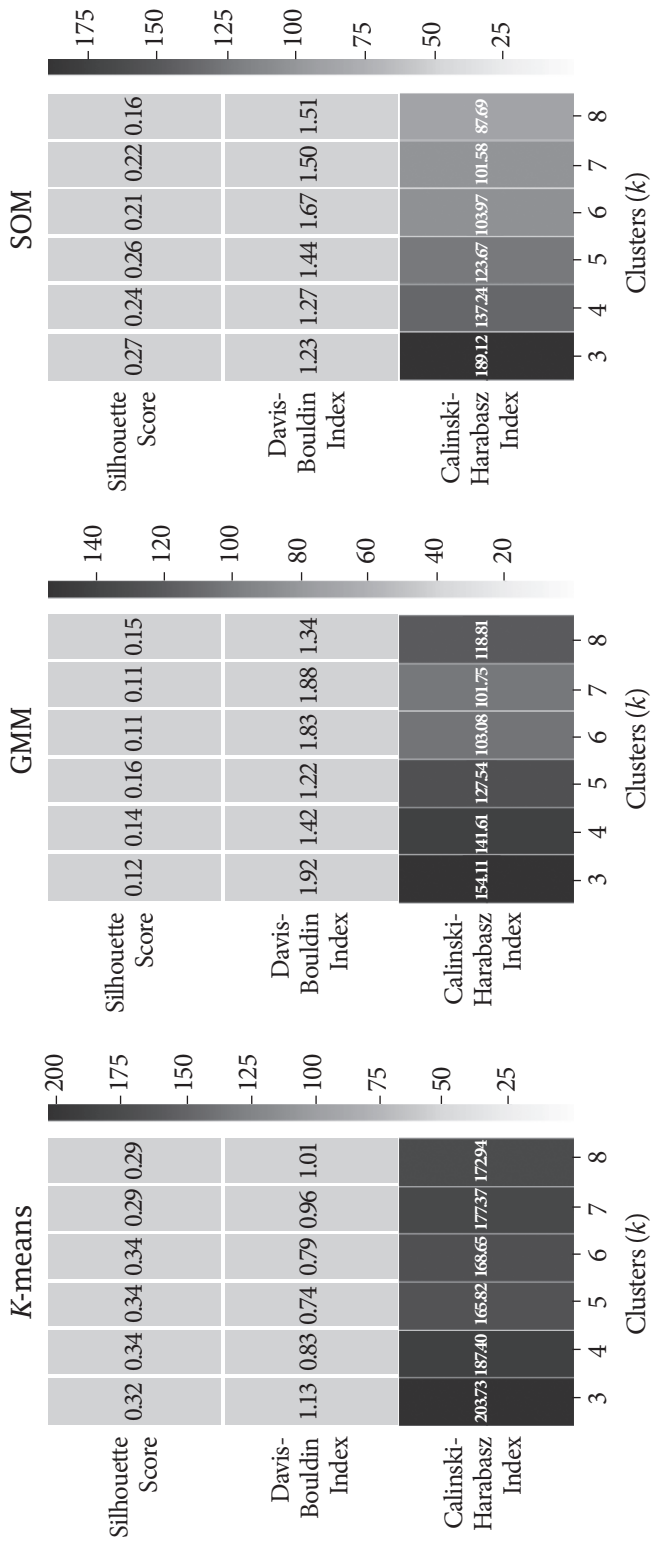

Fig. 1. The comparison of indexes for 3 algorithms with different number of clusters
Source: created by the author.

