EXPLANING VARIABILITY IN THE GLOBAL LABOUR MARKET USING PRINCIPAL COMPONENT ANALYSIS AND MULTIPLE IMPUTATION

1 Milica Marić, Faculty of Economics, University of Banja Luka, Banja Luka, Bosnia and Herzegovina
*Corresponding author's e-mail: milica.maric@ef.unibl.org
1 ORCID ID: 0009-0002-7228-2951

ARTICLE INFO

Review Scientific Paper
Received: 06.12.2024
Revised: 28.04.2025
Accepted: 28.05.2025
doi:10.63356/ace.2025.005
UDK
331.52:330.101.541(4)

Keywords: *labour market, variability, principal component analysis.*

JEL Classification: J01, J2, C82

ABSTRACT

This article investigates the underlying factors contributing to global labour market variability by applying Principal Component Analysis (PCA) and cluster analysis to data from 191 countries. Focusing on a broad set of economic, demographic and institutional indicators, the article seeks to uncover the primary dimensions shaping labour market dynamics worldwide. Key variables include GDP per capita, the Human Development Index (HDI), unemployment and poverty rates, labour freedom and corruption perception indices, average wages, and demographic characteristics such as population structure and migration rates. To ensure data completeness and robustness, multiple imputation was employed to address missing values. PCA was then used to reduce dimensionality and identify latent structures within the data. The resulting principal components were subsequently used in k-means clustering, which revealed four distinct clusters of countries sharing similar labour market profiles. The findings confirm that economic development and institutional quality are the dominant forces behind variations in labour market conditions across countries, while demographic variables, such as age distribution and migration, also play a meaningful role. These results support the hypothesis that clusters of countries with similar labour market profiles can be identified using the economic, demographic and institutional variables as inputs. The identification of country clusters further enables comparative insights and highlights region-specific challenges and opportunities. For policymakers, the study emphasises the importance of promoting economic stability, improving institutional frameworks and designing targeted interventions that consider demographic realities. It also calls for future research to incorporate additional socioeconomic dimensions and longitudinal data to more comprehensively capture the evolving nature of global labour markets.

1. INTRODUCTION

The functioning of the labour market is a complicated process and its mechanisms improve the economic soundness of the actors' economic decisions affecting the society's welfare (Lloret-Climent et al., 2020).

The labour market consists of three main actors: workers, companies and the state. Workers and companies have conflicting interests, as workers aim to maximise their earnings by working for higher wages, while companies seek to increase profits by hiring workers at lower wages (Borjas, 2020, p. 10). There are multiple approaches to studying the labour market, with no universal or comprehensive method. The opinions of various authors coincide in some areas but diverge in others (King, 1990, p. 10).

The goal of this article is to determine the principal components that explain the variability of a dataset containing variables which describe the labour market, such as workforce characteristics, economic, demographic, sociological and other variables, and to use the principal components as input in the cluster analysis. The variables were chosen based on the previous research on the factors that influence the labour market, and the hypotheses were formed accordingly.

The primary hypothesis is that the structure of global labour market variability can be described by a limited number of latent dimensions formed from economic, demographic and institutional variables. A secondary hypothesis is that these dimensions can be used to identify clusters of countries with distinct labour market profiles.

To test these hypotheses, the article employs Principal Component Analysis (PCA), a statistical method used to reduce the dimensionality of a dataset by transforming the original variables into a set of uncorrelated principal components that account for the maximum possible variance. (Greenacre et al., 2022). To enhance the analytical depth, cluster analysis, specifically k-means clustering, is performed on the PCA results to identify distinct groups of countries based on their labour market profiles (Umargono, Suseno, & Gunawan, 2020).

The next section reviews relevant literature on the determinants of labour market outcomes, laying the groundwork for the selection of variables and formulation of hypotheses. The methodology section describes the data sources, the treatment of missing data through multiple imputation, the procedures used to validate the assumptions for PCA, together with the k-means clustering methodology. The results section presents the extracted principal components, variance explained and findings from the cluster analysis. This is followed by a discussion of the substantive interpretation of the components and clusters, as well as their policy implications. The final section summarises key insights and outlines directions for future research, including the integration of longitudinal data and additional socioeconomic indicators to capture evolving labour market dynamics more comprehensively.

2. LITERATURE REVIEW

Economic theory suggests that an individual will be active in the labour market and, thus form a labour supply, if the difference between their actual earned wage and the minimum acceptable wage is positive (Boeri & Van Ours, 2021, p. 4). In other words, wages and earnings are the primary variables in the *labour*-leisure model, determining the allocation of time between work and leisure (Borjas, 2020, p. 19).

