

Uncovering the Underlying Patterns of Migration Drivers in the European Union: A Hierarchical Clustering Approach

Maria Berta BELU*

Bucharest University of Economic Studies, Bucharest, Romania

National Scientific Research Institute for Labour and Social Protection, Bucharest, Romania

**Corresponding author, belumaria19@stud.ase.ro*

Mădălina Ecaterina POPESCU

Bucharest University of Economic Studies, Bucharest, Romania

National Scientific Research Institute for Labour and Social Protection, Bucharest, Romania

madalina.andreica@csie.ase.ro

Abstract. *Labour migration has become a crucial strategy to address labour shortages across the European Union. While labour markets are an essential mechanism triggering migration, other factors may play a role in an individual's departure. This paper aims to investigate the underlying patterns of labour migration among the European Union states for the year 2024, using data from Eurostat and The Heritage Foundation. The variables were grouped by distinguishing between predisposing, proximate, precipitating and mediating factors of migration. As novelty, the dataset was extended compared to previous literature, by adding digitalization and other specific labour market and demographic factors into the analysis. Then, a principal component analysis was conducted to tackle the multi-dimensionality problem using RStudio. The selected principal components encompass diverse dimensions such as economic development and challenges, institutional performance, social outcomes and vulnerability, demographic and educational dynamics, digitalization performance and migration openness. A hierarchical cluster algorithm was finally developed to investigate patterns of migration drivers among six distinct EU country-clusters. These findings could enrich the existing literature on labour migration patterns in the EU, through adding new perspectives regarding influencing factors.*

Keywords: labour migration, drivers of migration, principal component analysis, Ward's agglomerative hierarchical clustering, European Union.

Introduction

In the persistent impact of the COVID-19 pandemic, labour migration has become a key strategy to address labour shortages driven by difficulties in recruitment, demographic changes and economic transformations across the European Union (EU). Since 2021, a series of extensive legal and policy updates have been implemented, including digitalization and fast track action plans to help harmonize migration with labour market demands (European Migration Network, 2025).

While labour markets represent one of the most influential mechanisms that triggers migration, the relocation decision cannot be solely attributed to isolated individual factors, but to an interconnected series of stimuli. According to Stark and Bloom (1985), apart from the difference in salaries between countries or regions, family or group influence, income maximization, adequate recognition of skills and the presence of other migrants at destination may also play a significant role in one's departure. Further, more recent labour migration theories add business opportunities (Massey & Parrado, 1998), trade openness, education, infrastructure and access to information (de Haas, 2007) to the previous mentioned migration determinants. More recent studies show that increasing education and income levels, which represent important and necessary sources for

individuals who decide to migrate, encourages the mobility of population (Dao et al., 2016). As well as this, extended access to diverse information and social networks fundamentally modulates one's desire to move, whereas macro-economic policies usually determine the actual aptitude to do so (Carling & Schewel, 2018).

Concerning the relevance of this topic, the main research question (RQ) of the paper consists of the following: *What are the underlying patterns of labour market migration among the EU states?*. To answer this RQ, a principal component analysis (PCA) is first conducted so to extract most of the information contained in a set of relevant migration factors considered. Next, a hierarchical cluster analysis is applied to identify country-groups with similar characteristics.

The novelty of the paper consists mainly in the integration of richer explanatory variables as compared to previous literature, by adding digitalization and other specific labour market and demographic factors into the investigation of the main migration patterns on the labour market in Europe.

The paper is organized as follows: Section 2 represents a brief literature review on the topic regarding different macro-econometric approaches on migration drivers, while Section 3 summarizes the methodological aspects of this research, specifically the data sources, the selection of variables and the application of the PCA and hierarchical clustering. Section 4 discusses the main findings of the analysis, and ultimately, the last section concludes, by also mentioning the limitations and future directions.

Literature review

Empirical research on migration drivers in the EU focuses relatively on macro-econometric approach, most often gravity-type or push-pull models, in which bilateral migration flows were regressed on a limited set of variables. Core explanatory variables typically included GDP per capita or wage gaps, unemployment differentials, geographic distance, common language, colonial ties, and shared borders (Mayda, 2010). Panel-data estimations confirmed the centrality of income and labor-market conditions as robust predictors of intra-EU mobility, while geographic and cultural proximity proxies captured migration costs and informational frictions (Afonso et al., 2025).

Over the past decade, however, both the conceptualization of migration drivers and the methodological toolkit have expanded considerably. First, empirical strategies have diversified beyond static gravity models to include dynamic panel models, instrumental-variable approaches addressing endogeneity between migration and economic performance, and micro-level analyses using survey or administrative data (Åslund, 2005). Second, scholars have incorporated institutional and policy variables, such as welfare-state generosity, recognition of qualifications, and labor-market regulation, into multi-variable frameworks, often combining macroeconomic indicators with meso-level network effects and micro-level characteristics (De Giorgi & Pellizzari, 2009). This shift reflects a broader theoretical move toward understanding migration as the outcome of interacting aspirations, capabilities, and structural constraints rather than a response to isolated economic signals (de Haas, 2021). Van Hear et al. (2012) proposed a broad classification of migration drivers and distinguished between predisposing, proximate, precipitating and mediating factors of migration, which was later followed by the study of Manafi and Roman (2022).