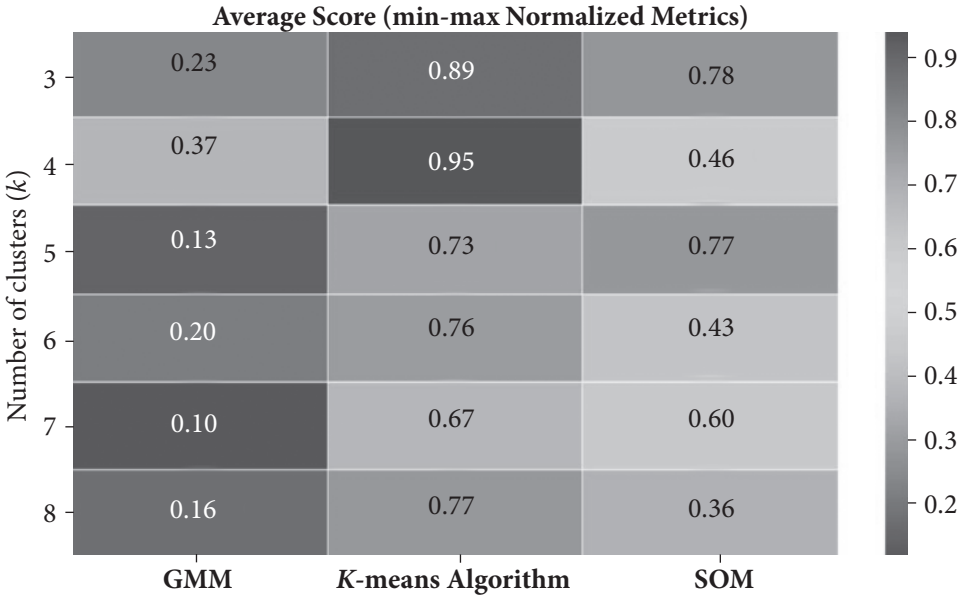

Fig. 2. The heatmap of *AVR* values for *K*-means, GMM, SOM algorithms with 3, 4, 5, 6, 7, and 8 clusters
Source: created by the author.

The normalised (min-max) results of *AVR* computing for 3 algorithms using formula (4) and different number of clusters are visualised in Power BI environment and presented in Fig. 2.

The highest *AVR* value indicates the best choice of the number of clusters for the different algorithms suitable for our dataset. Thus, we see that *K*-means with 4 and 3 clusters (*AVR* = 0.95, *AVR* = 0.89 respectively) determines the best result. For SOM $k = 3$ and $k = 5$ (*AVR* = 0.78, *AVR* = 0.77 respectively) also indicate good choice. GMM shows very poor results compared to *K*-means and SOM for the entire range of clusters, where AVR varies in the range from 0.37 to 0.1 (see Fig. 2).

Since SOM has more advantages compared to *K*-means, because it offers better possibilities for visualization of clustering results, so can be interpreted analytically, we confirmed our choice of clustering algorithm (SOM with 5 or 3 clusters) by calculating and comparing *AVR* (based on conducted metrics analysis).

The fourth phase is the application of the selected algorithm to the data with the parameters obtained in the previous stages. The fifth phase is visualization and validating of the results. These two phases with graphical representation of clustering results (SOM with 3 and 5 clusters) of countries according to their decarbonization potential are presented and described in detail in our works (Zhytkevych et al., 2023; Matviychuk et al., 2024).

For external validation in the presence of labels, ARI, NMI can be used and we will implement this phase in our future works.

In addition, monitoring and updating are important in the clustering process, so regular data updates and re-analysis to track changes can be indicated as a necessary clustering step.

Thus, the conducted analysis of clustering methods and metrics allowed us to confirm that SOM with 5 or 3 clusters is a suitable method for our pretrained dataset in the context of determining the decarbonization potential of countries, in particular Ukraine.

CONCLUSIONS

Therefore, our literature review outcomes determined that clustering can be used to identify the decarbonization potential of countries. Countries that are located in the same cluster, and therefore at similar stages of development, can create a mechanism for cooperation on building a clean energy base, distributing energy storage and developing decarbonization policies. This data can help policymakers better assess, formulate and research decarbonization strategies, as well as take into account historical (typical) common features of carbon emission patterns.

The authors also analyzed and identified the advantages and disadvantages of clustering algorithms, including Hierarchical clustering, *K*-means, DBSCAN, GMM and Kohonen Self Organizing Maps. This study highlights the practical value of applying clustering algorithms such as GMM, SOM and *K*-means to multidimensional socio-economic and environmental data.

The results demonstrate that SOM is well suitable for Big Data analysis due to their linear computational complexity and ability to preserve topological relationships, particularly when the number of iterations and map size are fixed.

A clustering quality score was also provided to compare the effectiveness of the three clustering algorithms. Thus, the application of the clustering methods (GMM, SOM and *K*-means) was tested on a database for assessing the decarbonization potential of various countries, including Ukraine. Hence, SOM showed its efficiency and interpretability, making it a reliable tool for exploratory data analysis and country profiling in the context of determining the decarbonization potential.

