DOI 10.33111/nfmte.2023.067

COMPARATIVE ANALISYS OF THE EFFECTIVENESS OF DIMENSIONALITY REDUCTION ALGORITHMS AND CLUSTERING METHODS ON THE PROBLEM OF MODELLING ECONOMIC GROWTH

Sergii Poznyak

Kyiv National Economic University named after Vadym Hetman
54/1 Beresteysky Ave., Kyiv, 03680, Ukraine
ORCID: 0009-0006-4894-7983, E-mail: poznyak.sergiy.w@gmail.com

Yurii Kolyada

Kyiv National Economic University named after Vadym Hetman
54/1 Beresteysky Ave., Kyiv, 03680, Ukraine
ORCID: 0000-0003-2516-9817, E-mail: koliada_yurii@kneu.edu.ua

This article is devoted to the research of economic growth of countries by identifying patterns in historical data sets on macroeconomic indicators. Using machine learning techniques, namely cluster analysis methodology in combination with data transformation algorithms, in particular dimensionality reduction, groups of countries with similar patterns in the structure of the economy, availability of production factors, internal and external economic activity and development dynamics were formed. The novelty of the article is the approach to selecting optimal clustering and dimensionality reduction algorithms by quantifying the results of their work. The evaluation of the dimensionality reduction methods was carried out using the cumulative variance indicator, and the clustering methods were assessed based on the aggregate indicator proposed in the article, which combines the standardized Davies-Bouldin, Calinski-Harabasz indices and the Silhouette coefficient. According to calculations, among the 11 considered methods of dimensionality reduction, the most effective is the Kernel PCA algorithm, while among the 7 clustering methods, K-means is the most effective for this task with a given set of indicators. The study was conducted on 6 five-year time intervals from 1991 to 2020 with a focus on the Ukrainian economy. According to the research, Ukraine's economy migrated from the "post-Soviet" cluster (first half of the 1990s) to the Eastern European cluster (second half of the 2010s) over the period under consideration, which indicates real economic growth and gradual integration with the European Union.

Keywords: *economic growth, cluster analysis, dimensionality reduction, machine learning*

JEL Classification: C38, F43, O41

Introduction

In today’s globalized world, economic growth is becoming a key component of the stability and prosperity of countries. The growing interdependence of economies and constant changes in the geopolitical environment make the topic of economic growth modelling extremely relevant. The factors that influence economic growth are diverse and cover innovation, technological progress, trade, infrastructure and many other aspects. Understanding these factors and their interactions is key to developing effective economic management strategies. In this context, research and modelling of economic growth becomes a necessity for achieving sustainable development of countries.

The nonlinearity and complexity of economic processes pose a challenge for accurate modelling and forecasting. Traditional methods can be limited in the face of a large number of variables and their interactions. Machine learning techniques provide powerful tools to address these challenges, allowing to analyze complex patterns and dependencies in economic data. This opens up new opportunities for improving forecasting accuracy and understanding the dynamics of economic growth.

In this context, country clustering is an effective method for grouping countries with similar economic characteristics. It allows for taking into account the resemblance of countries in large amounts of data and identify subgroups with similar development paths. In this article, we will examine the essence of country clustering, highlight the benefits of using it in the context of economic growth modelling, cluster countries using a wide range of corresponding machine learning models, compare their performance based on various metrics, and identify new opportunities for further research in this area.

The purpose of our study is to model the economic growth of countries by identifying patterns in their development based on a number of macroeconomic indicators using a wide range of clustering methods. We also conduct a comparative analysis of clustering models to solve the problem of modelling economic growth based on a number of specialized metrics.

The research objectives include several key steps that can be divided into theoretical and practical parts. In the theoretical part, we will focus on the literature review starting with the problem of economic growth and its various aspects to modern methods of economic growth research using artificial intelligence and machine learning, in particular cluster analysis, paying special attention to the latter. We will then conduct a detailed qualitative analysis of dimensionality reduction and clustering methods, identifying their main advantages and disadvantages. An important task will be to evaluate the quality of the formed clusters and their analysis in the context of economic and social realities. Thus, in the practical part we will focus on the selection of optimal set of macroeconomic variables that will most accurately reflect the economic realities of countries on the collected data, after which we will apply various algorithms for data dimensionality reduction and clustering in practice, select the most effective ones for our task and made appropriate conclusions.

Analysis of recent publications on the research topic

The issue of economic growth has been relevant since the emergence of trade relations as a phenomenon. However, this sphere received special attention in the second half of the 20th century as a result of globalization processes and increasing competition between economies.

Starting with the Solow-Swan model [1-3], which laid the foundation for most of the following models, the main indicator of economic growth is the growth rate of capital in the economy, while the production process of the country’s economy is described by a production function (usually two- or three-factor). In developing this approach, it is worth mentioning the Ramsey-Cass-Koopmans model [4-6], which added household consumption to the calculation, the Uzawa-Lucas model [7, 8] and the Mankiw-Romer-Weil model [9], which proposed to complicate the production function by adding human capital, and also the Romer model [10, 11], which transformed technological progress from an exogenous factor to an endogenous one. Despite the large number of assumptions of most models (such as a closed economy or the equality of savings to investment in any given period), which reduces their realism, they remain a good analysis tool.

Another approach to forecasting economic growth is based on identifying cycles in economic development, which differ in both their duration and causes. Different types of economic cycles are described in [12-14]. Economic cycles are regular fluctuations in economic activity that are reflected in changes in consumer preferences, technological innovation, and investment activity (Kitchin cycles, which last 3-5 years) [12], fluctuations in market supply and demand, changes in credit conditions and pricing policies (Juglar cycles, lasting 7-11 years) [14], fundamental changes in production and economic structure, new technologies, industry development, and competitive environment (Kuznets swings of 15-25 years) [13], profound economic transformations, demographic changes, geopolitical shifts (Kondratieff waves lasting 45-60 years) and other [14].

With the development of economic growth theory, mathematical models became increasingly complex and the transition to machine computing became a prerequisite for further progress. With the transition to computer modeling, it became possible to apply not only predictive mathematical models, but also other techniques of artificial intelligence and machine learning. Among these relatively new methods of studying economic growth, cluster analysis occupies a special place, as it is a convenient method for identifying behavioral similarities or differences between observations that can be used in various studies. For example, in the articles of I. Strelchenko at al. [15-17], to model economic dynamics, cluster analysis is used, in particular Kohonen self-organizing maps.

Since the clustering problem is quite complex, there are a large number of machine learning algorithms for solving it, such as K-means, MiniBatch K-means [18], Spectral clustering [19], Ward’s method [20], BIRCH [21], Agglomerative clustering [22], Gaussian mixture method [23], Affinity propagation [24], Mean shift [25], DBSCAN [26], HDBSCAN [27], OPTICS [28] and others. However, in many scientific studies that use clustering methods, the choice of the most effective methodology and the optimal number of clusters is often only the authors' expert opinion and is not sufficiently substantiated, and application of metrics is limited and partial, which may reduce the effectiveness of clustering.

When the number of clusters is not known in advance, a hierarchical method for the cluster analysis is often used, which allows combining clusters based on some selected metric. For example, the hierarchical clustering method is used by V. Deltuvaitė and L. Sinevičienė [29] for clustering EU countries by financial and economic indicators, P. Enzmann and M. Moesli [30] for clustering ASEAN countries, N.-S. Koutsoukis [31] for dividing countries into political and economic groups, M. Peruzzi and A. Terzi [32] conduct clustering by economic growth characteristics. However, all of the above articles have a common problem that stems from the choice of clustering method, namely sensitivity to data outliers and noise, which is often critical when clustering countries.

To solve a similar problem of clustering countries by the nature of economic growth, the authors of [33] J. Čadil at al, use the non-hierarchical K-means method, but do not sufficiently justify the choice of the number of clusters in the article. When solving such a complex problem, forcing the selection of only 4 clusters can lead to very blurred boundaries between clusters and large intra-cluster variance.

The authors R. Cerqueti and V. Ficcadenti [34] use statistical metrics to determine the optimal number of clusters, but metrics were not used to select a method. Similarly, the authors L. Wulandari and B. Yogantara [35] used a set of metrics to estimate the optimal number of clusters for the task of clustering countries by economic and health indicators.

Therefore, given the weaknesses of the studies reviewed, there is a need for a way to quantify the results of cluster analysis and dimensionality reduction that takes into account various metrics and the choice of algorithms based on it, which would be objective and sufficiently justified, to which we dedicate our study.

The statement of basic material and methodology

Cluster analysis is a powerful tool in the arsenal of data analysis methods that is widely used in various scientific fields, including economics. This method allows you to group objects (e.g. countries, regions, enterprises) into homogeneous subgroups or clusters based on their characteristics or indicators. Cluster analysis allows to identify similarities in data that may not be visible when using traditional methods of analysis. As a result, researchers can study the characteristics of each group in more detail, identify commonalities and differences between them, and identify key factors that influence the behavior of objects in each cluster.

In the context of economic growth research, cluster analysis provides an opportunity to better understand the diversity of economic systems and processes. For example, cluster analysis can be used to identify groups with similar models of economic development, which will allow for comparative analysis of development strategies, identify key factors of economic growth, analyze disparities in the array of socio-economic indicators across regions, countries or other groups, etc.

Conducting cluster analysis requires a researcher to have a deep understanding of statistical methods and clustering algorithms, as well as careful preparation of data for analysis. For effective clustering, it is important to perform feature engineering, which includes two main stages: data standardization and dimensionality reduction. Standardization is about bringing all attributes to the same scale, thus ensuring the correct interpretation of distances between them. Dimensionality reduction, in turn, simplifies data processing by separating important attributes from less important ones. This process helps to avoid the problem of model overload (when the algorithm is unable to process the given data set in adequate time) and improves its overall performance.

Dimensionality reduction is important not only to optimize computational processes, but also to avoid the problem of multicollinearity. Multicollinearity occurs when attributes are mutually dependent, which can lead to instability and incorrect clustering results. Reducing the dimensionality allows you to eliminate unnecessary dependencies and strengthen important relationships, thereby facilitating the process of identifying homogeneous groups of objects.

There are a large number of algorithms for dimensionality reduction, but the study tested the algorithms most often mentioned in the literature: Principal component analysis (PCA), Incremental PCA, Sparse PCA, Kernel PCA [36], Singular value decomposition (SVD) [37], Sparse random projection (SRP) [38], Independent component analysis (ICA) [39], ISOMAP [40], Multidimensional scaling (MDS) [41], Local linear embedding (LLE) [42] and t-SNE [43]. Short description of these algorithms presented in Table 1.