Importantly, methodological evolution also includes data-driven classification and dimensionality-reduction techniques, such as PCA and cluster analysis (Hausemer et al., 2019). These approaches are used to group countries or regions according to multidimensional driver profiles, to identify typologies of migration regimes, or to detect patterns in combinations of

economic, demographic, institutional, and socio-political factors (Kumpikaitė-Valiūnienė et al., 2025; Rosińska, 2024; Manafi & Roman, 2022).

Despite this broadening, a key gap remains, as more triggers of migration need to be taken into consideration when investigating the main migration patterns on the labour market in Europe, especially in the digital area. Thus, the explanatory variables list was enriched as compared to previous literature, by adding digitalization and other specific labour market and demographic factors into the analysis. Then PCA was conducted to tackle such multi-dimensional problems and plug the resulting principal components into the hierarchical cluster algorithm to further investigate the patterns of migration drivers in the EU.

Methodology

Given the main purpose of the paper, firstly, the data was gathered from Eurostat and The Heritage Foundation for the year 2024. The selection of the 18 variables (see Appendix, Table A1) follows the approach of Van Hear et al. (2012), by considering four main types of migration drivers: (1) predisposing factors of migration that create the high-probability context in which migration may happen; (2) proximate factors that influence migration more directly but derive from the structural features of the previous group; (3) precipitating factors as main triggers of migration; as well as (4) mediating factors. The criteria for categorizing the factors are based on their function in the migration process.

Therefore, as predisposing factors of migration which reflect the structural inequalities between the origin and the destination country, GDP per capita (*gdp_cap*), labor market patterns, income inequalities and demographic factors were selected. As proximate factors that comprise particular events related to economic, political, human security or environmental disparities, both in the origin and the destination place, the Economic freedom index (*econ_freedom*) and the Corruption perception index (*corruption*) were chosen. Next, as precipitating factors that are more observable than measurable, and more usually linked to the origin country, and less to the destination (Van Hear et al., 2012), life expectancy at birth (*life_exp*), poverty rate (*pov_soc_excl*) and the percentage of young persons neither in employment nor in education and training (NEET) were included. Finally, as mediating factors, which accelerate or restrict migration by mainly influencing transportation, communication and information infrastructure, indicators of digitalization and migration were considered.

Because of the high number of indicators taken into consideration, a PCA was next applied, followed by a hierarchical clustering analysis using RStudio.

Since the chosen variables had distinct measurement units and PCA is sensitive to the combination of variables with distinct units as it is computed from the covariance matrix of indicators (Jolliffe & Cadima, 2016; Greenacre et al., 2023), the variables were first standardized before applying PCA.

For the standardization of variables, the z-score normalization was implemented. By subtracting its mean, each variable is centered, then it is divided by its standard deviation. As a result, a set of variables with mean equal to 0 and standard deviation equal to 1 is obtained (Greenacre et al., 2023). The formula for z-score is, as follows (Jolliffe & Cadima, 2016):

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

where x_{ij} is the initial value of a variable, $\bar{x}_j$ is the mean, and s_j is the standard deviation. The z-scores are then introduced into the PCA.

PCA is a multivariate statistical technique most applied to reduce the data set size by extracting only the most important information from the data table (Abdi & Williams, 2010). The selection of several components can be made depending on a few methods. One of the most common methods is based on eigenvalues. For the correlation PCA, the components can be selected only if they present eigenvalues greater than 1 (Abdi & Williams, 2010). Another method of selection is represented by the “elbow” rule. The components are illustrated on a bar plot, called scree plot in this case, each bar corresponding to their explained variance. The decision on how many components should be considered is made by searching for the “elbow” in the bar series, indicated by a sharp change of slope (Greenacre et al., 2023). However, this method tends to be subjective since it is only based on visual elements, as Abdi and Williams (2010) state in their study.

In general, a clustering analysis is a classification scheme for organizing data in small groups of objects, presenting some patterns of similarities or differences in characteristics (Everitt et al., 2011). There are non-hierarchical and hierarchical methods. Focusing on hierarchical methods, these imply that the partitions are produced in a nested sequence rather than assigning observations to a given number of clusters in a single step. Hierarchical methods can be agglomerative, starting with a few observations that are merged into larger groups, or divisive, beginning with multiple observations that are split into smaller groups (Everitt et al., 2011).

For the clustering analysis, Ward’s agglomerative hierarchical clustering method was employed. Invented by Ward (1963), this method focuses on improving the sum of squares within clusters at each fusion step (Everitt et al., 2011) and presents two different implementations, Ward1 and Ward2 (Murtagh & Legendre, 2024). In this study, the second implementation was used for cluster analysis due to the fact that it is based more on variance rather than distance. Also, compared to Ward1, this version performs the computations correctly by squaring the distances to minimize the within-cluster variance.

Ward2 uses Euclidean distance for the input dissimilarities. In this case, the function to be minimized is given by the following formula (Murtagh & Legendre, 2014):

$$D(c_1, c_2) = \delta^2(c_1, c_2) = \frac{|c_1||c_2|}{|c_1| + |c_2|} \|c_1 - c_2\|^2$$

where c_1, c_2 are the centroids of two clusters, $|c_1|, |c_2|$ are the sizes of clusters, and $\|c_1 - c_2\|^2$ is the squared Euclidean distance between the centroids of clusters. The fraction $\frac{|c_1||c_2|}{|c_1| + |c_2|}$ represents the weight by cluster size for the distance, meaning that the increase in the sum of squares within clusters is dependent not only on distance, but also on the number of observations in each cluster.