Other researchers suggest that, among the macroeconomic factors, the level of income at the national level, the level of unemployment, social inequality and the share of the urban population have a significant influence on individuals' expectations about the benefits obtained through work (Zamfir et al., 2021). As relative earnings increase, labour supply also increases, beacuse workers compare their earnings with past earnings and the earnings of others (Bracha, Gneezy, & Loewenstein, 2015), or with reference points that reflect expected earnings or earnings which the individual aspires to (Kahneman & Tversky, 1979). Wage levels have been proven to play a significant role in boosting the innovative activities of workers (Pieroni & Pompei, 2008).

The significance of demographic characteristics is such that all changes in the labour market that cannot be explained by activity or employment rates can be attributed to changes in the population and its structure (Blanchard & Katz, 1992). Additionally, it is noted that differences in activity within a population group can be attributed to individual characteristics that are not directly measurable and cannot be included in the analysis (Ben-Porath, 1973).

The labour market is also shaped by the structure of the country's industry and the products it produces and exports, with countries that produce more sophisticated products generally having lower unemployment and higher employment rates (Adam et al., 2021).

3. METHODOLOGY

This study employs PCA to identify the major dimensions of variability in global labour market data. Before conducting PCA, we ensured that the data met key assumptions using Bartlett's Test of Sphericity and the Kaiser–Meyer–Olkin

(KMO) measure of sampling adequacy. In addition, a multiple imputation (MI) technique was used to handle missing data and provide robust estimates.

PCA is a factor-extraction technique that finds linear combinations of the observed variables which capture the maximum variance (Greenacre et al., 2022). To reduce the dimension of the dataset and to derive the new, uncorrelated variables, the original data are projected into a new coordinate system, where the first axis corresponds to the direction where the data varies the most; the second axis corresponds to the direction where the data varies the most after the first direction, etc. The first principal component is the projection of the original data to the first principal axis and captures the greatest amount of the variance in the data. The second principal component is the projection of the data on the second principal axis, explaining the greatest portion of variance remaining after the first principal component. Each subsequent principal component is uncorrelated with the other components and explains the greatest portion of variance, while being orthogonal to the preceding principal components (Kherif & Latypova, 2020).

The PCA analysis was conducted in the R programming language, using the *prcomp()* command. The command centres and standardises the data before performing singular value decomposition (SVD) to decompose the data matrix and compute the principal components and loadings (Harvey & Hanson, 2024). The data matrix X has samples in rows and values of respective variables in columns, so the SVD decomposition breaks the matrix X into three matrices:

$$X = UDV^T$$

where D is a diagonal matrix with all non-diagonal elements zero, and diagonal containing singular values, the columns of V give the principal axes that define the new coordinate system, and the scores, which are the projections of the data on principal axes, are obtained by XV or UD (Harvey & Hanson, 2024).

The requirements for the PCA were checked by running Bartlett's Test of Sphericity and calculating the KMO measure of sampling adequacy, using the command *check_factorstructure()* in the R programming language (R project, 2024). The Bartlett's Test of Sphericity checks whether the variables' correlation matrix is different from an identity matrix and calculates the probability that the correlation matrix has significant correlations among at least some variables using the test statistics:

$$\chi^2 = -\left(N - 1 - \frac{2p+5}{6}\right)\ln|R|$$

where R is a correlation matrix of p variables and N is the number of observations (Bartlett, 1951). This approach reduces the risk of obtaining random principal components and factors, and their incorrect interpretation (Tobias & Carlson, 1969).

The KMO measure of sampling adequacy, which ranges from 0 to 1, indicates the extent to which each variable in the dataset is predicted without error using the other variables (Kaiser, 1974). The measure is calculated as follows:

$$KMO = \frac{\sum_{j \neq k} r_{jk}^2}{\sum_{j \neq k} r_{jk}^2 + \sum_{j \neq k} p_{jk}^2}$$

where the is the correlation between the variables j and k, and is the partial correlation (R project, 2025). A higher KMO value indicates that patterns of correlations are relatively compact and thus that factor analysis (or PCA) should yield distinct and reliable factors. As a rule of thumb, KMO > 0.8 is considered **meritorious** (i.e., very good), 0.7–0.8 is **middling**, and below 0.6 indicates the need for remedial measures (Hair et al., 2018; Kaiser, 1974).

The dataset contained 10.58% of the missing data, with none of the variables exceeding 50%. If more than 50% of observations were missing for a particular variable, it was excluded from the analysis according to recommendations by Madley-Dowd et al. (2019). Since PCA cannot be conducted on the dataset containing missing data, multiple imputation (MI) was used to handle the missing data, as it was the most commonly recommended technique (Van Ginkel, 2023). Multiple imputation consists of three steps. The first step involves estimating the missing data multiple times (M) using a statistical model that describes the data structure, thereby resulting in M different versions of the dataset that differ only in the estimates of the missing data. The desired analysis is then applied to each of the M datasets, yielding an equal number of analysis results, which are then combined into a single consolidated result (Van Ginkel, 2023).