Table 1

DIMENSIONALITY REDUCTION ALGORITHMS

Algorithm	Short description
Principal component analysis	A statistical method of reducing the dimensionality of data that searches for new orthogonal variables, called principal components, that reflect the greatest variance in the data. This method allows complex data to be expressed in terms of fewer components while retaining as much information as possible. Principal components are chosen to capture the maximum variation in the original data, reducing the dimensionality of the data and simplifying its analysis [36]. The *disadvantages* of PCA may be its sensitivity to outliers and nonlinear dependencies in the data. PCA assumes linear relationships between features, so it may be ineffective in cases where data relationships do not follow linear models. In addition, it is important to keep in mind that principal components obtained by PCA are not always easy to interpret in terms of the original features.
Incremental PCA	A variant of the principal component method designed to efficiently process large amounts of data when it is impossible to fit the entire data set into RAM. Instead of calculating the covariance matrix for the entire data set, incremental PCA allows you to calculate the principal components incrementally, optimizing memory and resource usage [36]. The *disadvantages* of Incremental PCA may be some loss of accuracy compared to traditional PCA, as the model gradually adapts to new data, and this may lead to small changes in the principal components. It should also be taken into account that Incremental PCA requires additional computing resources to keep the model up-to-date, and in this case, it may be more difficult to interpret the principal components. Incremental PCA may also perform less well in cases where the data structure changes significantly over time.

Sparse PCA	A variation of the principal component method that considers the sparsity (presence of many zero coefficients in the principal component vectors). The main goal is to extract only a small number of meaningful principal components, which allows for an effective sparse representation of the data [36]. The *disadvantages* of Sparse PCA can be the difficulty of adjusting the parameters to achieve the optimal level of sparsity. Determining the correct level of sparsity may be a non-trivial task and require additional experimentation. Also, in cases of insufficient data or improper parameter selection, Sparse PCA can lead to the loss of important information or underestimation of the real data structure. It should also be borne in mind that Sparse PCA can be demanding on computing resources, especially when working with large amounts of data.
Kernel PCA	A variation of principal component analysis that uses kernels to transform data into a higher-dimensional space so that linearly inseparable data become separable. Instead of a simple linear transformation, Kernel PCA uses kernel functions to map data into a higher dimension, allowing for efficient consideration of nonlinear relationships between features. A variety of kernels, such as polynomial, Gaussian (RBF), sigmoid, cosine, etc., can be used to identify different forms of nonlinear relationships in the data [36]. The *disadvantages* of Kernel PCA can be high computational costs, especially when dealing with large amounts of data and complex kernel functions. Choosing a suitable kernel can also be a non-trivial task, and the wrong choice can lead to inadequate dimensionality reduction and loss of important information. Another problem is that the results of Kernel PCA can be less interpretable, as new principal components are represented as combinations of kernel functions that are not always understandable in terms of the original features.
Singular value decomposition	It is a mathematical algorithm for decomposing an original data matrix into the product of three matrices: a matrix of left singular vectors, a diagonal matrix of singular values, and a matrix of right singular vectors. This decomposition allows us to represent the original matrix as the product of these three components, which reveals the structural properties and relationships in the data. SVD is widely used in recommender systems, image processing, and other areas where it is important to highlight the main aspects of large amounts of data [37]. The *disadvantages* of the SVD method are the requirement of significant computing resources and memory, especially when working with large and sparse matrices. It is also important to keep in mind that SVD can be sensitive to outliers in the data, which can lead to inadequate representation of the principal components.

Sparse random projection	A data dimensionality reduction method that uses sparse projection matrices to transform input vectors into a lower dimensional space. This method is effective for large amounts of data because it applies a random projection to the input features, creating a compact and computationally efficient representation of the data [38]. The *disadvantages* of the Sparse random projection method can be a loss of accuracy compared to other dimensionality reduction methods such as PCA or t-SNE. Since the choice of projections is random, SRP may not always effectively take into account the features and structure of specific data. It should also be borne in mind that depending on the dimensionality and structure of the data, SRP may require additional parameter tuning to achieve optimal results, and this may be a challenge when using the method in practical applications.
Independent component analysis	A method of signal processing and data analysis aimed at identifying the independent components that make up an input signal or data set. The idea is to divide the total signal or data into a set of independent components, which allows you to identify the hidden factors that cause the output signal [39]. The *disadvantages* of the Independent component analysis method can be high computational costs, especially when working with large amounts of data. It is also important to keep in mind that ICA is sensitive to the assumptions of normal data distribution and independence of components. In cases where these conditions are not met, ICA results may be less effective or require additional processing. In addition, the interpretation of the resulting independent components can be difficult, especially if the statistical properties of the data are not clearly known.
ISOMAP	ISOMAP stands for Isometric Mapping. It is a nonlinear data dimensionality reduction method that preserves the geodesic distances (takes into account the curvature of space, allowing more accurate measurements of distances between points that cannot be adequately described by straight Euclidean distances) between all pairs of points in the original space [40]. Unlike linear methods such as PCA, ISOMAP captures nonlinear relationships that can be critical for datasets with complex structures. The *disadvantages* of the ISOMAP method can be high computational costs, especially when working with large amounts of data, as it requires the calculation of geodesic distances between all pairs of points. This can lead to a significant increase in computing time and resource requirements. It should also be borne in mind that ISOMAP is sensitive to noise in the data and the influence of outliers, which can affect the quality of the result. In addition, the choice of algorithm parameters, such as the number of nearest neighbors for calculating geodesic distances, may require careful tuning to achieve optimal results.

Multi-dimensional scaling	A dimensionality reduction method that maps distances between objects in the original space to their new, low-dimensional coordinates. The main goal of MDS is to maintain the similarity of distances between objects in the original and reduced spaces. MDS attempts to preserve Euclidean distances between data points. It minimizes the difference between the distances in the high-dimensional space and the corresponding ones in the reduced space [41]. The *disadvantages* of the Multidimensional scaling method can be computational costs, especially when working with large amounts of data, as it requires calculating all possible pairs of distances between objects. This can lead to a significant increase in computing time and resource requirements. It should also be borne in mind that MDS can be sensitive to noise and outliers in the data, which can affect the accuracy of the results. The choice of distance metric and other MDS parameters may require careful tuning to achieve optimal results in a particular analysis context.
Local linear embedding	A method of nonlinear data dimensionality reduction that preserves local linear structures in the data. LLE treats each data example as a linear combination of its nearest neighbors' parameters and finds the optimal weighting coefficients to represent this example in a low-dimensional space. The main idea is to preserve the local relationships between neighboring points while reducing the dimensionality, which allows to take into account nonlinear structures in the data [42]. The *disadvantages* of the Local linear embedding method can be sensitivity to the choice of parameters, such as the number of neighbors and the dimensionality of the low-dimensional space. Incorrectly chosen parameters can lead to inadequate data representation and loss of important information. It is also important to keep in mind that LLE can be computationally intensive, especially when working with a large amount of data or in problems with a large number of neighbors for each point. Another disadvantage may be that LLE is less effective in cases where the global data structure is more important than local relationships.
t-SNE	An embedding method for data dimensionality reduction that aims to map similar objects close to each other in a low-dimensional space. t-SNE works by attempting to preserve the neighborhood between points in the original and reduced space. The basic idea is to preserve the similarity between points by increasing the neighborhood probabilities for similar points and decreasing them for dissimilar points [43].

	The *disadvantages* of the t-SNE method can be its computational complexity and execution time, especially when working with large amounts of data. It is also important to keep in mind that t-SNE can selectively map distances between points in a high-dimensional space, which can lead to an over-focus on local structures. Also, t-SNE has the property of changing the similarity of objects, which can cause distortion of some distances and structures when reducing the dimensionality. In addition, the choice of algorithm parameters may require research and experimentation to achieve optimal results.

To assess the efficiency of the process of dimensionality reduction, we will use cumulative explained variance (1), which indicates the proportion of variation in the original data that remains after dimensionality reduction. The higher the cumulative explained variance remains, the more important information characteristics are preserved, and the more effective the dimensionality reduction is.

$$var_c = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{n} \lambda_j}, \tag{1}$$

where m is a number of selected factors (initial features, principal components, etc.), n – total number of factors, λ_i – the eigenvalue of covariance matrix for selected factor i.

To conclude the analysis of dimensionality reduction algorithms, it is worth noting that these methods play an important role in preparing data for clustering. Reducing the number of dimensions often improves the efficiency and accuracy of clustering algorithms. Now we turn to a qualitative analysis of clustering methods that allow us to identify the internal structure of data and group objects based on their similarities. Let's consider the main approaches to clustering, their advantages and disadvantages, as well as the criteria by which the quality of the resulting clusters is assessed.

For the study, the following clustering algorithms were chosen: K-means, MiniBatch K-means [18], Spectral clustering [19], Ward's method [20], BIRCH [21], Agglomerative clustering [22], Gaussian mixture method (GMM) [23], Affinity propagation [24], Mean shift [25], DBSCAN [26], HDBSCAN [27] and OPTICS [28]. Short description of these algorithms presented in Table 2.

Table 2

CLUSTERING ALGORITHMS

Algorithm	Short description
K-means	The K-means clustering algorithm is one of the most common grouping methods in machine learning. It is based on dividing a dataset into K clusters, where K is a predefined number of clusters. Starting from the initial cluster centroids, the algorithm iteratively updates the cluster membership of objects and the location of the centroids to minimize the sum of the squared distances between objects and their assigned centroids. The process continues until convergence, when the cluster composition and the location of the centroids stop changing significantly. The K-means algorithm is efficient and fast, but it is sensitive to the starting points, and its results depend on the set number of clusters [18]. The *disadvantages* of the K-means method can be the sensitivity to the initial location of the centroids, which can affect the final result and the variations in clustering in new runs of the algorithm. In addition, K-means assumes that the clusters are homogeneous and have a spherical shape, which may make it less effective for some types of data where clusters may have complex shapes or uneven densities. Also, the algorithm is sensitive to outliers and noise in the data, which can lead to inaccurate clustering. The choice of the number of clusters can also be a non-trivial task and require preliminary expertise or the use of additional methods to determine it.
MiniBatch K-means	It is a variant of K-means designed to process large amounts of data using mini-samples instead of the full data set. The algorithm is based on randomly selecting small subsets of data and using them to update cluster centroids and cluster membership of objects at each iteration. This makes the algorithm faster and more scalable, well suited for large datasets, while still demonstrating similar accuracy to regular K-means [18]. The *disadvantages* of MiniBatch K-means may be a loss in accuracy compared to the full version of K-means. Since the centroids are updated based on subsamples of data, this can affect the stability and convergence of the algorithm, especially in situations where there are large or sparse clusters in the data. It should also be borne in mind that due to the use of random subsamples, the results of MiniBatch K-means may be less detailed and sensitive to initial conditions compared to the full version of the algorithm. The choice of parameters, such as the size of subsamples (batch size), may require additional adjustment to achieve optimal results in a particular case.