For the optimal number of clusters, a simulation was applied, using the Calinski-Harabasz criterion. The formula for the criterion is the following (Cengizler & Kerem Un, 2017):

$$CH(k) = \frac{B_c(k)}{(k-1)} \bigg/ \frac{W_c(k)}{(n-1)}$$

where n is the number of clusters, k is the class, while B_c and W_c stand for between and within clusters sums of squares. This criterion evaluates each cluster solution depending on inter-cluster and intra-cluster distances. The higher the criterion, the better the solution.

Finally, Ward’s clustering method was applied to the principal components scores obtained through the correlation PCA, while the Euclidean distances between centroids were computed in the reduced-dimensional PCA space.

Results and discussions

Descriptive statistics and correlation analysis

The selected data was first investigated through descriptive statistics (see Appendix, Table A2) and correlation analysis (Figure 1). For life expectancy in Ireland and old age dependency ratio in Greece, the values were forecasted for 2024 using a linear forecast based on the period 2015-2023, due to missing data.

Noticeable differences between the minimum and maximum values are observed for a few indicators, such as *gdp_cap*, *lab_cost*, *corruption*, *neet*, *tax_rate*, *digit_pub_serv_cit*, *asyl_appl* and *net_migr_rate*, pointing to a substantive heterogeneity between the EU-27 countries. The highest *gdp_cap* value is in Luxembourg (127,030 euro), while top-down is in Bulgaria (with only 16,260 euro as compared to the EU average of 41,548 euro). The *lab_cost* varied across the EU from 10.6 euro in Bulgaria to 55.2 euro in Luxembourg, with an EU average of 28.44 euro. The lowest *tax_rate* applied to gross wage earnings was recorded in Slovakia (11.11%), while the highest in Belgium (34.2%) as compared to the EU average of 23.98%. Netherlands had the lowest percentage of *neet* (4.9%), as compared to Romania's highest percentage of 19.4%, that is almost double as the EU average (10.5%).

Regarding *corruption*, the highest level was recorded in Hungary, and the lowest in Denmark, while on average, the value of 62.33 highlights a moderate perception of corruption in the EU in 2024. The variable *digit_pub_serv_cit* performed worst in Romania, scoring 62.7 in comparison to the EU average score of 82.32, and best in Malta, with a score of 99.7.

Regarding *asyl_appl*, Hungary recorded the minimum of only 5 applicants, while on top was Germany with 103,600 applicants, compared to the EU average of 16,994. As for *net_migr_rate*, Latvia had the lowest net rate of -2.5, meaning that more people emigrated rather than immigrated, whereas Malta had the highest net rate of 18.7, receiving more immigrants than any other country per 1,000 persons of average population. The EU average rate was 6.18.

The analysis of correlation between the 18 indicators (Figure 1) generally brought to the fore moderate correlations in magnitude, most of them being situated between 0 and 0.6 in absolute values. It is observed that the strongest correlations were positive, between 0.62 and 0.82.

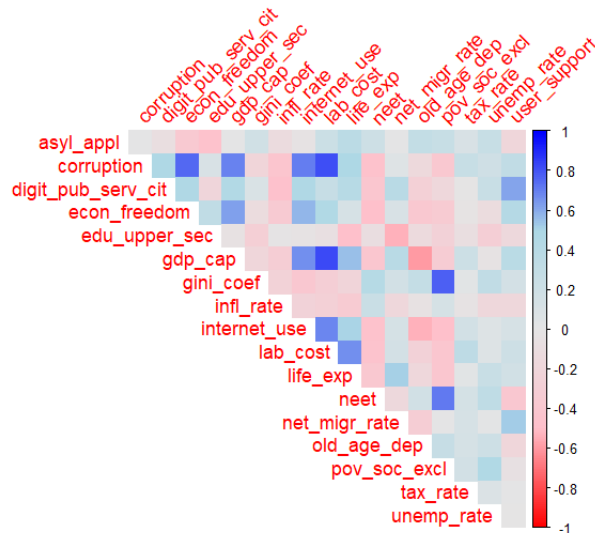


Figure 1. Correlation plot

Source: Authors' own research.

Of all variables, *lab_cost* recorded the highest correlation (0.82) with *corruption* and *gdp_cap*. Other positive strong correlations were found between *gini_coef* and *pov_soc_excl* (0.78), *corruption* and *econ_freedom* (0.76), *corruption* and *internet_use* (0.71), *pov_soc_excl* and *neet* (0.71), and *corruption* and *gdp_cap* (0.7). On the other hand, *gdp_cap* and *old_age_dep* had the greatest negative correlation (-0.59).

Principal component analysis

The PCA was applied to the scaled variables to retain the most relevant information of the selected variables. Looking at the eigenvalues (Table 1), the first six components presented eigenvalues greater than 1, with a cumulative variance of 82.7%.