3.1. Data and Variables

For the PCA, data on a broad set of economic, demographic and institutional variables were collected for 191 countries. The variables were chosen based on the results of the previous analyses, in the sense that the author included variables that had previously been found to have an impact on the labour market. Over 50 different variables from more than 10 sources were considered, but only a 28 of them were retained due to the unavailability of a certain number of observations.

The data for the study was mostly collected from *The Global Economy* database, which compiles statistical data on over 300 indicators from several reliable sources, such as national statistical institutes, the World Bank, the International Monetary Fund, the United Nations, the World Economic Forum and other sources (The Global Economy, 2024). Additionally, data on the Human Development Index (HDI) for 2021 was obtained from the United Nations Development Programme (UNDP) website, which publishes this index (UNDP, 2024). Data on average weekly working hours, average and minimum monthly wages, output per hour, average age and the percentage of young people not employed or in education was collected from the International Labor Organization (International Labor Organization, 2024). Data on migration rates was sourced from the Central Intelligence Agency (CIA, 2024).

Table 1 provides an overview of the variables included in the analysis, along with their descriptions and data sources.

Table 1: Overview of the Variables Included

Abbreviation	Description	Source
Country	Observed Country	*The Global Economy*
LPR	Labour Market Participation Rate	*The Global Economy*
GDPpcUSD	GDP per Capita, Current USD	*The Global Economy*
GDPpcPPP	GDP per Capita, Purchasing Power Parity	*The Global Economy*
HouseCons	Household Consumption as a Percentage of GDP	*The Global Economy*
UnempR	Unemployment Rate	*The Global Economy*
YUnempR	Youth Unemployment Rate, Ages 15-24	*The Global Economy*
FlabourF	Female Labour Participation Rate	*The Global Economy*
CorrIND	Corruption Perceptions Index (0-100)	*The Global Economy*
FreeCorrIND	Freedom from Corruption Index (0-100)	*The Global Economy*
BusIND	Business Freedom Index (0-100)	*The Global Economy*
LFreeIND	Labour Freedom Index (0-100)	*The Global Economy*
URB	Urban Population Percentage	*The Global Economy*
DEPEND	Dependency Ratio	*The Global Economy*
PerREFUG	Refugees as a Percentage of Total Population	*The Global Economy*
BrainDrainIND	Human Flight and Brain Drain Index (0-10)	*The Global Economy*
HappyIND	Happiness Index (0-10)	*The Global Economy*
DifMF_EMP	Male-Female Unemployment Rate Difference	*The Global Economy*
Popul	Population in Millions	*The Global Economy*
Per_F_Popul	Female Population Percentage	*The Global Economy*
HDI21	Human Development Index	*HDI*
PovertyR	In-Work Poverty Rate	*ILO*
MeanHRS	Average Weekly Working Hours	*ILO*
AvgWage	Average Monthly Wage	*ILO*
StatWage	Statutory Minimum Gross Monthly Wage	*ILO*
YNotEET	Percentage of Youth Not in Employment or Education	*ILO*
OutPerH	Output per Hour Worked	*ILO*
MeanAge	Average Population Age	*ILO*
Migr	Net Migration Rate	*CIA World Factbook*

Source: Developed by the author

Other variables that could have contributed to the analysis and provided additional insights were considered but ultimately not included due to a percentage of missing observations exceeding 50%. Excluded variables were the shadow economy as a percentage of GDP, health expenditure, the GINI index, percentage of impoverished population, percentage of GDP allocated to education, literacy rate, average years of education, percentage of highly educated population, percentage of religious population (by religion), cost of living index and level of social protection.

3.2. Sample Size

In the literature, various recommendations and opinions are given regarding sample size when conducting principal component analysis. However, formal guidelines for sample size in PCA and factor analysis are not extensive and often lack strong empirical support (Osborne & Costello, 2004).

Broadly, rules of thumb for sample size range from 3 to 6 observations per variable, with a minimum of 250 observations in total (Cattell, 1978, p. 508). Some authors state that the sample must have more observations than variables in the dataset and that 50 observations should be an absolute minimum for factor and principal component analysis, with a recommended ratio of 5 observations per variable (Hair et al., 2018, p. 101). Others suggest the sample size should be five times the number of variables (Gorsuch, 2014). Yet another common recommendation is a 10:1 ratio of observations to variables (Nunnally, 1978, p. 421), although this recommendation is not backed by publicly published research (Osborne & Costello, 2004).