Spectral clustering	A data grouping method based on the use of the properties of spectral graph theory. The algorithm is based on converting the similarity matrix between objects into a graph structure, where vertices represent objects and edges reflect the levels of their similarity. By analyzing the eigenvectors of the graph matrix, objects are divided into clusters using eigenvalue and eigenvector method. Spectral clustering is effecttive for detecting non-standard and non-linear structures in data [19]. The *disadvantages* of the spectral clustering method can be high computational costs, especially when working with a large amount of data or high dimensional space. The calculation of eigenvectors and eigenvalues can be a resource-intensive task. In addition, the choice of parameters, such as the number of clusters or the type of similarity between objects, may require expert judgement and careful tuning to achieve optimal results. In the case of large amounts of noisy data or outliers, spectral clustering may perform less well, as it may try to incorporate noise into the cluster structure.
Ward's method	Ward's method for hierarchical clustering is a method based on minimizing intra-cluster variances in the process of data grouping. The essence of the method is to calculate the variance of objects within clusters and combine those clusters that lead to minimizing the increase in the total intra-cluster variance after their union [20]. The *disadvantages* of the Ward's method can be its sensitivity to outliers and the complexity of calculations on large data volumes. In the presence of outliers or noise, this method may produce incorrect or less stable clusters. Also the Ward's method is not always effective for data with complex non-elliptical cluster shapes. In addition, the choice of distance or similarity measure between clusters can affect the result, and even with careful selection of these parameters, Ward's method may underperform other clustering methods in some scenarios.
BIRCH	It is a method designed for efficient clustering of large amounts of data. The algorithm is based on building a tree known as a CF-tree (Clustering Feature tree), where each node represents a cluster and the leaves represent a group of objects. BIRCH uses the technique of calculating the main characteristic vector (CF-values) for each node of the tree, which allows efficient processing and updating of clusters, even with a large amount of data [21]. The *disadvantages* of the BIRCH method may be lower clustering accuracy compared to some other methods, especially in situations where clusters have complex shapes or are arranged hierarchically. Due to the approach of using global and local thresholds to define clusters, BIRCH may not always effectively recognize and separate clusters with different densities and shapes. In addition, the choice of parameters such as the radius of influence and the separation threshold may require extensive tuning to achieve optimal results in a particular context.

Agglo-merative clustering	The essence of the method is that each object is initially considered as a separate cluster, and then at each step the most similar clusters are combined until one common cluster remains. The method of agglomerative clustering is based on a dendrogram to determine the optimal number of clusters. On the dendrogram, it is necessary to find the largest vertical difference between nodes, and draw a horizontal line in the middle. The number of vertical lines crossing it is the optimal number of clusters. The *advantages* of Agglomerative clustering include ease of implementation, no need to determine the number of clusters before starting the analysis, and the ability to use different similarity metrics. *Disadvantages* include high computational complexity, especially with a large number of objects, and vulnerability to outliers or incorrectly chosen similarity metrics, which can lead to suboptimal results [22].
Gaussian mixture method	It is a statistical approach that models a data set as the sum of several Gaussian (normal) distributions. The essence of the method is that each cluster is associated with a Gaussian distribution with its own parameters – mean, covariance matrix, and weight. Using the EM (Expectation-Maximization) algorithm, GMM seeks to find the most likely set of parameters that best explain the observed data. One of the key features of GMM is that each object can belong to several clusters with different probabilities [23]. The *disadvantages* of the Gaussian mixture method can be high computational costs, especially when dealing with large amounts of data and high dimensional space. Since the EM algorithm used to estimate the model parameters is iterative, a large amount of data can lead to increased computational time. In addition, the GMM can be sensitive to the choice of initial parameters, which can affect the stability and convergence of the algorithm. In cases where the data distribution does not correspond to a Gaussian function, or where clusters have nonlinear relationships, GMM may not produce optimal results because it assumes only Gaussian components.
Affinity propagation	Instead of a predefined number of clusters, as is typical for many other algorithms, Affinity propagation determines clusters by selecting a set of "epicenters" – representatives with inter-object "affinity" determined by the similarity matrix between objects. During the iterative process, the epicenters are selected and updated based on the interaction between the objects, and the number of clusters is determined automatically. The algorithm allows to effectively detect non-standard and irregular data structures [24].

	The *disadvantages* of the Affinity propagation method can be high computational complexity, especially when working with large amounts of data. The algorithm requires computing and updating the similarity matrix between all pairs of objects, which can lead to increased runtime for large datasets. It is also important to keep in mind that Affinity propagation is prone to forming many clusters in cases where other methods can combine them into more aggregated groups. The choice of parameters, such as the propensity score, may require research and experimentation to achieve optimal results in a particular case.
Mean shift	A method that identifies clusters by finding the maximum of the probability density function in the feature space. The algorithm is based on selecting a starting point (usually random) in the data space, and then calculating the average value (mean shift) of all points that are in a certain window around the starting point. This process is repeated iteratively, moving the selected points in the direction of the shift until the points converge to a local maximum of the density function. Thus, when the points converge to stable positions, or the shift becomes too small, the regions of point convergence will be defined as clusters [25]. The *disadvantages* of the Mean shift method can be high computational costs, especially when processing large amounts of data or high dimensional space. Since Mean shift requires a kernel width parameter, choosing the right value can be a non-trivial task and affect the quality of clustering. In addition, the algorithm can be sensitive to initial conditions, and the results may vary depending on the starting point set. In cases where the clusters have unequal densities or when they are located at different levels of the hierarchy, Mean shift may produce less optimal results compared to other clustering methods.
DBSCAN	A method that identifies clusters in the data space based on their density. The algorithm is based on the fact that for each object, its number of neighbors is determined using the radius and the minimum number of neighbors. Objects that have a sufficient number of neighbors within the specified radius form clusters. An important feature of DBSCAN is the ability to detect clusters of any shape and identify noise points that are not part of any cluster [26]. The *disadvantages* of DBSCAN may be its sensitivity to the choice of parameters, such as the radius and the minimum number of points in the neighboring region. An incorrect choice of these parameters can lead to incorrect clustering or cluster merging. It can be especially difficult to determine the optimal values of the parameters in high-dimensional spaces. DBSCAN can also have problems detecting clusters of different densities. In addition, the variability in cluster shapes and sizes can lead to difficulties in selecting appropriate parameter values for a particular data set.

HDBSCAN	The essence of the algorithm is to use the density and distance between objects to define clusters, but HDBSCAN adds the ability to define hierarchical structures in the data. Using the "minimum density" parameter and different density levels, the algorithm forms clusters and their hierarchy. It also takes into account noise points and objects that cannot be uniquely assigned to any cluster [27]. The *disadvantages* of the HDBSCAN method can be high computational complexity, especially when working with large amounts of data or high dimensional space. The algorithm requires the construction and analysis of a hierarchical structure for clustering, which can affect the execution time. The choice of parameters, such as the minimum number of objects in the cluster and the radius, can be a non-trivial task and may depend on the specific dataset. HDBSCAN may also perform less well in areas where clusters have different densities and shapes and require careful parameter tuning to achieve optimal results.
OPTICS	It is a method designed to identify the structure of clusters in a data set without a predefined number of clusters. The essence of OPTICS is to map the data space into a so-called "reachability plot", similar to a dendrogram, in which objects are arranged according to their "reachability" from other objects. Based on this dendrogram, clusters and their hierarchy can be identified, as well as outliers. OPTICS can detect clusters of different densities and shapes, adapting to complex data structures [28]. The *disadvantages* of the OPTICS method can be high computational complexity, especially when dealing with large amounts of data or high dimensional space. Since the algorithm depends on building and analyzing a graph of relationships between points, this can lead to a significant increase in runtime for large data. OPTICS can also prove difficult in cases where clusters have different densities and sizes, or where the data space has a large number of dimensions. In addition, it is not always easy to determine the optimal parameters, such as the radius and minimum number of objects in a cluster, to achieve the best results for a particular dataset.

After getting acquainted with the main clustering methods, there is a need to choose the most optimal one. Expert opinion will not always help in this matter, as there are many nuances with the subject area and research data, so there is a risk that the most optimal method will not be chosen. For this reason, there is a need for a numerical assessment of the quality of clustering results. This study used an aggregate measure of clustering quality based on 3 indicators, namely the Davies-Bouldin index [44], the Calinski-Harabasz index [45], and the Silhouette coefficient [46].

The Davies-Bouldin index [44] is a metric for evaluating the quality of clustering in machine learning. This index takes into account distances within clusters and between cluster centroids to determine how well defined and separated the clusters are in a clustering task. More specifically, the Davies-Bouldin index is calculated through the ratio of the weighted average of internal distances between points in the same cluster and the external distances between clusters. The lower the index, the better the clusters are separated and defined. The index is calculated by the formula:

$$DBI = \frac{1}{N}\sum_{i=1}^{N} max_{j\neq i}\frac{S_i + S_j}{M_{i,j}}, \tag{2}$$

where N – number of clusters, S_i – average distance from each point in the cluster C_i to the cluster centroid A_i, $M_{i,j}$ – is the distance between the centroids A_i and A_j of clusters C_i and C_j, respectively.

S_i calculated as:

$$S_i = \frac{1}{T_i}\sum_{l=1}^{T_i}\left\|X_i^l - A_i\right\|, \tag{3}$$

where T_i is the size of the cluster C_i, X_i^l is a point l belonging to a cluster C_i, $\left\|X_i^l - A_i\right\|$ denotes the Euclidean distance between the vectors of the point X_i^l and of the centroid of the cluster A_i.

$M_{i,j}$ is calculated as the Euclidean distance between centroids of clusters C_i and C_j:

$$M_{i,j} = \left\|A_i - A_j\right\|. \tag{4}$$

The Calinski-Harabasz index [45] is based on the intra- and inter-cluster distances in the feature space and is used to determine how well defined and separated the clusters are. Higher values of the index indicate better clustering quality. The index is calculated by the formula:

$$CH = \frac{\sum_{i=1}^{N} T_i \|A_i - A\|}{N-1} \cdot \frac{T-N}{\sum_{i=1}^{N} \sum_{l=1}^{T_i} \left\| X_i^l - A_i \right\|}, \tag{5}$$

where T – size of the entire data set, A – centroid of the entire data set.

In formula (5), the first multiplier corresponds to the inter-cluster variance, and the second multiplier to the inverse of the intra-cluster variance.

The Silhouette coefficient [46] takes into account the degree of separation between clusters and homogeneity within clusters, and is expressed as a numerical value between –1 and 1. High value of the Silhouette coefficient indicates that objects in each cluster are well separated from each other and belong to a homogeneous group, while low value indicates inconsistency and overlap between clusters. The Silhouette coefficient is calculated by the formula:

$$SC = max_i \ \frac{1}{T_i} \sum_{l=1}^{T_i} \frac{b(X_i^l) - a(X_i^l)}{max\{a(X_i^l), b(X_i^l)\}}, \tag{6}$$

where $a(X_i^l)$ is the average distance between point X_i^l of cluster C_i and all other points in this cluster:

$$a(X_i^l) = \frac{1}{T_i - 1} \sum_{k=1, k\neq l}^{T_i} \|X_i^l - X_i^k\|, \tag{7}$$

$b(X_i^l)$ is the minimum average distance between point X_i^l of cluster C_i and all points of each of the other clusters:

$$b(X_i^l) = min_{j\neq i} \frac{1}{T_j} \sum_{k=1}^{T_j} \|X_i^l - X_j^k\|. \tag{8}$$

After calculating the performance indicators (2), (5) and (6), we will have their data array in the context of each clustering method and each variable parameter of the method (in particular, the number of clusters, etc.). Since all the calculated indices have different orders of magnitude, it is not correct to add them together or calculate the average value. Therefore, we propose an aggregate indicator (9) that will take into account the magnitude and the direction of each of the clustering quality indices (the higher the SC and CH indices, the better the results, and vice versa with the DBI index):

$$AV = \frac{SC - \overline{SC}}{\sigma_{SC}} + \frac{CH - \overline{CH}}{\sigma_{CH}} - \frac{DBI - \overline{DBI}}{\sigma_{DBI}}, \tag{9}$$

where $\overline{SC}$ – mean value of the index SC among all experiments (for different algorithms and their parameters), σ_{SC} – standard deviation of the index SC for all experiments conducted, similarly $\overline{CH}$ – mean value of the index CH, σ_{CH} – standard deviation of the index CH, $\overline{DBI}$ – mean value of the index DBI, σ_{DBI} – standard deviation of the index DBI.