Table 1. Eigenvalues and variances for principal components

Principal component	Eigenvalue	Variance explained	Cumulative variance
PC1	5.9203	32.89%	32.89%
PC2	3.0452	16.92%	49.81%
PC3	1.9479	10.82%	60.63%
PC4	1.6645	9.25%	69.88%
PC5	1.2899	7.17%	77.05%
PC6	1.0162	5.65%	82.7%
PC7	0.7799	4.33%	87.03%

Source: Authors' own research.

By comparison with the eigenvalue method, the elbow rule represented through the scree plot (Figure 2) indicated an optimal number of three components, with a cumulative variance of 60.63%, as visually indicated by the sharp change of slope. Starting with the fourth component, the slope becomes flatter.

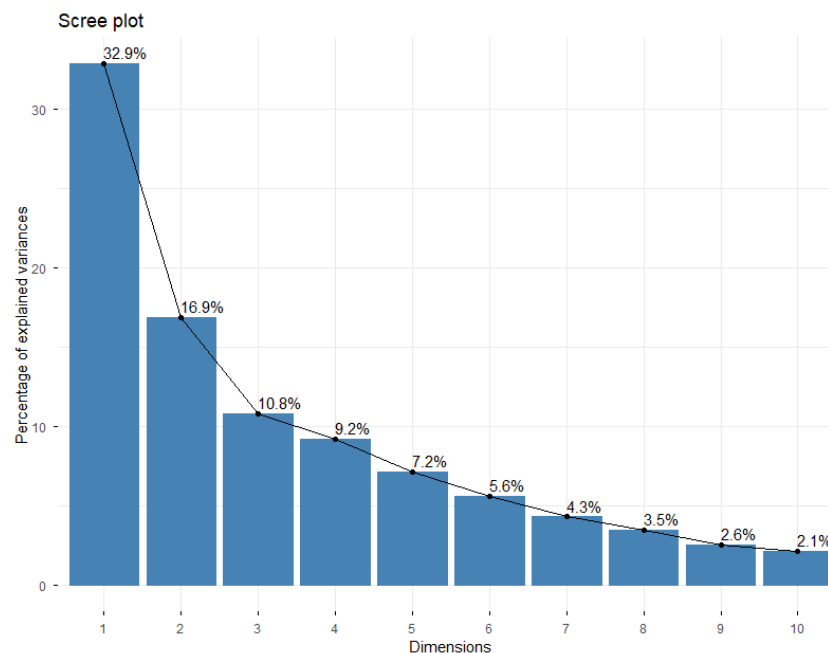


Figure 2. Scree plot for the explained variance of components

Source: Authors' own research.

Taking into consideration all the above-mentioned, the first five components were chosen to be used in the next step of the research, with a cumulative variance of 77.05%. Although the scree plot showed only the first three, this method is prone to subjectivity, not being backed up by a statistical test or a cut-off value. On the other hand, given the context of analyzing macro and institutional indicators for the EU-27 and the explained variance of 5.65%, the sixth component was dropped even though the eigenvalue was slightly over 1. In their study, Saccenti and Camacho (2015) supported the idea that the optimal number of principal components should depend more on the interpretability and the use of the PCA model rather than rigid methods.

The variables with over-average contributions for each of the five principal components are presented in Table 2 from the highest to the lowest contribution.

Table 2. Variables with over-average contributions for each of the first five components

Principal component	Variables with contributions over average
PC1: Economic development and institutional performance	gdp_cap, corruption, internet_use, lab_cost, econ_freedom, neet, life_exp, digit_pub_serv_cit, pov_soc_excl
PC2: Labour market and socio-demographic outcomes	edu_upper_sec, unemp_rate, gini_coef, asyl_appl, pov_soc_excl, net_migr_rate, life_exp
PC3: Digitalization and migration	user_support, asyl_appl, gini_coef, digit_pub_serv_cit, lab_cost, net_migr_rate, life_exp
PC4: Demographic, educational and institutional dynamics	net_migr_rate, edu_upper_sec, corruption, econ_freedom, life_exp, infl_rate, unemp_rate
PC5: Social vulnerability and economic challenges	old_age_dep, tax_rate, infl_rate, neet, gdp_cap, pov_soc_excl

Source: Authors' own research.

The first component was marked by the contributions of nine variables which are primarily related to economic development and institutional capacity regarding corruption and digitalization. The second component showed a distinct structure, with seven indicators mainly describing educational, labour market and income inequality phenomena. The third component contained a mixed information of digitalization, demography and income inequality, while the fourth one was more composed of demography, education and institutional perception. The last component, with only six variables with over-average contributions, principally focused on demography, taxation and inflation.

Hierarchical clustering

For hierarchical clustering, Ward's second agglomerative method was implemented. The five principal components were included in the process of clustering. For the selection of clusters, a simulation was carried out to choose an optimal number between 2 and 6, calculating the Calinski-Harabasz criterion. The results are written in Table 3.

Table 3. Optimal number of clusters using Calinski-Harabasz criterion

Number of clusters	Calinski-Harabasz criterion
2	8.872
3	8.257
4	8.839
5	8.518
6	8.911

Source: Authors' own research.

The highest value of the criterion was found for an optimal solution of six clusters, with a value of 8.911. Other solutions that had notably close criterion values to the optimal were the ones with two or four clusters. Further, the EU-27 countries were separated into six clusters, being represented in a dendrogram (Figure 3), as well as in a reduced-dimensional PCA space (Figure 4).