The ratio of elements in the sample to the number of variables used in this study is 6.8:1, indicating that there are nearly seven observations per variable. This meets the oft-recommended ratio of 5:1 or higher, meaning the sample size is five times greater than the number of variables. The total number of elements in the sample is 191, which also aligns with the guideline that the minimum sample size should be at least 50. An important aspect of this analysis is that technically no sampling was performed - all internationally recognised countries were considered. The dataset could not be expanded further because it already encompasses all available data given the scope of this research. Therefore, within the current research scope and available data, it is not possible to obtain a larger sample or more observations.

3.3. Missing Data and Multiple Imputation

After confirming that the sample size meets the basic recommendations for PCA and assembling the dataset with 28 variables and 191 observations, the next step in the analysis was to address the missing observations. Due to the secondary nature of the data and its unavailability, it was not possible to collect values for all 28 variables in all 191 observed countries, resulting in gaps in the dataset.

In the entire dataset, there is a total of 10.58% missing values. Among the variables included in the analysis, the variable AvgWage had the highest percentage of missing data at 47%, followed by PovertyR at 37%, HouseCons with 35%, HappyIND at 30% and PerREFUG at 18%. All other variables have less than 10% missing values. For the variable StatWage, missing values were replaced with 0, assuming that the legal minimum wage is not defined in countries where data is unavailable.

According to the recommendations of Van Ginkel (2023), the data will be completed using multiple imputation with M=100 iterations. In this way, 100 different datasets with filled missing values were created. The method used for filling in missing values was predictive mean matching (PMM), which imputes values by using available data and weighting them with an appropriate metric (Van Buuren, 2012).

3.4. Adequacy for Principal Component Analysis

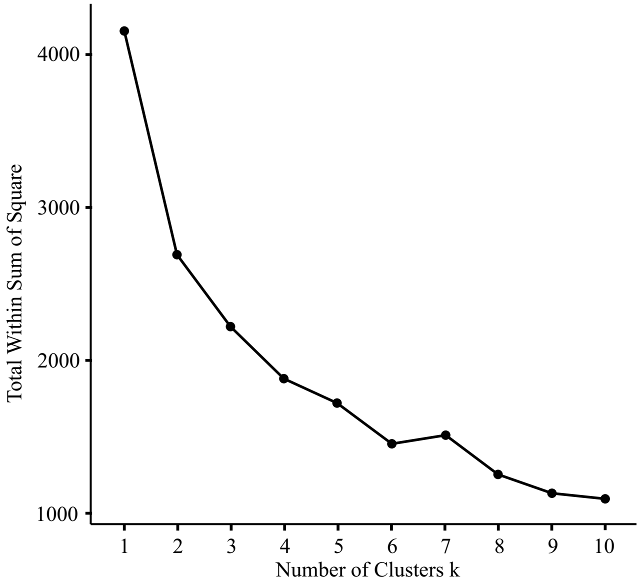
To verify the adequacy of the 100 imputed datasets for PCA, Bartlett's test of sphericity and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy were conducted on all 100 created datasets (R Project, 2024). In all 100 cases, Bartlett's test of sphericity had a p-value < 0.001, indicating a sufficient level of correlation among the variables to conduct PCA.

The overall KMO measure of sampling adequacy for the datasets ranged from 0.81 to 0.85, which places our data in the category of *good* in terms of suitability for factor analysis (Hair et al., 2018). Such KMO values confirm that the patterns of correlations are compact enough and that each variable has a significant amount of common variance to be explained by underlying factors. In other words, the imputed datasets are appropriate for PCA (Kaiser, 1974). Thus, both Bartlett's test and the KMO measure indicate that the data meet the key assumptions for PCA.

3.5. Cluster Analysis

Following the PCA, cluster analysis was applied to the principal component scores to identify groups of countries with similar labour market characteristics. The scores of the seven extracted principal components served as the input for clustering.

The clustering process used the k-means clustering technique. The optimal number of clusters was determined using the elbow method, which visually shows the differences in the sum of square error of each cluster. The most extreme difference forms the angle of the elbow showing the best cluster number (Umargono, Suseno, & Gunawan, 2020). This visual method indicated that four clusters were optimal for the dataset, as after *k=4* the rate of decrease slows down and the curve starts to flatten. The results are shown in Graph 1.



Graph 1: The Elbow Method
Source: Author's calculation

K-means clustering was then performed using the *kmeans ()* function in R, with the number of clusters set to four. This algorithm iteratively partitions the data into clusters by minimising the total intra-cluster variance (R Documentation, 2025). Each country was assigned to one of the four clusters based on the Euclidean distance to cluster centroids in the principal component space.

After assigning countries to clusters, average values of key labour market variables were calculated for each group. This allowed for an interpretation of the substantive characteristics of each cluster, including levels of economic development, institutional quality and demographic composition. The integration of PCA and clustering enables both dimensionality reduction and pattern recognition, providing a richer understanding of the global landscape of labour markets.