Results and discussion

Economic growth is a complex and multidimensional category defined by the increase in a certain indicator (real GDP per capita or capital intensity of production, for example) of a country over a certain period. This process can be caused by various factors that interact and influence economic activity. There are several key factors that determine economic growth [47]. Productive factors, such as capital and labor, play an important role in creating a productive economic base. Human capital, which includes the education and skills of the population, is also an important component, contributing to innovation and productivity. International trade opens up new markets and opportunities for expanding economic interconnections. Natural resources, the tax burden, public spending and public capital also influence economic growth, determining its pace and sustainability. Taking these factors into account collectively determines the success of a development strategy and contributes to sustainable economic growth.

In addition, when modelling economic growth, it is important to take into account such a key aspect of macroeconomic dynamics as savings (investment) and consumption [1-11]. Savings, which are converted into investments, stimulate economic development by providing additional capital to support entrepreneurship and implement new projects. Investments contribute to production, technological development and job creation, which in turn strengthens economic activity. On the other hand, an increase in consumer spending can boost production and stimulate business activity. In this case, an important factor is the interdependence of consumption and savings: the higher the share of consumption in GDP, the lower the level of savings, and, consequently, domestic investment.

When selecting variables for cluster analysis of countries in the context of economic growth, a number of factors need to be carefully considered to ensure reliable and high-quality results. Firstly, it is important to avoid absolute indicators as much as possible, since economic conditions may differ widely between countries. Instead, relative indicators such as GDP per capita growth should be considered. But a country's size and population can also affect economic growth, so it is advisable to keep a few absolute indicators.

The second aspect to consider is the diversification of variables. It is important to take into account the various characteristics of the economy, as economic growth depends on a complex set of factors. Additionally, time series and dynamics of change should be taken into account, as economic indicators can change over time, and this is important for identifying trends and sustainability. Clustering methods allow for a balanced analysis and help to create a comprehensive picture of the economic situation of countries in the context of their growth.

Taking into account the data requirements for clustering and their availability in the World Bank database [48], which is the most comprehensive for analyzing the economic growth of countries worldwide, we selected the following indicators: GDP and GDP growth to reflect the overall scale of the economy and its dynamics; GDP per capita and its growth to take into account the level of development without the influence of the size of the economy [49]; shares of the main sectors of the economy (agriculture, industry and services) to take into account the specialization of the country and the type of society (pre-industrial, industrial, post-industrial); capital intensity as the main indicator of economic growth used in the relevant models [1-11]; indicators illustrating the availability of basic factors of production such as capital, labor and land; population growth and its age structure, which determine human capital (another important factor of production); the level of integration into world trade and the nature of integration (exporting or importing country) are explained by such indicators as the share of foreign trade in GDP and the ratio of imports to exports [16]; by analogy, we add consumption and savings indices that characterize the domestic consumer; as well as the tax burden, inflation, the level of expenditures and revenues in the economy to account for the level of governance. A detailed list of variables can be seen in Table 3.

Most indicators from Table 3 have a maximum observation period from 1960 to 2021 at the time of the study. Taking into account various internal conditions, such as socio-political situation, data privacy, the level of statistics management in different countries, the completeness of the data for many countries is not sufficient to conduct research on the entire specified interval. Therefore, the study was limited to the period from 1991 to 2020. The choice of the lower limit is explained by the fact that 1991 is a key year in political terms (the collapse of the Soviet Union and the loss of its influence on other countries), and the choice of the upper limit is explained by the significantly lower completeness of the data for 2021 compared to previous years. In this interval, the vast majority of countries have sufficient official data for modelling (see Fig. 1).

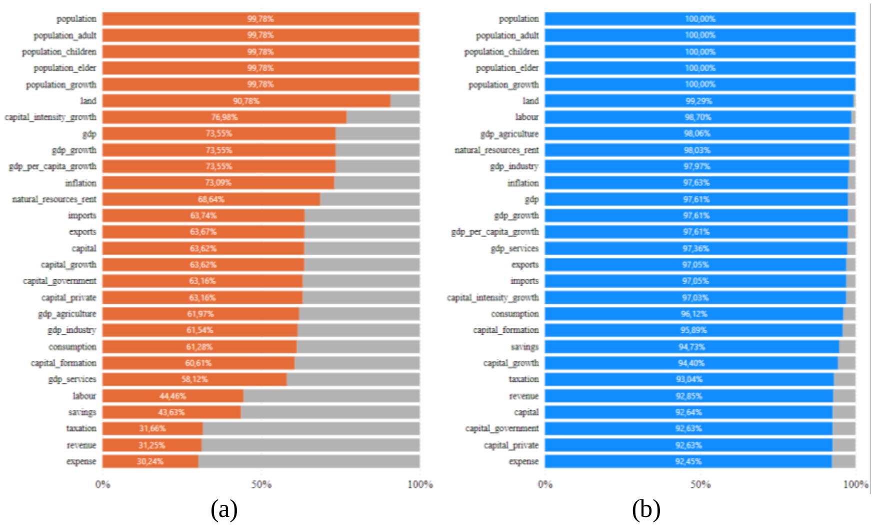
Similarly, countries with less than 85% data completeness (i.e., more than 15% missing data) were excluded. This left 150 countries out of 217 in the study (see Fig. 2). Thus, small island countries, dependent territories of other countries, and some underdeveloped countries were excluded from the overall database.

Table 3

VARIABLES FOR CLUSTER ANALYSIS

Variable name	Variable description
average_gdp	Average GDP (constant 2015 US$)
gdp_growth	Average GDP growth (annual %)
gdp_per_capita	Average GDP per capita (constant 2015 US$)
gdp_per_capita_growth	Average GDP per capita growth (annual %)
gdp_agriculture_perc	Average agriculture, forestry, and fishing, value added (% of GDP)
gdp_industry_perc	Average industry (including construction), value added (% of GDP)
gdp_services_perc	Average services, value added (% of GDP)
capital_intensity	Ratio of gross capital (constant 2015 US$) to labor force (total)
capital_intensity_growth	Average ratio of gross capital (constant 2015 US$) to labor force (total) growth
capital_capacity	Ratio of gross capital (constant 2015 US$) to GDP (constant 2015 US$)
capital_private_rate	Ratio of gross government capital (% of GDP) to gross private capital (% of GDP)
population	Average population (total)
land	Average agricultural land (sq. km)
population_growth	Average population growth (annual %)
population_asymmetry	Ratio of population ages 0-14 (% of total population) to population ages 65 and above (% of total population)
trade_rate	Average trade (% of GDP)
trade_asymmetry	Ratio of imports of goods and services (% of GDP) to exports of goods and services (% of GDP)
consumption_rate	Average final consumption expenditure (% of GDP)
savings_rate	Average gross savings (% of GDP)
consumption_asymmetry	Ratio of final consumption expenditure (% of GDP) to gross savings (% of GDP)
tax_burden	Average tax revenue (% of GDP)
inflation	Average inflation, consumer prices (annual %)
expense_level	Average expense (% of GDP)
revenue_level	Average income, excluding grants (% of GDP)
res_outturn	Average total natural resources rents (% of GDP)

Note. The average value means the annual arithmetic mean for the considered period.



(a) (b)

Fig. 1. Completeness of data (a) for the full period from 1960 to 2021, and (b) for the sample represented in the study – 1991-2020

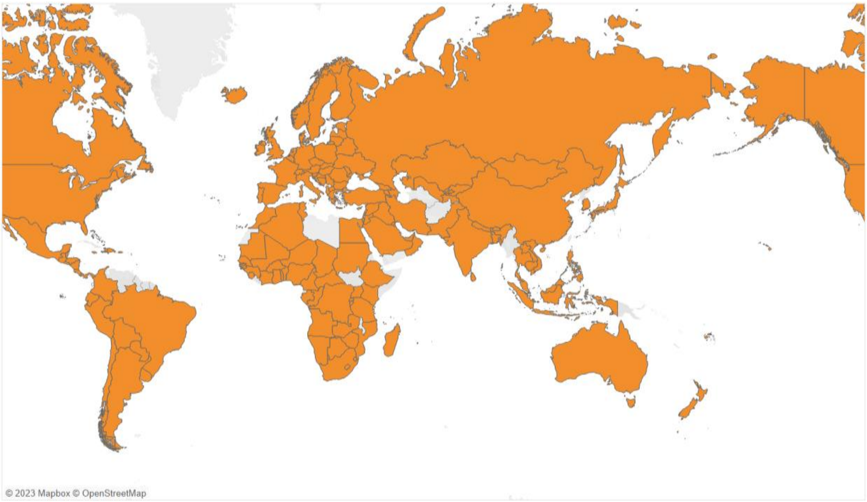

© 2023 Mapbox © OpenStreetMap

Fig. 2. Countries participating in the research

The use of clustering and dimensionality reduction algorithms requires 100% data completeness for all indicators. But if we continue to reduce the number of countries, we will not get a significant increase in data completeness, while greatly truncate the database. In this case, it was decided to replace the gaps in the data with plausible values. The remaining gaps in the data were replaced by the following formula:

$$I_i = \frac{I_j I_{wi}}{I_{wj}}, \tag{10}$$

where I_i – missing value for the period i, I_j – the value of the same indicator in the non-empty period j closest to i, I_{wi} – the total value of the desired indicator in the period i for all countries of the group to which, according to the World Bank methodology, the country with the missing indicator belongs (for example, for the Asian countries with Low Income), I_{wj} – the total value of the desired indicator for all countries of the same group in the period j.

After the list of variables is formed and the data is cleaned, in order to continue the study, data preparation, which includes data standardization and dimensionality reduction, is necessary before clustering. Standardization is performed to ensure that variables with different magnitudes (such as GDP in billions of dollars and consumption as a percentage of GDP) have the same impact on the clustering result. And dimensionality reduction is aimed at getting rid of multicollinearity of variables (for example, when high GDP per capita is correlated with the share of services in GDP, as both indicators often indicate a developed economy and if this phenomenon is not eliminated, certain characteristics of countries will affect the results much more than others), as well as reducing the amount of data as much as possible with minimal loss of information for faster work at further stages of the study.

To standardize the data, we used the Z-standardization method, which is very common in statistics and modeling. For dimensionality reduction, we tested 11 algorithms from Table 1 and 5 customization options of algorithms, the full list of which is presented in Fig. 3.

