

References

1. Узагальнені дані надання послуг перенесення номерів за 2024 рік (2024). Вилучено з <https://nkek.gov.ua/static-objects/nkek/sites/1/dani-mnp/dani-mnp-2024.pdf>
2. Роман Мирончук, (2026) Українці активно змінюють операторів: lifecell отримав найбільший приплив номерів. Вилучено з <https://minfin.com.ua/ua/2026/03/07/169520362/>
3. Kaggle. (n.d.). Telco customer churn [Data set]. <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>
4. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why tree-based models still outperform deep learning on tabular data. *Advances in Neural Information Processing Systems*, 35, 507–520. <https://arxiv.org/abs/2207.08815>
5. Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6(1), Article 28. <https://doi.org/10.1186/s40537-019-0191-6>

Matviychuk A. V.,

DSc (Economics), Professor,

Skisov V.,

First-cycle student (Bachelor's level),

Kyiv National Economic University named after Vadym Hetman

MODELLING THE CLUSTER STRUCTURE OF EU REGIONS USING ML METHODS: A PCA-SOM APPROACH

The study of regional heterogeneity in the European Union is an important direction of modern spatial econometrics. EU cohesion policy aims to reduce disparities between NUTS-2 regions; however, despite substantial financial transfers (EUR 392 billion in 2021-2027), regional inequality persists: the gap in GDP per capita (PPS) between the richest and poorest regions exceeds eightfold [1]. This underlines the need for effective methods of regional space typification capable of detecting hidden structural patterns in large time panels.

The aim of the study is to construct a robust typology of 191 EU NUTS-2 regions by combining Principal Component Analysis (PCA) with Kohonen Self-Organising Maps (SOM). The data source is Eurostat statistics for 2012-2023 [2] covering four dimensions of regional development: innovation potential, urbanisation structure, sectoral employment structure, and tourism specialisation.

The indicator system comprises 13 variables in four groups: innovation potential (HRST (Human Resources in Science and Technology) share in employment, hourly and per-person labour productivity, life expectancy at birth); urbanisation (shares of urban, suburban, and rural population, population density); employment structure (shares of IT and communications, science and technology, manufacturing, and education); and tourism activity (number of overnight stays in collective accommodation establishments). Data preparation included the exclusion of regions with a critical share of missing values (above 41%) and imputation of remaining gaps using a Random Forest algorithm with predictive mean matching (PMM, missRanger, seed = 123).

At the first stage, Principal Component Analysis was applied to each of the four indicator groups. A distinctive feature of the approach is the use of the Aitchison Centred Log-Ratio (CLR) transformation [3] for compositional variables (urbanisation and employment shares), which ensures geometric correctness when working with closed-system data. The sign of PC1 is anchored by a reference indicator (HRST for Ind_Innov; clr_cities for Ind_Urban; Employment_Science for Ind_Employ), guaranteeing meaningful interpretability of the indices [4].

Table 1

Variance explained by the first principal component across indicator groups

Group / Index	PC1 share, %	Cumulative (PC1+PC2), %	PC1 sign anchor
Ind_Innov	78.5	93.7	HRST
Ind_Urban	64.9	93.7	clr_cities
Ind_Employ	57.2	86.6	Employment_Science
Ind_Tourism	-	-	z-score(log Tourism)

Source: calculated by the author in R

For the Urban and Employ groups, where PC2 accounts for more than 28% of additional variance, the second component was also included in the SOM feature space. The neural network input space therefore consists of six dimensions: Ind_Innov, Ind_Urban (PC1), Urban_PC2, Ind_Employ (PC1), Employ_PC2, and Ind_Tourism.

At the second stage, the Kohonen Self-Organising Map was trained on the entire dataset of 2,292 observations (191 regions $\times$ 12 years) simultaneously. This ensures a single coordinate system for tracking regional trajectories over time. Map architecture: hexagonal grid $8 \times 9 = 72$ nodes (size determined by the formula $\text{units} = 5 \cdot \sqrt{N}$); learning rate α decreases linearly from 0.05 to 0.01 over 1,000 iterations. Clustering of the 72 map nodes was performed using k-means with 25 random starting points. The optimal value $k = 5$ was selected based on the elbow test, silhouette analysis, and verification using the AV integral criterion [5], where the k-means configuration ($k = 5$) achieved the highest AV value among all non-binary models.