4. RESULTS

After collecting the data and checking their suitability for the PCA, the next step is to conduct the analysis. We created 100 imputed datasets using the multiple imputation method, which only differ in the values of the imputed observations. It is recommended to perform PCA on all of the imputed datasets separately and then pool the results into one outcome (Van Ginkel, 2023).

Following these recommendations, we carried out 100 separate PCAs (one for each imputed dataset), and then combined the results by averaging the PCA outcomes. In practice, this pooling was done by averaging the eigenvalues and component loadings across the 100 PCA runs.

Table 2 shows the eigenvalues of the correlation matrix for each principal component, based on the combined analysis.

Table 2: Eigenvalues

	Eigenvalues
PC1	10.911
PC2	3.328
PC3	2.267
PC4	1.948
PC5	1.273
PC6	1.105
PC7	1.035
PC8	0.894
PC9	0.730
PC10	0.630
PC11	0.582
PC12	0.522
PC13	0.472
PC14	0.397
PC15	0.358
PC16	0.288
PC17	0.255
PC18	0.222
PC19	0.169
PC20	0.149
PC21	0.118
PC22	0.099
PC23	0.082
PC24	0.050
PC25	0.042
PC26	0.038
PC27	0.019
PC28	0.015

Source: Author's calculation

According to Kaiser's criterion, principal components with eigenvalues greater than one (as seen in Table 2) should be included in the analysis since this criterion provides the number of interpretable components in empirical research (Kaiser, 1960). The number of retained principal components in this analysis should accordingly be seven.

Table 3 provides the standard deviations of each principal component and the proportion of total variance explained, both for each component individually and cumulatively. The standard deviation (SD) of a component is the square root of its eigenvalue. The proportion of variance is the eigenvalue divided by the total number of variables (28), and the cumulative percentage shows how much of the total variance is explained by all components up to the last one.

Table 3: Explained Variance

	Standard Deviation	Proportion of Variance	Cumulative Percentage
PC1	3.303	0.390	0.390
PC2	1.824	0.119	0.509
PC3	1.506	0.081	0.590
PC4	1.396	0.070	0.659
PC5	1.128	0.045	0.705
PC6	1.051	0.039	0.744
PC7	1.017	0.037	0.781
PC8	0.946	0.032	0.813
PC9	0.854	0.026	0.839
PC10	0.794	0.023	0.862
PC11	0.763	0.021	0.882
PC12	0.722	0.019	0.901
PC13	0.687	0.017	0.918
PC14	0.630	0.014	0.932
PC15	0.598	0.013	0.945
PC16	0.537	0.010	0.955
PC17	0.505	0.009	0.964
PC18	0.471	0.008	0.972
PC19	0.411	0.006	0.978
PC20	0.386	0.005	0.984
PC21	0.343	0.004	0.988
PC22	0.314	0.004	0.991
PC23	0.286	0.003	0.994
PC24	0.225	0.002	0.996
PC25	0.204	0.001	0.997
PC26	0.194	0.001	0.999
PC27	0.136	0.001	0.999
PC28	0.122	0.001	1.000

Source: Author's calculation

Based on Table 3, the first principal component alone explains about 39.0% of the variability in the dataset, and the second component explains about 11.9%, therefore together the first two components account for roughly 50.9% of the variance. Each of the remaining individual components explains less than 10% of the variance. All 28 principal components collectively explain 100% of the variability, but in practice, one would focus on the subset of components with larger eigenvalues. In our case, the first seven components have eigenvalues above 1.0 (per Kaiser's criterion), and together they explain about 77.9% of the total variability. Therefore, we retain these seven principal components for further analysis.

Table 4 shows the structure of the retained principal components, namely the loadings of the original variables within each component on each of the first seven principal components. A higher absolute loading indicates that the variable is more strongly associated with that component.