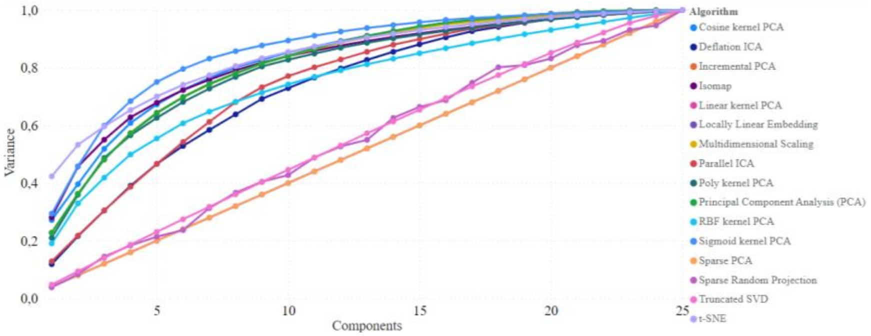


Fig. 3. Results of testing dimensionality reduction algorithms on the study data

The Fig. 3 shows the cumulative explained variance calculated by formula (1) for each of the dimensionality reduction methods depending on the number of principal components, indicating what percentage of the information is retained from the original volume.

We found that the most effective method is Kernel principal component analysis with sigmoid kernels, as can be seen in Fig. 3. Any method of dimensionality reduction cannot avoid loss of information, but usually 5% information loss is acceptable. The above method allows to reduce the dimensionality to 15 principal components while retaining 95.76% of the information.

After completing the previous stage, the data is ready for clustering. However, first it is necessary to filter out algorithms that are obviously not going to provide high results, and also to select the required number of clusters. This is done in order to reduce the time and computing power required for the research.

The automatic selection of the number of clusters in our case is a disadvantage, since the observations in the data are very scattered and it is often difficult to draw a clear boundary between them. Therefore, the use of automatic algorithms will lead to unrepresentative results. Thus, this factor led to the rejection of the Affinity propagation and Mean shift algorithms.

Since this study is primarily concerned with the analysis of the economy of Ukraine, it is necessary to ensure clustering of all observations without exception, otherwise there is a risk that the Ukrainian economy will not be included in any of the clusters. In the test dataset of the study [50], presented in Fig. 4, which is the most similar to the data of this research, black dots correspond to unclustered observations. To avoid such a situation, we exclude the algorithms DBSCAN, HDBSCAN, and OPTICS.

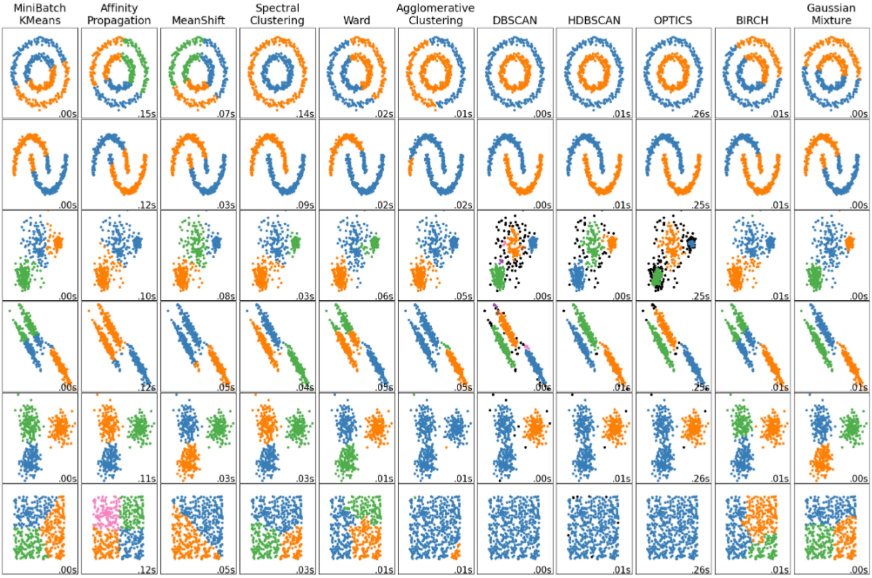


Fig. 4. Testing clustering algorithms on different types of data samples [50]

We will determine the optimal number of clusters by the aggregate indicator given in formula (9). It is worth noting that we decided to set the optimal number of clusters within the range from 7 to 20, firstly, to reduce the time spent on calculations, and secondly, it makes no sense to consider less than 7 clusters because of too large groups of countries, which may include very different observations. If there are more than 20 clusters, then it is difficult to visualize such data correctly, moreover, many clusters will be unbalanced and consist of 1-2 observations.

At this stage, everything is ready for the cluster analysis, the input data for which will be the macroeconomic indicators from Table 3, averaged over the period from 1991 to 2020, normalized and reduced to 15 principal components, with each of the 150 countries represented by a separate entry. The top 20 clustering results, evaluated by formula (9), are presented in Table 4.

Table 4

TOP-20 MOST EFFECTIVE CLUSTERING ALGORITHMS ON THE STUDY DATA

Algorithm	*N*	*DBI*	*CH*	*SC*	*AV*
K-means	17	1,00657	18,43062	0,18542	2,18810
BIRCH	20	1,01926	18,20109	0,17919	1,95720
Agglomerative clustering	20	1,01926	18,20109	0,17919	1,95720
Ward’s method	20	1,01926	18,20109	0,17919	1,95720
K-means	16	1,04798	17,93731	0,17993	1,79440
K-means	20	1,08612	18,52437	0,18057	1,77329
BIRCH	19	1,07294	18,35439	0,17595	1,69737
Agglomerative clustering	19	1,07294	18,35439	0,17595	1,69737
Ward’s method	19	1,07294	18,35439	0,17595	1,69737
MiniBatch K-means	18	1,06567	17,92405	0,17800	1,67745
K-means	18	1,08950	17,74273	0,18115	1,60237
BIRCH	17	1,08250	18,53183	0,16927	1,55808
Agglomerative clustering	17	1,08250	18,53183	0,16927	1,55808
Ward’s method	17	1,08250	18,53183	0,16927	1,55808
K-means	7	1,24845	20,37201	0,18181	1,51237
BIRCH	18	1,08840	18,44755	0,16782	1,48533
Agglomerative clustering	18	1,08840	18,44755	0,16782	1,48533
Ward’s method	18	1,08840	18,44755	0,16782	1,48533
MiniBatch K-means	16	1,12667	18,98103	0,16914	1,46594
K-means	10	1,14692	19,03277	0,17030	1,41546

According to Table 4, besides the fact that the K-means method being the most effective for the described problem, it was included among the top 20 clustering results 6 times with different number of clusters, which is the maximum among the algorithms considered. Also, the Agglomerative clustering and BIRCH algorithms, as well as

Ward's method, gave fairly good results on the simulation data, which were included in the table 4 times each.

Having chosen the most effective method with the optimal parameters (K-means with division into 17 clusters), we now proceed to describe the results of the cluster analysis. First, we will cluster the countries for the full period of the study – based on the average values for 30 years: from 1991 to 2020. Fig. 5 (as well as all subsequent ones) shows 2 maps of the world: the larger one highlights the cluster to which Ukraine belongs, and the map in the lower right corner shows all 17 clusters in different colors.

In general, Ukraine falls into a cluster of countries with which it has very little in common, as the averaging period was too long, and periods of economic growth overlapped with periods of unstable military and political situation and economic crises, which distorted the results to some extent. In this cluster, together with Ukraine, are Belize, Ecuador, Jordan, Kyrgyzstan, North Macedonia, South Africa, Tunisia, Uruguay, Venezuela and a number of other countries in Africa and Central America (see Fig. 5).

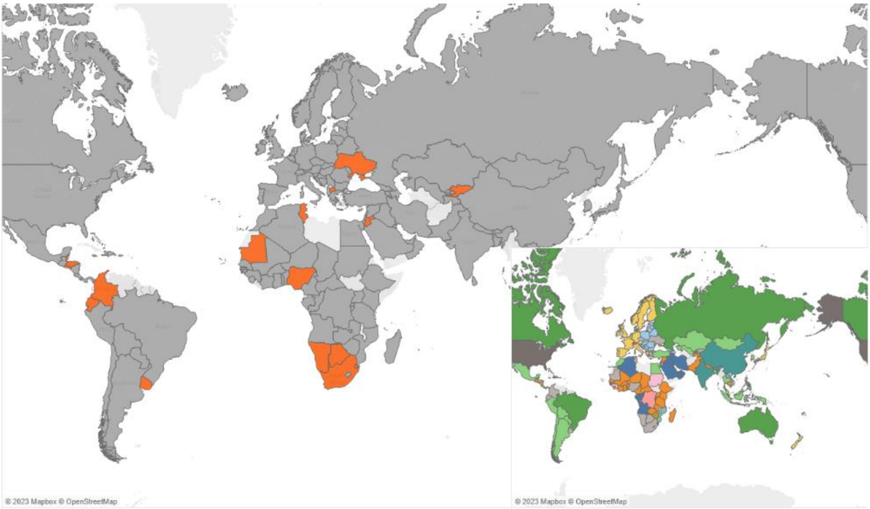


Fig. 5. Results of clustering countries based on average values of indicators for the full period (1991-2020)

Analysis of clustering results in Fig. 5 shows that averaging the indicators does not lead to the desired level of discrimination between clusters. One of the main reasons for this is the significant variability of the data over a long period of time, which leads to a loss of detail where the most important events and changes may be missed or undercounted in the overall average approach.

To solve this problem, it is proposed to divide a large period into smaller time intervals and perform clustering in each of them separately. This approach will avoid loss of detail due to over-averaging and take into account the specifics of each time period, while removing data outliers that may occur in certain years. Each segment can represent a certain stage of development, economic cycle or other important changes, which helps to improve understanding of the dynamics and structure of the data.

Thus, we divide the 30-year period into 6 5-year intervals. Fig. 6 shows the result of clustering on the data of the first of them, which is from 1991 to 1995. For the clustering data of this interval and the following ones, the same method is used as for the full period, namely K-means with division into 17 clusters.