However, it is important to mitigate one of the drawbacks of applying SOM by pre-determining the number of clusters in advance. This can be achieved by applying evaluation metrics, in particular formulas (1—3) and determining the exact number of clusters for effective and efficient application. The SOM with 3 and 5 cluster were identified as a balanced trade-off between computational efficiency and interpretability of the dataset used.

In contrast, algorithms such as the GMM have demonstrated the ability to consider each cluster to describe a specific distribution and DBSCAN is able to detect irregularly shaped clusters and separate noise. However, GMM is sensitive to parameters, while DBSCAN requires parameter selection and presents problems with clusters of different densities. This suggests that additional preprocessing is required before applying them to real datasets, making these algorithms inefficient and impractical for our dataset.

Concluding, the clustering in general allows us to identify groups of countries with similar decarbonization characteristics, which can be used for the iden-

tification of appropriate benchmarks for Ukraine. The results obtained can be used to adapt international best practices to the Ukrainian energy infrastructure, which has undergone and is undergoing significant transformations as a result of the war. This provides an analytical basis for the formation of national targeted decarbonization policy.

Overall, the results of the study suggest that no single clustering method is universally optimal. Rather, a comparative approach tailored to the nature of the dataset may provide the most robust conclusions. Future research should focus on combining clustering with supervised learning methods to enhance policy relevance and predictive power.

REFERENCES

Kosse, I. (2023). Rebuilding Ukraine's infrastructure after the war. *Policy Notes and Reports 72*. 24 p. URL: https://wiiw.ac.at/rebuilding-ukraine-s-infrastructure-after-the-war-dlp-6621.pdf

Zhou, W., Zhou, J., Hu, G. (2023). Research of varying patterns of CO_2 emissions in 182 countries based on K-means method. *Applied and Computational Engineering.* No. 6(1). P. 1597—1606. https://doi.org/10.54254/2755-2721/6/20230480

Novo, R., Marocco, P., Giorgi, G., Lanzini, A., Santarelli, M., Mattiazzo, G. (2022). Planning the decarbonisation of energy systems: The importance of applying time series clustering to long-term models. *Energy Conversion and Management: X*. Vol. 15. 100274. https://doi.org/10.1016/j.ecmx.2022.100274

Inekwe, J., Valadkhani, A., Smyth, R. (2020). Drivers of carbon dioxide emissions: An empirical investigation using hierarchical and non-hierarchical clustering methods. *Environmental and Ecological Statistics.* No. 20(4). P. 1—40. https://doi.org/10.1007/s10651-019-00433-4

Li, Z., Wang, C., Li, Y. (2020). Using clustering algorithms to characterise uncertain long-term decarbonisation pathways. *Applied Energy.* Vol. 268. 114947. https://doi.org/10.1016/j.apenergy.2020.114947

Hu, Y., Weng, L. (2024). Net-zero energy transition in ASEAN countries: The evolutionary model brings novel perspectives to the cooperative mechanism of climate governance. *Journal of Environmental Management.* Vol. 351. 119999. https://doi.org/10.1016/j.jenvman.2023.119999

Chontanawat, J. (2018). Decomposition analysis of CO_2 emissions in ASEAN: An extended IPAT model. *Energy Procedia.* Vol. 153. P. 186—190. https://doi.org/10.1016/j.egypro.2018.10.057

Anser, M., Ali, S., Umair, M., Javid, R., Mirzaliev, S. (2024). Energy consumption, technological innovation, and economic growth in BRICS: A GMM panel VAR framework analysis. *Energy Strategy Reviews.* Vol. 56. 101587. https://doi.org/10.1016/j.esr.2024.101587

Jain, A. (2010). Data clustering: 50 years beyond *K*-means. *Pattern Recognition Letters*. No. 31. Iss. 8. P. 651—666. https://doi.org/10.1016/j.patrec.2009.09.011

Murtagh, F., Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *WIREs Data Mining and Knowledge Discovery.* Vol. 2. Iss. 1. P. 86—97. https://doi.org/10.1002/widm.53

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory.* Vol. 28. Iss. 2. P. 129—137. https://doi.org/10.1109/TIT.1982.1056489

Ester, M., Kriegel, H., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press. URL: https://cdn.aaai.org/KDD/1996/KDD96-037.pdf

McLachlan, G., Peel, D. (2000). Finite mixture models. John Wiley & Sons. 427 p. https://doi.org/10.1002/0471721182