Table 4: Retained Principal Components

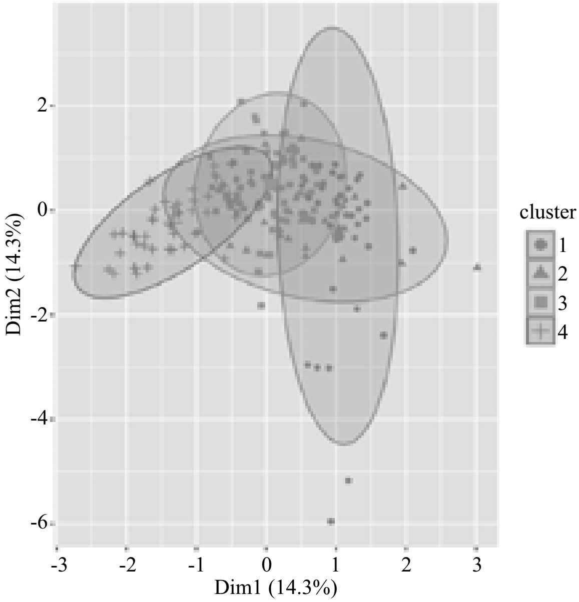
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
GDPpcUSD	0.265	-0.001	-0.003	0.011	-0.023	-0.012	0.003
GDPpcPPP	0.277	0.000	0.011	0.010	-0.014	-0.006	0.006
HouseCons	-0.195	-0.027	-0.030	-0.003	0.015	-0.013	0.028
UnempR	-0.088	-0.006	-0.015	-0.012	-0.026	0.049	0.001
YUnempR	-0.050	-0.003	-0.011	-0.010	-0.033	0.076	0.000
LPR	0.065	0.005	0.008	0.004	-0.050	-0.010	0.004
FLabourF	0.066	-0.001	-0.055	-0.009	0.020	0.012	-0.007
CorrIND	0.257	0.003	-0.014	-0.005	-0.026	-0.009	0.003
FreeCorrIND	0.262	0.004	-0.017	-0.007	-0.026	-0.016	0.003
BusIND	0.256	-0.013	-0.012	-0.011	-0.008	-0.020	-0.008
LFreeIND	0.138	-0.020	-0.032	-0.023	-0.109	-0.088	-0.017
URB	0.203	0.003	0.019	0.016	0.038	0.015	0.010
DEPEND	-0.172	-0.008	-0.043	0.004	-0.006	-0.069	-0.014
PerREFUG	-0.024	-0.002	-0.008	-0.005	0.066	-0.208	0.013
BrainDrainIND	-0.256	-0.003	-0.013	-0.024	-0.003	-0.035	0.021
HappyIND	0.242	-0.001	-0.024	0.005	0.009	0.056	0.017
DifMF_EMP	0.086	0.003	-0.018	-0.023	0.015	0.035	0.007
Popul	-0.006	0.000	0.026	-0.005	-0.016	0.033	-0.087
Per_F_Popul	-0.032	-0.008	-0.058	-0.023	0.076	-0.015	-0.017
HDI21	0.277	-0.002	0.009	-0.005	0.023	-0.005	0.005
PovertyR	-0.092	-0.017	-0.047	0.028	-0.032	0.008	-0.011
MeanHRS	-0.079	0.007	0.064	-0.009	0.018	-0.076	0.012
AvgWage	0.232	-0.004	-0.031	0.012	0.000	-0.006	0.008
StatWage	0.195	-0.008	-0.011	-0.010	-0.005	-0.028	-0.006
YNotEET	-0.224	-0.007	-0.002	-0.014	-0.054	0.017	0.012
OutPerH	0.269	-0.004	-0.005	0.007	-0.005	-0.005	0.002
MeanAge	0.244	0.002	-0.001	-0.012	0.040	0.006	-0.004
Migr	0.066	-0.019	-0.013	0.045	-0.008	-0.040	-0.044

Source: Author's calculation

The loadings in Table 4 are used to interpret each principal component in substantive terms. High positive or negative loadings indicate which variables are most strongly associated with a component.

To identify country groupings with similar labour market profiles, k-means clustering was applied using the scores from the seven retained principal components.

The optimal number of clusters was determined using the elbow method, which showed a clear inflexion point at four clusters. The k-means algorithm was run with *k=4*, resulting in four distinct clusters of countries. The four clusters are given in Graph 2.



Graph 2: Cluster Plot
Source: Author's calculation

To visually represent the cluster structure, a two-dimensional projection of the high-dimensional PCA scores was employed using multidimensional scaling (MDS) of the Euclidean distance matrix (Hout, Papesh, & Goldinger, 2013). The resulting cluster plot displays the observations (countries) in a reduced space defined by Dim1 and Dim2, which together capture approximately 28.6% of the total variation in inter-country distances within the PCA space. Notably, Dim1 and Dim2 each account for 14.3%, indicating a balanced distribution of the variation preserved in the two-dimensional view.

The visual separation of clusters supports the findings from the PCA. Cluster 1 (red) is well-separated along Dim1, reflecting countries with high levels of economic development, institutional quality and productivity. Cluster 2 (green) and Cluster 3 (blue) exhibit partial overlap in the centre of the plot, consistent with their intermediate socio-economic profiles and transitional labour market structures. Cluster 4 (purple) is positioned distinctly on the left side, capturing low-income countries characterised by lower development indices, higher demographic dependency and labour market exclusion.