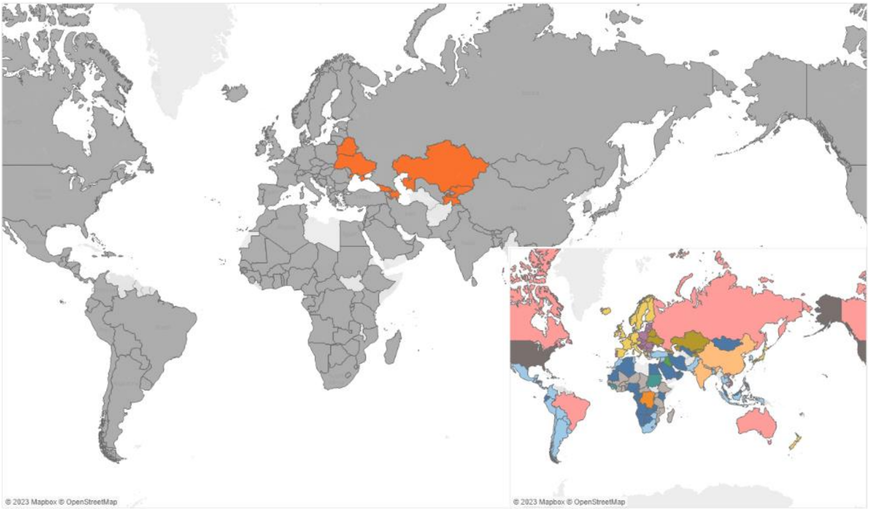


Fig. 6. Results of clustering countries based on data for the period 1991-1995

Fig. 6 shows that in 1991-1995 Ukraine was a part of the “post-Soviet cluster”, which is characterized by economic problems in the transition from a planned to a market system. This cluster is characterized by high inflation, generally low economic growth, negative GDP growth, low savings, and a high level of economic industrialization. Ukraine, located in this cluster, is indicative of the complex challenges encountered in adapting to the new market conditions, as defined by the economic difficulties that often arise during the transition period.

By analogy, we will perform a cluster analysis of the next 5-year period from 1996 to 2000 (see Fig. 7).

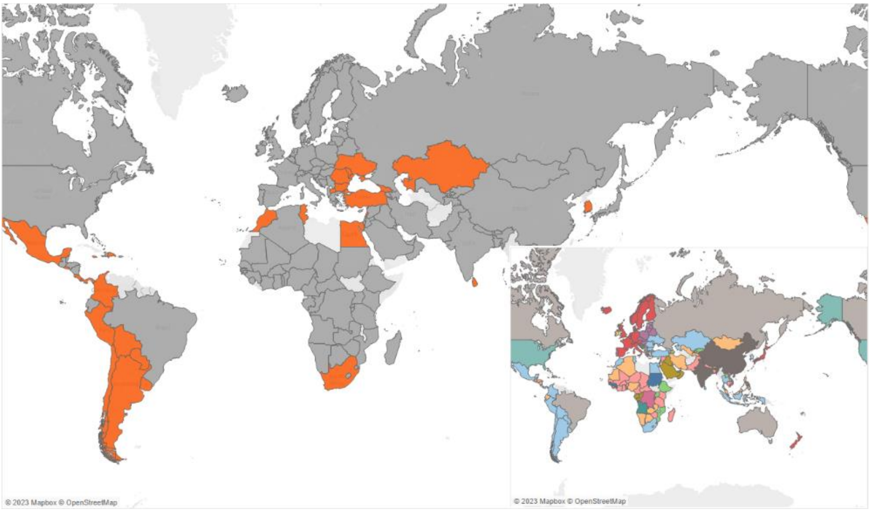


Fig. 7. Results of clustering countries based on data for the period 1996-2000

As can be seen from Fig. 7, between 1996 and 2000 Ukraine was in a cluster with most of Latin American countries, some countries in South-Eastern Europe, Kazakhstan and developed African countries. This cluster includes countries characterized by accelerated economic growth due to the effect of a low base and the introduction of market mechanisms into the economy.

The next period of clustering is from 2001 to 2005, the results of which are shown in Fig. 8.

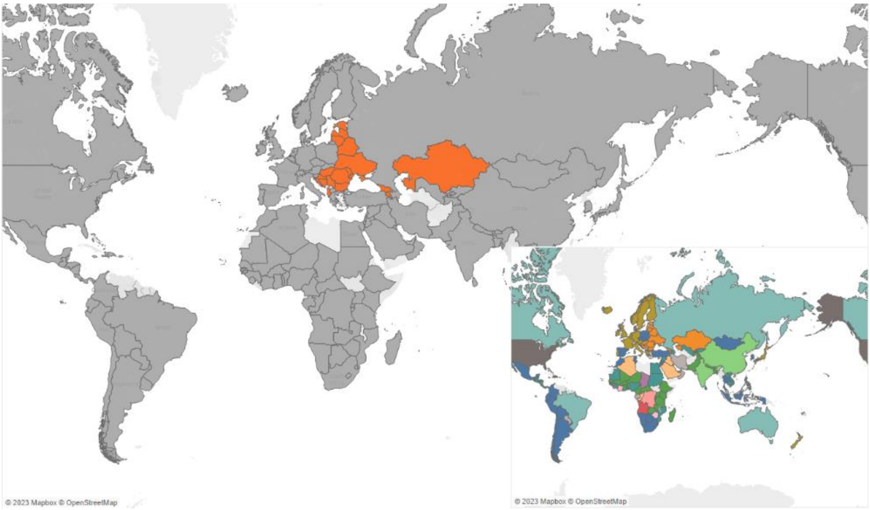


Fig. 8. Results of clustering countries based on data for the period 2001-2005

Fig. 8 shows Ukraine in a cluster with the Balkan, Baltic, Caucasian countries and Kazakhstan. This cluster is characterized by certain common features, such as an emphasis on integration into the global market, economic reforms and liberalization. Ukraine, being a member of this cluster, demonstrates attempts to address economic challenges through active participation in global economic processes.

Next, similarly to previous periods, we cluster countries based on average indicators from 2006 to 2010, which can be seen in Fig. 9. In the period 2006-2010, Ukraine was part of a cluster that unites most of the Eastern European countries (see Fig. 9). This cluster is defined by common characteristics, such as the implementation of economic reforms, active participation in European integration, and similar reactions to the 2008 global crisis.

After that, the next interval for clustering is the period from 2011 to 2015 (see Fig. 10).

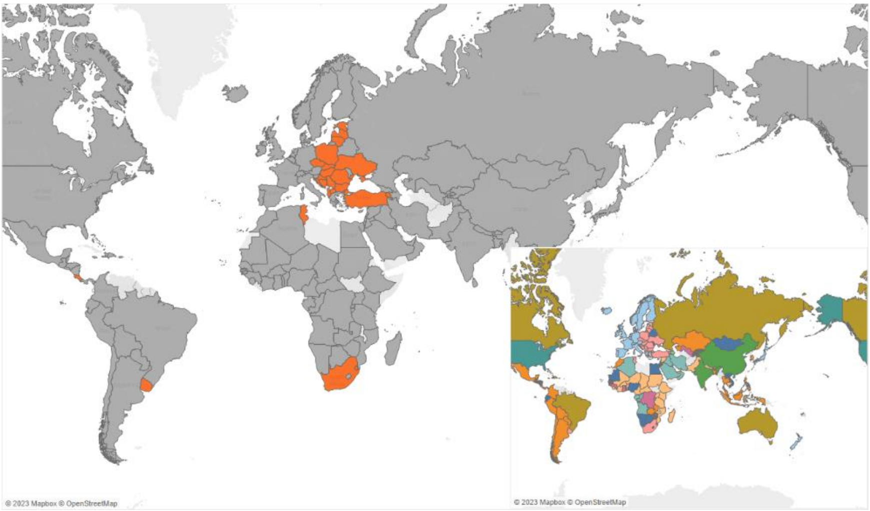


Fig. 9. Results of clustering countries based on data for the period 2006-2010

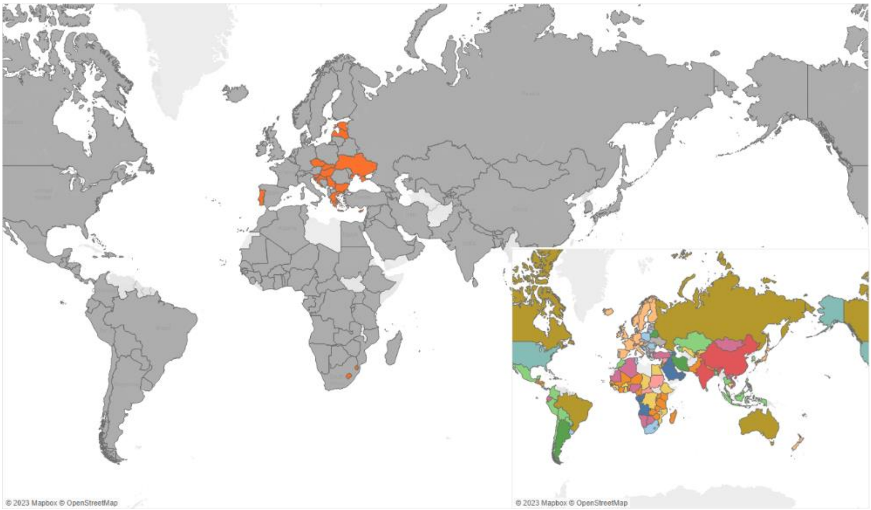


Fig. 10. Results of clustering countries based on data for the period 2011-2015

During 2011-2015 Ukraine is part of a cluster that includes mainly Balkan and Baltic countries (see Fig. 10). In this period, these countries demonstrated certain common characteristics, such as structural reforms, economic changes and responses to the impact of global events on their development.

The last 5-year period for clustering is from 2016 to 2020, the results of which are shown in Fig. 11.

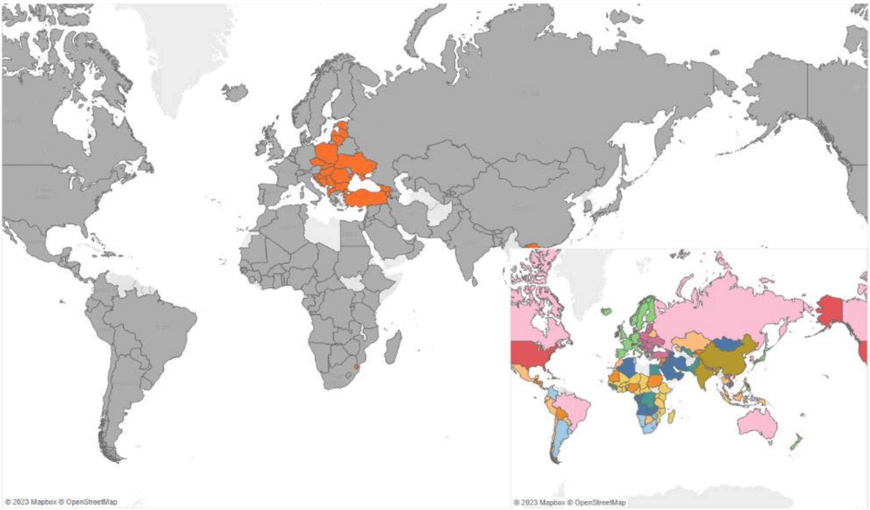


Fig. 11. Results of clustering countries based on data for the period 2016-2020

In the period from 2016 to 2020, Ukraine was in cluster that unites most of the countries of Eastern Europe, as can be seen in Fig. 11. This group of countries was characterized by significant economic development dynamics, which can be identified by several key factors, among them are rapid GDP growth compared to highly developed countries, low population growth or even depopulation, average inflation (lower than in underdeveloped countries but higher than in highly developed ones), rapid capital accumulation, predominance of the service sector in the structure of branches of economy, and high human capital. Economic growth in this cluster was driven by the implementation of economic reforms in many countries, including Ukraine. The introduction of structural changes and policy reforms contributed to improved economic efficiency and stability in the region. Ukraine, like the other countries in this cluster, demonstrated positive economic growth rates, which contributed to the recovery from the economic and political crisis and the beginning of the russian invasion in 2014.