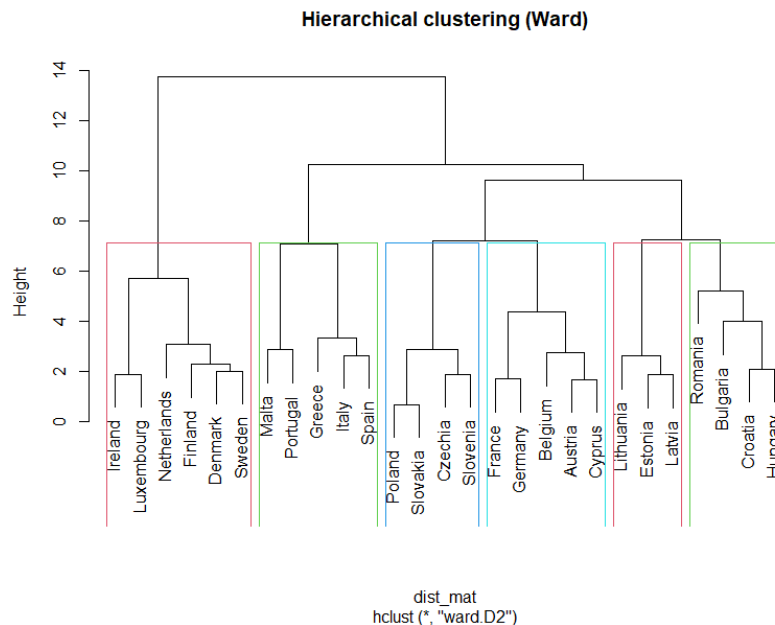


Figure 3. Dendrogram for the hierarchical clustering of the EU-27 countries

Source: Authors' own research.

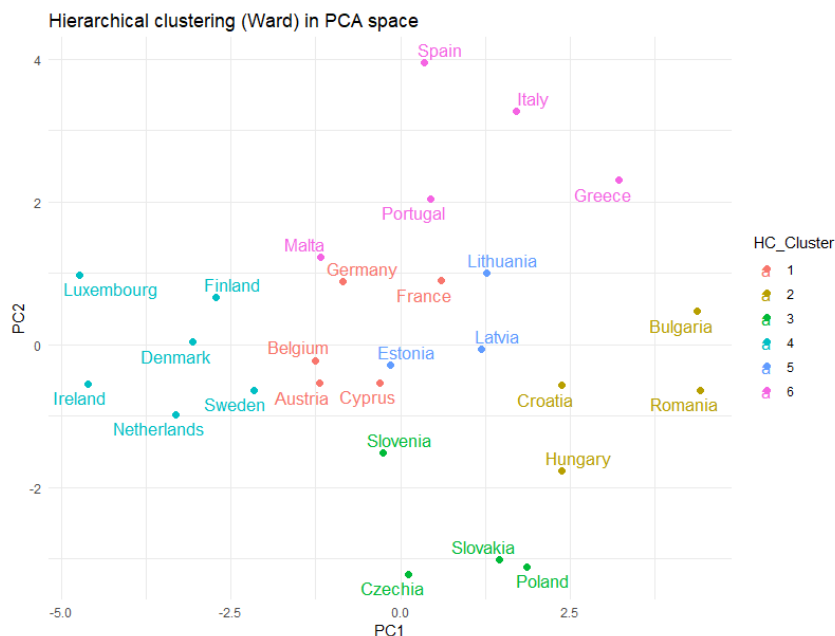


Figure 4. Hierarchical clustering of the EU-27 countries in PCA space

Source: Authors' own research.

The countries in Cluster 1 (Core states) are Austria, Belgium, Cyprus, France and Germany. These are best described by a very high rate of individuals who weekly use the internet for different purposes (92%-94.9%), relatively low income inequality among their citizens (the Gini coefficient taking values between 24.6 and 30.1), a high life expectancy at birth (81.5-83.2 years), and a high percentage of persons with at least upper secondary education (82.8%-86.2%). On the other hand, compared to the mean, the states in this cluster recorded moderate rates of persons who were at risk of poverty or social exclusion (16.9%-21.1%), as well as moderate to high old age dependency ratios (26.5-35.2). Moreover, the performances of the digital public services were below the EU-27 average (71.2-81.4), while the economic freedom index presented values between 62.5-72.2, highlighting a limited to moderate government interference.

Cluster 2 is composed of four Eastern European countries, respectively Bulgaria, Croatia, Hungary and Romania. These countries are characterized by low economic development, having GDP per capita between 16,260 and 22,200 euro, and a moderate government interference in the economy. The perception of corruption in the public sector is amongst the highest, the index having values between 41 and 47. Regarding labour market, the hourly labour costs are small, between 10.6 and 16.5 euro. Unemployment is situated below the EU average level, with rates between 4.2% and 5.4%. However, the rates of young people not in employment nor education recorded over-average levels (10.6%-19.4%). Life expectancy presents lower levels compared to other EU countries, people being expected to live between 75.9 and 79.1 years. Considering digitalization of public services, their performances are regarded as of moderate quality, with a score between 62.7 and 77.7.