While the clusters are not entirely non-overlapping, their spatial arrangement in this plot illustrates the relative similarity and dissimilarity of countries based on their labour market features as represented by the PCA dimensions. The compactness of Cluster 1 and Cluster 4 suggests greater internal homogeneity, whereas the dispersion of Clusters 2 and 3 indicates more heterogeneity in transitional economies.

Additionally, Table 5 presents the average values of selected indicators by cluster, offering insight into the defining characteristics of each group.

Table 5: Cluster Means by Variable

Cluster	1	2	3	4
GDPpcUSD	54615	3856	11893	2443
GDPpcPPP	57416	7822	20357	3992
HouseCons	46.70	71.46	62.09	72.34
UnempR	4.90	16.95	7.12	4.54
YUnempR	13.16	34.42	17.74	8.25
LPR	63.68	47.62	60.61	64.77
FLabourF	42.36	32.65	43.76	44.35
CorrIND	67.76	31.84	44.18	29.75
FreeCorrIND	74.93	32.74	45.69	26.31
BusIND	78.88	47.90	65.04	44.50
LFreeIND	61.85	53.06	56.79	49.40
URB	83.89	59.06	61.94	41.04
DEPEND	49.98	63.61	49.91	75.48
PerREFUG	0.009	0.037	0.006	0.004
BrainDrainIND	2.31	6.53	5.25	6.62
HappyIND	6.67	4.58	5.66	4.43
DifMF_EMP	-1.20	-8.02	-0.98	-0.50
Popul	27.36	24.08	63.03	33.34
Per_F_Popul	48.25	50.20	50.81	50.14
HDI21	0.91	0.64	0.77	0.55
PovertyR	1.26	13.15	3.60	27.48

Cluster	1	2	3	4
MeanHRS	35.84	40.11	39.20	38.84
AvgWage	3573.59	1087.82	964.77	604.54
StatWage	1050.39	105.88	282.31	67.87
YNotEET	10.13	31.84	17.98	25.38
OutPerH	62761	13809.93	22344.46	5027.08
MeanAge	40.47	26.24	35.80	21.38
Migr	2.81	0.82	-1.71	-0.89

Source: Author’s calculation

The cluster analysis reveals that economic development, demographic dynamics and institutional frameworks jointly shape national labour outcomes. Importantly, the clusters offer actionable groupings for comparative analysis, enabling countries to benchmark their performance and learn from peers facing similar structural conditions.

5. DISCUSSIONS

The PCA revealed that economic development, institutional quality and demographic composition are the most significant factors contributing to global labour market variability. The first principal component (PC1), which explains 39% of the total variance, captured a composite of variables including GDP per capita, the HDI, output per hour and institutional indicators such as business freedom and the perception of corruption. Countries scoring highly on this component tend to enjoy higher living standards, robust institutional frameworks and more efficient labour markets. This aligns with established economic theory, which posits that developed economies with effective governance structures tend to offer greater labour market stability and inclusion.

To translate these abstract dimensions into actionable insights, a k-means cluster analysis was applied to the scores from the first two principal components, together accounting for over 50% of the total variance. This step enabled the grouping of countries into four empirically grounded clusters that reflect distinct labour market regimes. The elbow method was used to determine the optimal number of clusters, identifying a clear inflexion point at four.

Cluster 1 is composed of high-income economies with advanced institutional infrastructure. These countries report the highest GDP per capita and HDI scores, coupled with low unemployment rates, high average wages and strong governance indicators. Labour force participation is high, particularly among women, and labour productivity (measured by output per hour) is among the highest globally. Countries in this cluster also have relatively older populations, highlighting demographic ageing as a policy priority.

Cluster 2 represents lower-middle-income countries grappling with deep-seated labour market challenges. These include very high rates of youth unemployment, low female participation, weak business environments and moderate-to-low levels of HDI. Despite a younger demographic profile, the institutional barriers in these countries hinder effective labour market integration. Poverty levels remain elevated, and informal employment is likely pervasive.

Cluster 3 encompasses a mix of emerging economies with transitional characteristics. These countries show moderate levels of economic development and governance, and they occupy a middle ground in terms of unemployment, labour participation and wage levels. Urbanisation is more advanced than in Cluster 2, and female participation and governance scores are generally higher. However, the variability within this cluster suggests heterogeneous policy needs.

Cluster 4 includes the world's lowest-income countries, marked by systemic disadvantage. These nations score the lowest across nearly all indicators: GDP per capita, HDI, institutional quality and productivity. Although unemployment rates appear low, this likely reflects high levels of informal labour. The population in these countries is among the youngest globally, and working poverty is particularly acute. Educational and healthcare infrastructure may also be underdeveloped, compounding labour market exclusion.

The cluster plot in Graph 2 provides visual confirmation of these groupings. Cluster 1 is differentiated along the axis associated with economic and institutional development. Cluster 4 appears at the opposite end, encapsulating structural disadvantage. Clusters 2 and 3 show partial overlap, reinforcing their transitional nature but also hinting at internal diversity in development trajectories.