Kohonen, T. (2012). Self-organizing maps. Berlin, Springer. 502 p. https://doi.org/10.1007/978-3-642-56927-2

Datta, S., Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*. Vol. 19. Iss. 4. P. 459—466. https://doi.org/10.1093/bioinformatics/btg025

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. Vol. 20. P. 53—65. https://doi.org/10.1016/0377-0427(87)90125-7

Manning, C., Raghavan, P., Schьtze, H. (2023). Introduction to Information Retrieval. Cambridge, Cambridge University Press. 542 p. URL: https://nlp.stanford.edu/IR-book/

Davies, D., Bouldin, D. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 1. Iss. 2. P. 224—227. https://doi.org/10.1109/TPAMI.1979.4766909

Xu, R., Wunsch, D. (2008). Clustering. Wiley IEEE Press. 368 p. URL: https://www.wiley.com/en-us/Clustering-p-9780470276808

Calicski, T., Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics: Theory and Methods*. Vol. 3. Iss. 1. P. 1—27. https://doi.org/10.1080/03610927408827101

Halkidi, M., Batistakis, Y., Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*. Vol. 17. P. 107—145. https://doi.org/10.1023/A:1012801612483

Hubert, L., Arabie, P. (1985). Comparing partitions. *Journal of Classification*. Vol. 2. P. 193—218. https://doi.org/10.1007/BF01908075

Zhou, S., Xu, H., Zheng, Z., Chen, J., Li, Z., Bu, J., Wu, J., Wang, X., Zhu, W., Ester, M. (2024). A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions. *ACM Computing Surveys*. Vol. 57. Iss. 3. P. 1—38. https://doi.org/10.1145/3689036

Strehl, A., Ghosh, J. (2002). Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*. Vol. 3. P. 583—617. https://dl.acm.org/doi/10.1162/153244303321897735

Received on July 14, 2025
Reviewed on August 18, 2025
Revised on August 19, 2025
Signed for printing on August 25, 2025

Олена Житкевич, канд. екон. наук,
докторантка кафедри штучного інтелекту, моделювання та статистики
Київський національний економічний університет імені Вадима Гетьмана
Берестейський просп., 54/1, 03680, Київ, Україна

ОГЛЯД ТА ВИБІР АЛГОРИТМІВ КЛАСТЕРИЗАЦІЇ ДЛЯ НАБОРІВ ДАНИХ У КОНТЕКСТІ ДЕКАРБОНІЗАЦІЇ КРАЇН

Проблема оцінки та моделювання потенціалу декарбонізації країн має вирішальне значення для забезпечення сталого економічного і соціального розвитку на мікро- та макрорівнях. Зважаючи на зростання актуальності великих даних у дослідженні декарбонізації та інтеграцію алгоритмів кластеризації, украй важливо визначити методи кластеризації, які б були масштабованими, надійними й придатними для використовуваного набору даних. Проведений аналіз відповідної літератури свідчить про те, що немає єдиного оптимального методу кластеризації, тому порівняльний підхід, адаптований до характеру набору даних, може забезпечити найкращі результати. Розглянуто широко використовувані методології кластеризації, застосовані до підготовлених наборів даних, зокрема в контексті декарбонізації. Виконано оцінювання якості кластеризації за допомогою її внутрішніх метрик. Дослідження було проведено на попередньо визначеному наборі даних з 14 нормалізованих ключових показників для встановлення потенціалу декарбонізації 41 країни протягом 10-річного періоду. Застосування трьох методів кластеризації (*K*-середніх, GMM і самоорганізованих карт) було протестовано на базі даних для оцінювання потенціалу декарбонізації різних країн, у тому числі України, і сформульовано важливі висновки, зокрема, що самоорганізована карта з трьома і п'ятьма кластерами є найбільш придатною кластеризацією для набору даних, що використовується для визначення потенціалу декарбонізації країн, у тому числі й України. Отримані результати кластеризації можуть бути використані для адаптації передового міжнародного досвіду до української енергетичної інфраструктури, яка зазнала і зазнає значних трансформацій унаслідок війни.

***Ключові слова**: потенціал декарбонізації країн; алгоритми кластеризації; самоорганізовані карти; метрики валідації кластеризації.*

Надійшла 14.07.2025
Прорецензована 18.08.2025
Доопрацьована 19.08.2025
Підписана до друку 25.08.2025