Cluster 3 (Emerging economies) encompasses four countries: Czechia, Poland, Slovakia and Slovenia. The economic development of these states was better than the countries in Cluster 2 but still remaining below the EU average. As well as Cluster 2, they present a mild government economic interference. Income inequality is characterized by below-mean values between 21.7 and 26. People from these countries have relatively high internet use, 87.7% to 92.2% using it at least once per week. Another important feature of this cluster is the very high rates of persons with at least upper secondary education, between 89% and 94.8%. In general, people live between 78.6 and 82.3 years, with Slovenia recording a life expectancy over the EU mean.

Countries in Cluster 4 (Top performers: Denmark, Finland, Ireland, Luxembourg, Netherlands, Sweden) are the most economically developed and free, with Luxembourg taking the first place at GDP per capita in the EU and Ireland having the highest economic freedom index. Except Netherlands, the inflation rates were below the average. Corruption perception is relatively low, with a score between 77 and 90. Income inequality is situated slightly around the mean value, while the rates of poverty risk and social exclusion are below average. People from these states extensively use the internet in everyday chores: between 95.3% and 99.2%. Also, apart from Sweden, online user support is available and very effective, scoring 100 in three countries belonging to the current group. Although the high economic development, these states present at least upper secondary education attainments below 90%. Regarding life expectancy, people may expect to live between 82 and 84.1 years.

Cluster 5 comprises the Baltic states (Estonia, Latvia, Lithuania). Even though the GDP per capita levels place these countries in the bottom part, economic freedom was over the EU average. Income inequality amongst citizens is above the mean, with values between 30.8 and 35.3. Internet use is relatively high, between 88.7% and 92.3% of individuals accessing it regularly. Digital public services and online user support present high-performance scores of over 87. Regarding demography, citizens are expected to live between 76.7 and 79.6 years, compared to the EU

average of 81.1 years, while old dependency ratio revolves around the mean. Poverty risk and social exclusion tend to become a social problem, with rates between 22.2% and 25.8%. Besides this, unemployment recorded values between 6.9% and 7.6%, slightly higher than the average of 5.8%, whereas young people not employed or in education rates being situated in the interval of 10.7% and 14.7% respectively. The Baltic countries have well-educated citizens, between 88.8% and 94.4% of population aged 25-64 holding at least upper secondary education degrees.

Cluster 6 is formed of five Mediterranean states (Greece, Italy, Malta, Portugal, Spain), characterized by relatively high corruption perception and income inequality levels above average. These countries present quite similar tax rates on gross wage earnings, between 21% and 25.8%. Apart from Italy, inflation rates revolved around the EU mean of 2.6%. Regarding digitalization, between 85.5% and 95% of population regularly use the internet. Individuals live slightly above the EU average life expectancy, respectively between 81.9 and 84.1 years.

Being part of Cluster 2, Romania was among the EU countries with the lowest GDP per capita and the highest public sector corruption perception in 2024. Concerning labour market, the hourly labour costs were almost the smallest, at just 12.5 euro. While unemployment rate was situated slightly below the average, at 5.4%, Romania recorded the highest NEET rate, with 19.4% of young people aged 15-29 not employed nor in education. The level of income inequality, expressed by the Gini coefficient, placed this country among the 50% with lower inequality. However, the tax level on gross wage earnings was just 0,9 percentage points below the maximum value recorded in 2024. Regarding demographic dynamics, life expectancy at birth is among the lowest, Romanian citizens being expected to reach 76.6 years. The digital public services and online user support performance was moderate, despite 90.4% of individuals accessing the internet weekly.

Considering the 2015-2023 period, Romania presented a fluctuating trend concerning immigration and emigration. Between 2015 and 2021, it recorded a higher number of emigrants rather than immigrants, but after 2021, the number of immigrants managed to exceed the one of emigrants, reaching over 320,000 immigrants in 2023. In comparison, the EU always had a higher immigration during the selected period, with a peak of almost 7 million immigrants in 2022.

The results are consistent with previous studies in the field. In this sense, Hausemer et al. (2019) used a set of socio-economic indicators and found that European regions may also be grouped into six clusters related to the migration trajectory of asylum seekers. The six clusters were: (1) strong economic growing regions; (2) highly innovative and growing regions; (3) attractive manufacturing regions; (4) EU average-growing, demographically balanced regions; (5) EU average-growing but increasingly depopulating regions; and (6) lagging and depopulating regions, which include the southern borders of Europe.

Conclusion

This paper aimed to investigate the primary patterns in labour market migration for the EU-27 countries in 2024. For this, a PCA was carried out, followed by hierarchical clustering. Both types of analysis allowed us to respond to the research question that was raised at the beginning of the study. First, through PCA, five components were identified, each having a specific theme according to the contributions of the variables. Then, these components were introduced into the Ward's agglomerative hierarchical clustering, resulting in six clusters of the EU-27 states.