Table 2

Profiles of five EU regional macro-clusters (mean Z-scores, 2013-2023)

Indicator	C1 Non-central	C2 Metropolis	C3 Tourist	C4 Catch-up	C5 Diversified
Ind_Innov	0.37	0.66	0.21	-1.51	0.55
Ind_Urban (PC1)	-0.41	4.35	-0.23	-0.41	0.2
Urban_PC2	-0.57	-1.55	-0.18	-0.40	0.9
Ind_Employ (PC1)	-0.13	1.79	0.36	-1.10	0.49
Employ_PC2	0.6	-0.09	1.1	-0.39	-0.57
Ind_Tourism	-0.43	0.25	1.28	-0.72	0.31
No. of regions (2023)	47	7	23	38	76

Source: calculated by the author in R

The resulting typology comprises five macro-clusters. Cluster 1 ("Moderately developed non-central regions", 47 regions) - a profile close to the average, with a moderate educational employment component and lower urbanisation (Sjælland, Nordjylland, Castilla y León). Cluster 2 ("Large urban cores", 7 regions) – a sharply pronounced urbanisation profile (Urban PC1 = +4.35), concentration of managerial and science-and-technology functions (Wien, Berlin, Hamburg, Praha, Budapest,

Brussels, Bremen). Cluster 3 ("Tourism and service regions", 23 regions) - the highest tourism index (+1.28) and educational employment component (+1.10), typical of coastal, island, and alpine territories (Illes Balears, Canarias, Tirol, Kriti, Jadranska Hrvatska). Cluster 4 ("Catch-up regions", 38 regions) - the lowest innovation scores (-1.51) and employment PC1 (-1.10), predominantly CEE regions (Małopolskie, Közép-Dunántúl, Nord-Vest, Latvija). Cluster 5 ("Developed diversified regions", 76 regions) - the largest cluster, combining a suburbanised spatial structure (Urban PC2 = +0.90) with a well-developed post-industrial base (Stuttgart, Oberbayern, Lombardia, Noord-Brabant).

Analysis of cluster structure dynamics over 2013-2023 reveals considerable stability: most regions retained their cluster membership throughout the study period. Cluster 5 gradually increased in size, indicating an expansion of the diversified-region group. The highest GRP growth rates are recorded in Cluster 4 (above 10% in peak years), confirming the presence of β -convergence among the catch-up regions of CEE. The deepest contraction in 2020 is observed in the tourism-and-service cluster (approximately -7%), reflecting the extreme sensitivity of these regions to pandemic-related restrictions.

The proposed methodological chain "PCA $\rightarrow$ CLR $\rightarrow$ SOM $\rightarrow$ k-means" combines the advantages of linear dimensionality reduction, geometrically correct analysis of compositional data, and non-linear neural-network topological mapping. Unlike standard k-means, the SOM preserves the topological structure of the multidimensional regional space, providing more stable clustering under conditions of high data dispersion. The use of a global map trained on the full panel dataset enables analysis of inter-cluster regional mobility in the time dimension, which is a key advantage for assessing structural transformations under external shocks.

References

1. European Commission. (2022). *Cohesion in Europe towards 2050: Eighth report on economic, social and territorial cohesion*. Publications Office of the EU. https://ec.europa.eu/regional_policy/sources/reports/cohesion8/8cr.pdf
2. Eurostat. (n.d.). *NUTS - Nomenclature of territorial units for statistics*. European Commission. <https://ec.europa.eu/eurostat/web/nuts/>
3. Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall.
4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer. <https://doi.org/10.1007/978-0-387-84858-7>
5. Poznyak, S., & Kolyada, Yu. (2023). Comparative analysis of the effectiveness of dimensionality reduction algorithms and clustering methods on the problem of modelling economic growth. *Neuro-Fuzzy Modeling Techniques in Economics*, 12, 67-110. <https://doi.org/10.33111/nfmte.2023.067>

Savina S. S.,
PhD in Economics, Associate Professor,
Kyiv National Economic University named after Vadym Hetman

CLUSTER ANALYSIS OF THE COUNTRIES OF THE WORLD BY HUMAN DEVELOPMENT INDICATORS

Assessing the quality of life of the population is an important area of modern socio-economic research, as it makes it possible to analyze the effectiveness of state social policy, the level of access to education, healthcare, and other components of public welfare. The use of modern information technologies and digital data processing methods creates opportunities for identifying current social problems and developing effective strategies for societal development. One of the key indicators for assessing quality of life is the level of human development, which characterizes the overall socio-economic progress of a country.