Next, we combine the principal components for all 6 five-year periods and re-cluster the countries for the full period (90 factors in total).

From 1991 to 2020, Ukraine demonstrated close economic and political ties with the countries of Eastern Europe, most of the time belonging to the same cluster as most of these countries (see Fig. 12).

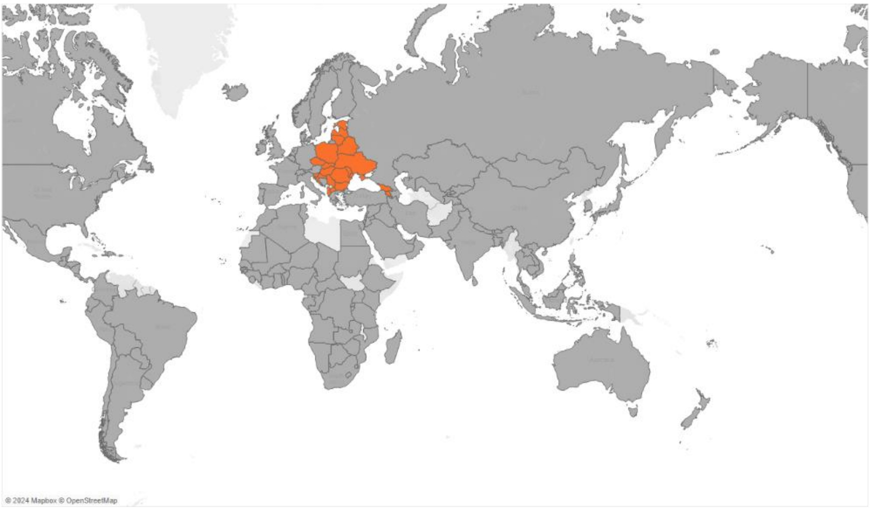


Fig. 12. Results of clustering countries based on data averaged over periods for 1991-2020 (Ukraine's cluster)

The main trends in the development of this cluster were determined by several key aspects that influenced the economic landscape of the region and marked the transition from a socialist to a market economy. During this period, the cluster countries actively implemented economic reforms aimed at liberalizing markets, privatizing state-owned enterprises and stimulating entrepreneurship.

These measures were intended to create a more flexible and competitive economic systems, facilitating the exit from the complex structures of socialist governance.

Another important aspect was the trend towards European integration. Many countries in this cluster, including Ukraine, were actively working to strengthen economic and political ties with the European Union. Integration with the EU included adapting to European standards, promoting mutual trade and economic ties. This process also meant increased cooperation with other European countries, which facilitated the exchange of experience and the formation of joint strategic initiatives to achieve stability and sustainable development in the region. Thus, European integration was a step towards deepening cooperation and rapprochement between the countries, which contributed to the formation of one economic and political space.

According to Fig. 13, a large number of clusters unite neighboring countries, which can be explained primarily by similar natural conditions, common historical events that have had an impact on

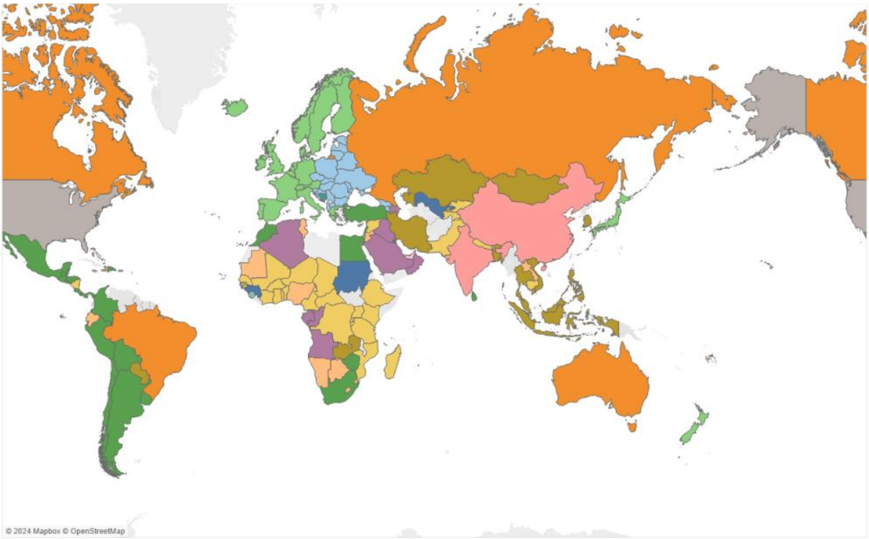


Fig. 13. Results of clustering countries based on data averaged over periods for 1991-2020

social and economic life, and the fact that through trade and cultural interpenetration, living standards in neighboring countries are gradually levelling off. Large countries often have large reserves of natural resources and a corresponding economic specialization (cluster 2 in Table 5), and a similar situation may be typical for neighboring countries that are located on large natural resource deposits (as in the oil-producing cluster 12/15/16 in Table 5).

A brief description of the clusters from Fig. 13 is given in Table 5, where some clusters are combined with each other. The need to group certain clusters of countries arises from the practical purposefulness and convenience of analysis in the study of international relations and economic processes. With a large number of clusters, a situation may arise when some countries form their own isolated clusters consisting of only one state. For the purpose of a comprehensive analysis and obtaining more representative information, it is important to group such clusters that demonstrate similar economic, political or socio-cultural characteristics.

Table 5

COMPOSITION OF COUNTRY CLUSTERS BASED ON AVERAGED 5-YEAR PERIODS FOR 1991-2020

№	List of countries in the clusters	Characteristics of clusters
1, 8	Albania, Armenia, Bosnia and Herzegovina, Belarus, Bulgaria, Croatia, Czech Republic, Estonia, Georgia, Hungary, Latvia, Lithuania, Malta, Moldova, North Macedonia, Poland, Romania, Serbia, Slovakia, Slovenia, Ukraine	Transition to a market economy, involvement in global markets, aspirations for European integration, strategy for infrastructure modernization and economic diversification. Some of the countries on this list are members of the European Union, while others are developing their economies in the context of geopolitical challenges, but general trends indicate that they are mutually adapting to the requirements of the modern global economic environment.
2	Australia, Brazil, Canada, Russia	United in the economic context by their vast territory and extensive natural resources, they are known for their significant role in the global market for the production of raw materials, including energy, agricultural products and metals.

3	Botswana, Cape Verde, Comoros, Ecuador, Eswatini, Fiji, Haiti, Jamaica, Jordan, Laos, Lesotho, Mauritania, Namibia, Nigeria, Seychelles, Tunisia	Highly dependent on agriculture and commodity imports many of these countries face socio-economic challenges such as poverty and income inequality. These countries are working on development and reforms to stimulate economic growth.
4, 14	Antigua and Barbuda, Argentina, Bahamas, Belize, Bolivia, Chile, Colombia, Costa Rica, Dominican Republic, Egypt, El Salvador, Guatemala, Honduras, Lebanon, Mauritius, Mexico, Morocco, Panama, Peru, South Africa, Sri Lanka, Turkey, Uruguay, Zimbabwe	Characterized by a diversity of production and significant contributions to their economies from such sectors as agriculture, tourism, industry and natural resource exports, the countries of these two clusters, the vast majority of which are from Latin America and the Middle East, interact in global markets and their economies have often been affected by global economic and trade trends.
5	Austria, Barbados, Belgium, Cyprus, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Japan, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom	The countries of this cluster have highly developed economies, characterized by high living standards, efficient socio-economic infrastructure and a developed sectoral division of labor. The basis of the economy of these countries is high technology, innovation, finance and trade. They actively cooperate internationally, having an important impact on the global economy.
6	Bangladesh, Bhutan, Indonesia, Iran, Kazakhstan, Malaysia, Mongolia, Paraguay, Philippines, Thailand, Vietnam, Zambia	Countries of this cluster are often defined as emerging or developing economies. They actively engaged in global markets with a focus on industrial production, export strategies, efforts to attract foreign investment and implement innovative programs to support economic growth.
0, 7, 9	Benin, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Republic, Chad, Democratic Republic of the Congo, Ethiopia, Gambia, Ghana, Guinea, Guinea-Bissau, Ivory Coast, Kenya, Kyrgyzstan, Madagascar, Mali, Mozambique, Nepal, Nicaragua, Niger, Pakistan, Rwanda, Senegal, Sierra Leone, Sudan, Syria, Tajikistan, Tanzania, Uganda, Uzbekistan	Emerging economies with common features such as low levels of industrialization, high dependence on agriculture, and problems with poverty and inequality.

10	Hong Kong, Luxembourg, Singapore, South Korea, Taiwan	Countries with highly developed, innovative and competitive economies, with a large role of financial services and high technology, efficient infrastructure, high level of education and sustainable development strategies.
11	China, India	Two of the world's most populous countries, large domestic markets, high economic growth, a wide range of industries and a high level of infrastructure and technology investment.
13	United States of America	It is the largest economy in the world, characterized by a high level of industryalization, a developed technological base and a wide variety of industries and services. A leader in many key sectors, such as information technology, finance, science and innovation.
12, 15, 16	Algeria, Angola, Azerbaijan, Bahrain, Gabon, Iraq, Kuwait, Oman, Qatar, Saudi Arabia, United Arab Emirates	Possessing significant reserves of energy resources, in particular oil and gas, the countries of these clusters specialize in the production and export of energy, which makes them key players in the global energy markets.

To summarize the results of the clustering, it can be noted that in many respects neighboring countries share common development patterns, largely due to foreign trade and mutual integration. Exceptions may be countries that have their own unique characteristics in terms of natural resource endowments or the existing political regime in the country. The size of the labor force and the size of the country's territory are also important factors determining the position of countries in clusters.

Conclusions

Cluster analysis is a comprehensive and promising method for studying economic growth. In this area, it can be used to identify groups of countries or regions based on a set of factors, analyze economic diversification and regional inequality, select key factors of economic growth, optimize the allocation of production factors, assess the impact of economic policies, identify opportunities for international cooperation and more. In this article, the main objective was to group countries by identifying similar patterns in their development based on macroeconomic indicators, with a special focus on the Ukrainian economy.

An important component of the research was the development of the methodology, which included defining the problem area through a detailed study of literature sources, a qualitative analysis of clustering and dimensionality reduction methods that revealed the advantages and disadvantages of each, determining approaches to the selection of modeling parameters, data processing, their standardization and dimensionality reduction. The scientific novelty of the article is in the development and application of a new methodology for quantifying the results of cluster analysis and dimensionality reduction algorithms. It is proposed to assess clustering methods using an aggregate indicator based on normalized quality metrics, such as the Davies-Bouldin index, the Calinski-Harabasz index, and the Silhouette coefficient, and to evaluate dimensionality reduction methods – use cumulative explained variance.