Cluster 1 included countries with low income inequality, high life expectancy, high percentage of well-educated people, and very high rate of internet users. Cluster 2 sums up low economic developed countries, with high public sector corruption perception, low labour costs,

reduced life expectancy, and over-average levels of NEET. Cluster 3 gathered better economic developed countries than the previous, but still below average. This cluster's countries had high internet use, very high rates of well-educated citizens and below-mean income inequality values. Cluster 4 comprised countries with the highest economic development and freedom, low corruption perception, extensive internet use and online support and high life expectancy. Cluster 5 featured three states with low GDP per capita, but over-average economic freedom. Apart from this, internet use is relatively high, accompanied by superior performance of digital public services and support. However, these countries presented social and demographic issues regarding income inequality, poverty risk and social exclusion, unemployment and life expectancy. Finally, Cluster 6 was defined by high corruption perception and above-average income inequality. On the other hand, these countries had a moderate to low inflation rate revolving around the EU mean, high life expectancy and a relatively high percentage of citizens using the internet weekly.

As main limitations of the study, the analysis was carried out for a single year, portraying a static image of the drivers of migration. Additionally, more comprehensive research would be achieved by taking in more variables describing the migration context depending on availability and relevance. Future directions would comprise conducting a social network analysis of the migration flows and linking them to the factors influencing this phenomenon.

Acknowledgement

This paper was co-financed by The Bucharest University of Economic Studies during the PhD program.

References

- Abdi, H., & Williams, L.J. (2010). Principal component analysis. *WIREs Computational Statistics*, 2, 433-459.
- Afonso, A., Alves, J., & Beck, K. (2025). Drivers of migration flows in the European Union: earnings or unemployment?. *International Labour Review*, 164(2), 1-23.
- Åslund, O. (2005). Now and forever? Initial and subsequent location choices of immigrants. *Regional Science and Urban Economics*, 35(2), 141-165.
- Carling, J., & Schewel, K. (2018). Revisiting aspiration and ability in international migration. *Journal of Ethnic and Migration Studies*, 44(6), 945-963.
- Cengizler, C., & Kerem Un, M. (2017). Evaluation of Calinski-Harabasz Criterion as Fitness Measure for Genetic Algorithm Based Segmentation of Cervical Cell Nuclei. *British Journal of Mathematics & Computer Science*, 22(6), 1-13.
- Dao, T.H., Docquier, F., Parsons, C., & Peri, G. (2016). Migration and Development: Dissecting the Anatomy of the Mobility Transition. *IZA Discussion Papers No. 10272*, Institute for the Study of Labor (IZA), Bon. Retrieved from <https://www.econstor.eu/bitstream/10419/147958/1/dp10272.pdf>
- De Giorgi, G., & Pellizzari, M. (2009). Welfare migration in Europe. *Labour Economics*, 16(4), 353-363.
- de Haas, H. (2007). Turning the Tide? Why Development Will Not Stop Migration. *Development and Change*, 38(5), 819-841.
- de Haas, H. (2021). A theory of migration: the aspirations-capabilities framework. *Comparative Migration Studies*, 9, 8(2021).

- European Migration Network (2025). *Labour migration in times of labour shortages*. Retrieved from https://emn.ie/wp-content/uploads/2025/12/2024_EMN-OECD_study_labour_migration_in_times_of_labour_shortages.pdf
- Everitt, B.S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis* (5th ed.). West Sussex, United Kingdom: John Wiley & Sons Ltd.
- Greenacre, M., Groenen, P.J.F., Hastie, T., d'Enza, A.I., Markos, A., & Tuzhzhilina, E. (2023). Principal component analysis. *Economics Working Papers 1856*, Department of Economics and Business, Universitat Pompeu Fabra. Retrieved from <https://repositori.upf.edu/server/api/core/bitstreams/9a4459af-9a9a-4c96-8eae-d97731cbf0c6/content>
- Hausemer, P., Todaro, L., Bocci, M., Plasilova, I., Rabuel, L., Williams, R., Vicari, V., Samek, M., Melloni, E., Pesce, F., Dufruca, S., Maranzano, P., Patrizio, M., Vasilescu, C., Kraler, A., Wagner, M., Baumgartner, P., Üllen, S., Bilger, V., Hofmann, M., Miah, P.A., Ricol, T.G., Mijačić, D., Žarković, J., Kahanec, M., Ulceluse, M., Althoff, J., & Steenberghen, T. (2019). *MIGRARE – Impacts of Refugee Flows to Territorial Development in Europe*. ESPON, Luxembourg. Retrieved from https://archive.espon.eu/sites/default/files/attachments/MIGRARE_Final_Report.pdf
- Jolliffe, I.T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374, 20150202.
- Kumpikaitė-Valiūnienė, V., Žičkutė-Daugelavičienė, I., Glińska, E., Rollnik-Sadowska, E., & Mockaitis, A. I. (2025). Exploring pull migration drivers to the European Union: A systematic literature review and future research agenda. *International Journal of Intercultural Relations*, 109, 102296.
- Manafī, I., & Roman, M. (2022). A Way to Europe: New Refugees' Migration Patterns Revealed. *Sustainability*, 14(2), 748.
- Massey, D.S., & Parrado, E.A. (1998). International Migration and Business Formation in Mexico. *Social Science Quarterly*, 79(1), 1-20.
- Mayda, A.M. (2010). International migration: A panel data analysis of the determinants of bilateral flows. *Journal of Population Economics*, 23, 1249-1274.
- Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?. *Journal of Classification*, 31(3), 274-295.
- Murtagh, F., & Legendre, P. (2024). *Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm*. Retrieved from <https://arxiv.org/pdf/1111.6285>
- Rosińska, A. (2024). *Drivers of Migration: Dimensions and Indicators. A Policy-Oriented Literature Review*. Publications Office of the European Union. Retrieved from <https://data.europa.eu/doi/10.2760/5771973>
- Saccenti, E., & Camacho, J. (2015). Determining the number of components in principal components analysis: A comparison of statistical, crossvalidation and approximated methods. *Chemometrics and Intelligent Laboratory Systems*, 149, 99-116.
- Stark, O., & Bloom, D.E. (1985). The New Economics of Labor Migration. *The American Economic Review*, 75(2), 173-178.
- Van Hear, N., Bakewell, O., & Long, K. (2012). *Drivers of Migration*. The Institute of Development Studies and Partner Organisations. Retrieved from <https://hdl.handle.net/20.500.12413/14821>

Ward, J.H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301), 236-244.

Appendix

Table A1. Variables grouped according to Van Hear et al. (2012) methodology

Variable	Abbreviation	Description
Predisposing factors		
Gross Domestic Product per capita	gdp_cap	The total economic output produced per person on an economic territory, output which meets the final demand of the economy. It is measured as euro/person at current market prices.
Gini coefficient of equivalised disposable income	gini_coef	The relationship of cumulative shares of the population arranged according to the level of equivalized disposable income, to the cumulative share of the equivalized total disposable income received by them. It takes values from 0, meaning perfect equality, to 100, meaning maximum inequality.
Inflation rate	infl_rate	The change over a period of the prices of goods and services acquired by households. It is calculated as annual average rate of change, expressed as percentage.
Unemployment rate	unemp_rate	The percentage of the unemployed persons aged 15-74 in relation to the total labour force aged 15-74.
Labour cost	lab_cost	The main expenditure an employee must make in order to employ staff. It comprises wages and salaries, employers' social security contributions and employment taxes, excluding any subsidies received. It is calculated as an average cost per hour worked, expressed in euro.
Persons aged 25-64 with at least upper secondary educational attainment level	edu_upper_sec	The percentage of persons aged 25-64 who have successfully completed at least upper secondary education.
Tax rate	tax_rate	The income tax on gross wage earnings including the employee's social security contributions, but excluding universal cash benefits, expressed as percentage of gross wage earnings, for two-earner couples with two children.
Old age dependency ratio	old_age_dep	The ratio between persons aged 65+ and persons aged 15-64.
Proximate factors		
Economic freedom index	econ_freedom	A measure of the liberty and free markets impact, comprising 12 freedoms, from property rights to financial freedom. It takes values from 0 (repressive government interference) to 100 (no government interference).
Corruption perception index	corruption	A composite index based on 13 different sources and scores, ranking countries according to people's perception of how corrupt the public sector is. A score of 0 represents a very high level of corruption, while 100 means a very low level.
Precipitating factors		
Life expectancy at birth	life_exp	The mean number of years a newborn is expected to live, given the current mortality conditions.
Persons at risk of poverty or social exclusion	pov_soc_excl	The percentage of persons living in households experiencing at least one of the following: poverty risk, severe material and social deprivation, or very low work intensity. It is calculated in relation to the total population.

Variable	Abbreviation	Description
Young persons neither in employment nor in education and training	neet	The percentage of persons aged 15-29 who are neither employed nor in a formal or non-formal education or training. It is calculated in relation to the total population aged 15-29.
Mediating factors		
Digital public services for citizens	digit_pub_serv_cit	The online provision of key public services for citizens, measured as the share of administrative steps that can be completed fully online for major life events. A score of 0 means a very low digital performance, while 100 a very high one.
Internet use	internet_use	The percentage of individuals aged 16-74 who use the internet at least once a week in relation to the total population aged 16-74.
User support	user_support	The extent to which online support, help features, and feedback mechanisms are available to both national as well as cross-border users. A score of 0 means a very low digital performance, while 100 a very high one.
Asylum applicants	asyl_appl	The number of third-country national or stateless persons having applied for international protection for the first time.
Net migration rate	net_migr_rate	The difference between total population change and natural change, plus statistical adjustment, expressed per 1,000 persons of the average population. A positive value indicates immigration, while a negative one emigration.

Source: Authors' own grouping based on Eurostat and The Heritage Foundation data.

Table A2. Descriptive statistics of the unscaled data

Variable	Min.	Median	Mean	Max.
gdp_cap	16,260	32,630	41,548	127,030
gini_coef	21.7	29.5	29.07	38.4
infl_rate	0.9	2.6	2.63	5.8
unemp_rate	2.6	5.4	5.81	11.4
lab_cost	10.6	21	28.44	55.2
edu_upper_sec	61.5	86.2	84.21	94.8
econ_freedom	55.1	68.5	69.51	82.6
corruption	41	59	62.33	90
life_exp	75.9	82.3	81.06	84.1
pov_soc_excl	11.3	19.7	20.2	30.3
neet	4.9	10	10.5	19.4
tax_rate	11.11	24.05	23.98	34.2
old_age_dep	21.7	32.2	32.25	38.4
digit_pub_serv_cit	62.75	81.43	82.32	99.7
internet_use	81.92	92.19	91.79	99.22
user_support	70.63	88.89	88.75	100
asyl_appl	5	2,865	16,994	103,600
net_migr_rate	-2.5	5.4	6.18	18.7

Source: Authors' own research.