As a result of applying the methodology described in the article to the array of selected and processed data, it was determined that the most effective algorithm for dimensionality reduction is Kernel principal component analysis with sigmoid kernels, which allowed preserving 95.76% of the information while reducing the data set by 40%. Regarding the optimal clustering method, K-means with the division of the full data array into 17 clusters turned out to be the most effective, and Agglomerative clustering, BIRCH, and Ward's methods were also quite effective. It is worth noting that the methodology described in this article can be applied to other areas as well.

The cluster analysis was conducted on data for the period from 1991 to 2020, divided into 6 5-year intervals, and finally on consolidated data from all six intervals. Over this period, the Ukrainian economy migrated from the post-Soviet cluster, where it was in the first time interval from 1991 to 1995, to the cluster of Eastern European countries in the last period from 2016 to 2020. This can be explained by the real economic growth of Ukraine, gradual integration with the European Union, and deepening of foreign economic ties with neighboring European countries, despite regular political and economic crises. As for other clusters and other countries, it is worth noting that clusters often unite neighboring countries, which can be explained by similar natural conditions, common historical events that have affected socio-economic life, as well as foreign trade between neighboring countries. Also, the size of countries, the number of labor force, natural conditions and location on large deposits of natural resources often determine the economic specialization of a country.

As for further research, we plan to extend the experiment by adding new clustering and dimensionality reduction algorithms with different parameters, try to use other data normalization methods, increase the number of factors in the study, and modify the methodology so that we can correctly compare clustering algorithms with automatic cluster selection and algorithms with a given number of clusters. It is also planned to use the results of clustering to train models for forecasting economic growth in countries of separate clusters.

References

1. Solow, R. M. (1956). A Contribution to the Theory of Economic Growth. *The Quarterly Journal of Economics*, *70*(1), 65-94. https://doi.org/10.2307/1884513

2. Solow, R. M. (1957). Technical Change and the Aggregate Production Function. *The Review of Economics and Statistics*, *39*(3), 312-320. https://doi.org/10.2307/1926047

3. Swan, T. W. (1956). Economic Growth and Capital Accumulation. *Economic Record*, *32*(2), 334-361. https://doi.org/10.1111/j.1475-4932.1956.tb00434.x

4. Ramsey, F. P. (1928). A mathematical theory of saving. *The Economic Journal*, *38*(152), 543-559. https://doi.org/10.2307/2224098

5. Cass, D. (1965). Optimum Growth in an Aggregative Model of Capital Accumulation. *The Review of Economic Studies*, *32*(3), 233-240. https://doi.org/10.2307/2295827

6. Koopmans, T.C. (1963). On the Concept of Optimal Economic Growth. *Cowles Foundation Discussion Papers*, Article 163. https://elischolar.library.yale.edu/cowles-discussion-paper-series/392

7. Uzawa, H. (1965). Optimal Technical Change in an Aggregative Model of Economic Growth. *International Economic Review*, *6*(1), 18-31. https://doi.org/10.2307/2525621

8. Lucas, R. E. (1988). On the Mechanics of Economic Development. *Journal of Monetary Economics*, *22*(1), 3-42. https://doi.org/10.1016/0304-3932(88)90168-7

9. Mankiw, N. G., Romer, D., & Weil, D. N. (1992). A Contribution to the Empirics of Economic Growth. *The Quarterly Journal of Economics*, *107*(2), 407-437. https://doi.org/10.2307/2118477

10. Romer, P. M. (1989). *Human Capital and Growth: Theory and Evidence* (Working Paper No. 3173). National Bureau of Economic Research. https://doi.org/10.3386/w3173

11. Romer, P. M. (1989). *Endogenous Technological Change* (Working Paper No. 3210). National Bureau of Economic Research. https://doi.org/10.3386/w3210

12. Kitchin, J. (1923). Cycles and Trends in Economic Factors. *The Review of Economics and Statistics*, *5*(1), 10-16. https://doi.org/10.2307/1927031

13. Kuznets, S. (1960). Economic Growth of Small Nations. In E.A.G. Robinson (Ed.), *International Economic Association Series. Economic Consequences of the Size of Nations* (pp. 14-32). Palgrave Macmillan. https://doi.org/10.1007/978-1-349-15210-0_2

14. Korotayev, A. V., & Tsirel, S. V. (2010). A Spectral Analysis of World GDP Dynamics: Kondratieff Waves, Kuznets Swings, Juglar and Kitchin Cycles in Global Economic Development, and the 2008-2009 Economic Crisis. *Structure and Dynamics*, *4*(1), Article 3306. https://doi.org/10.5070/sd941003306

15. Strelchenko, I. (2019). Modeling of cross-border spreading of financial crisis. *Neuro-Fuzzy Modeling Techniques in Economics*, *8*, 147-174. http://doi.org/10.33111/nfmte.2019.147

16. Matviychuk, A., Strelchenko, I., Vashchaiev, S., & Velykoivanenko, H. (2019). Simulation of the Crisis Contagion Process Between Countries with Different Levels of Socio-Economic Development. *CEUR Workshop Proceedings*, *2393*(II), 485-496. http://ceur-ws.org/Vol-2393/paper_423.pdf

17. Lukianenko, D., & Strelchenko, I. (2021). Neuromodeling of features of crisis contagion on financial markets between countries with different levels of economic development. *Neuro-Fuzzy Modeling Techniques in Economics*, *10*, 136-163. http://doi.org/10.33111/nfmte.2021.136

18. Kriegel, H.-P., Schubert, E., & Zimek, A. (2017). The (black) art of runtime evaluation: Are we comparing algorithms or implementations? *Knowledge and Information Systems*, *52*(2), 341-378. https://doi.org/10.1007/s10115-016-1004-2

19. Zare, H., Shooshtari, P., Gupta, A., & Brinkman, R. R. (2010). Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics*, *11*, Article 403. https://doi.org/10.1186/1471-2105-11-403

20. Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, *58*(301), 236-244. https://doi.org/10.1080/01621459.1963.10500845

21. Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. *ACM SIGMOD Record*, *25*(2), 103-114. https://doi.org/10.1145/233269.233324

22. Nielsen, F. (2016). Hierarchical Clustering. In *Undergraduate Topics in Computer Science. Introduction to HPC with MPI for Data Science* (pp. 195-211). Springer. https://doi.org/10.1007/978-3-319-21903-5_8

23. Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, *3*(1), 72-83. https://doi.org/10.1109/89.365379

24. Frey, B. J., & Dueck, D. (2007). Clustering by Passing Messages Between Data Points. *Science*, *315*(5814), 972-976. https://doi.org/10.1126/science.1136800

25. Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(8), 790-799. https://doi.org/10.1109/34.400568

26. Ester, M., Kriegel, H.-P., 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* (pp. 226-231). AAAI. https://aaai.org/papers/kdd96-037-a-density-based-algorithm-for-discovering-clusters-in-large-spatial-databases-with-noise/

27. Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Transactions on Knowledge Discovery from Data*, *10*(1), Article 5. https://doi.org/10.1145/2733381

28. Ankerst, M., Breunig, M., Kriegel, H.-P., & Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. *ACM SIGMOD Record*, *28*(2), 49-60. http://dx.doi.org/10.1145/304181.304187

29. Deltuvaitė, V., & Sinevičienė, L. (2014). Investigation of Relationship between Financial and Economic Development in the EU Countries. *Procedia Economics and Finance*, *14*, 173-180. https://doi.org/10.1016/s2212-5671(14)00700-x

30. Enzmann, P., & Moesli, M. (2022). Seizing opportunities: ASEAN country cluster readiness in light of the fourth industrial revolution. *Asia and the Global Economy*, *2*(1), Article 100021. https://doi.org/10.1016/j.aglobe.2021.100021

31. Koutsoukis, N.-S. (2015). Global Political Economy Clusters: The World as Perceived through Black-box Data Analysis of Proxy Country Rankings and Indicators. *Procedia Economics and Finance*, *33*, 18-45. https://doi.org/10.1016/s2212-5671(15)01691-3

32. Peruzzi, M., & Terzi, A. (2021). Accelerating Economic Growth: The Science beneath the Art. *Economic Modelling*, *103*, Article 105593. https://doi.org/10.1016/j.econmod.2021.105593

33. Čadil, J., Petkovová, L., & Blatná, D. (2014). Human Capital, Economic Structure and Growth. *Procedia Economics and Finance*, *12*, 85-92. https://doi.org/10.1016/s2212-5671(14)00323-2

34. Cerqueti, R., & Ficcadenti, V. (2022). Combining rank-size and k-means for clustering countries over the COVID-19 new deaths per million. *Chaos, Solitons & Fractals*, *158*, Article 111975. https://doi.org/10.1016/j.chaos.2022.111975

35. Wulandari, L., & Yogantara, B. O. (2022). Algorithm Analysis of K-Means and Fuzzy C-Means for Clustering Countries Based on Economy and Health. *Faktor Exacta*, *15*(2), 109-116. https://doi.org/10.30998/faktorexacta.v15i2.12106

36. Jolliffe, I. (2002). *Principal Component Analysis* (2nd ed.). Springer New York. https://doi.org/10.1007/b98835

37. DeAngelis, G. C., Ohzawa, I., & Freeman, R. D. (1995). Receptive-field dynamics in the central visual pathways. *Trends in Neurosciences, 18*(10), 451-458. https://doi.org/10.1016/0166-2236(95)94496-r

38. Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '01)* (pp. 245-250). Association for Computing Machinery. https://doi.org/10.1145/502512.502546

39. Hyvärinen, A. (2013). Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371*(1984), Article 20110534. https://doi.org/10.1098/rsta.2011.0534

40. Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, *290*(5500), 2319-2323. https://doi.org/10.1126/science.290.5500.2319

41. Mead, A. (1992). Review of the Development of Multidimensional Scaling Methods. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *41*(1), 27-39. https://doi.org/10.2307/2348634

42. Roweis, S. T., & Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, *290*(5500), 2323-2326. https://doi.org/10.1126/science.290.5500.2323

43. Linderman, G. C., & Steinerberger, S. (2017). *Clustering with t-SNE, provably*. arXiv. https://doi.org/10.48550/ARXIV.1706.02582

44. Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224-227. https://doi.org/10.1109/tpami.1979.4766909

45. Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1-27. https://doi.org/10.1080/03610927408827101

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

47. Rahman, M. M., & Alam, K. (2021). Exploring the driving factors of economic growth in the world's largest economies. *Heliyon*, *7*(5), Article E07109. https://doi.org/10.1016/j.heliyon.2021.e07109

48. The World Bank. (2023). *World Development Indicators* [Data set]. Retrieved February 1, 2023, from https://databank.worldbank.org/source/world-development-indicators

49. Zhylinska, O., Bazhenova, O., Zatonatska, T., Dluhopolskyi, O., Bedianashvili, G., & Chornodid, I. (2020). Innovation processes and economic growth in the context of European integration. *Scientific Papers of the University of Pardubice, Series D: Faculty of Economics and Administration*, *28*(3), Article 1209. https://doi.org/10.46585/sp28031209

50. Scikit-learn. (n.d.). *2.3. Clustering*. Retrieved February 20, 2023, from https://scikit-learn.org/stable/modules/clustering.html

The article was submitted on 2023, March 08