These empirically derived clusters offer a practical typology for comparative labour market analysis. For policy design, this typology supports a more tailored approach: countries in Cluster 1 might prioritise innovation, labour market flexibility and managing demographic ageing, Cluster 2 would benefit from targeted strategies to enhance youth employment and reduce gender disparities, Cluster 3 may focus on strengthening governance and skills development to sustain economic momentum, and Cluster 4 requires foundational investment in education, healthcare and institution-building to address widespread informality and exclusion.

Moreover, the clustering results highlight how macroeconomic and institutional indicators interact with demographic factors to shape labour market outcomes. For instance, countries with similar GDP per capita may still diverge significantly in labour inclusion due to governance quality or demographic structure. This reinforces the importance of integrated policy approaches that go beyond economic growth to address structural and institutional barriers.

In sum, the PCA-cluster integration offers a robust framework for understanding the structural diversity of global labour markets. It not only confirms the centrality of economic and institutional development but also elucidates how these factors combine with demographic and social variables to produce distinct labour market regimes. The findings offer a valuable lens for international benchmarking and policy planning.

Future research could expand this framework by incorporating time-series data to capture the evolution of labour market structures over time. Additionally, disaggregated regional analyses could explore intra-cluster variation, enhancing the precision and applicability of the typology. Exploring the impacts of digitalisation, climate change or migration shocks on cluster dynamics could also yield important insights into future labour market resilience.

6. CONCLUSIONS

This article identifies the key latent dimensions that explain variability in labour market conditions on a global level. Through the application of PCA on a comprehensive dataset of 191 countries, seven principal components that explain 77.9% of the total variability in the dataset on the labour market were identified.

In support of our main hypothesis, these principal components align with macroeconomic, demographic and institutional factors. Moreover, the auxiliary hypothesis - that living standards, as measured by indicators such as GDP per capita and the Human Development Index, are primary drivers - was substantiated by the strong loadings of these variables on the first principal component. This component alone accounted for 39% of the variance, reflecting a powerful structural axis that differentiates countries according to their economic development and institutional quality.

The cluster analysis further strengthened these hypotheses by demonstrating that the principal components not only reduce dimensionality but also provide a meaningful basis for classifying countries into distinct labour market categories. The four clusters were differentiated through k-means clustering, offering compelling empirical support for the theoretical propositions. Specifically, the existence of well-defined clusters confirmed that countries do, in fact, group together in ways consistent with the structural dimensions captured by the PCA: primarily development level, institutional strength and demographic composition. Thus, the cluster findings validate and reinforce the hypothesis that structural macroeconomic and institutional differences underpin global labour market variability, with living standards emerging as the most decisive axis of differentiation.

Beyond dimensionality reduction, the article's major theoretical and empirical contribution lies in the integration of PCA with cluster analysis. Using the scores from the extracted principal components, a k-means clustering algorithm was used to reveal four distinct clusters of countries, each representing a unique labour market profile. This classification system adds interpretative depth to the PCA findings by translating abstract statistical patterns into tangible country groupings. Cluster 1 grouped high-income, institutionally robust countries with productive and inclusive labour markets. Cluster 2 represented lower-middle-income economies struggling with youth unemployment, gender inequality in labour participation and institutional fragility. Cluster 3 included transitional emerging markets with diverse demographic and institutional features. Cluster 4 contained the lowest-income countries with weak governance, high working poverty and young populations, where informality likely dominates labour market activity.

These clusters offer a powerful typology for understanding and comparing labour markets across diverse socio-economic contexts. The cluster plot differentiated these groups along the principal component axes, reinforcing the interpretive strength of the PCA and underscoring the robustness of the clustering procedure. Importantly, the clarity of the cluster separation lends further credibility to the PCA results by showing that countries naturally aggregate along key structural dimensions. The internal coherence of each cluster and the external distinctiveness between them provide strong support for the usefulness of this typology in comparative labour market research and policy analysis.

The used framework of combining PCA, multiple imputation to address missing data, and k-means clustering demonstrates the utility of statistical integration in capturing the complexity of labour market structures. The use of imputation ensured the integrity and completeness of the dataset, enabling a full exploitation of available information. PCA reduced multidimensional complexity while retaining explanatory power, and cluster analysis translated these results into an actionable classification. This integrative approach enhances the generalisability and analytical utility of the findings, offering a replicable model for future cross-country labour market studies.

Future research could expand on the cluster analysis by examining its temporal stability. Applying this PCA-cluster framework to panel data would allow researchers to track how countries move between clusters over time and identify the forces driving those transitions. Such longitudinal analysis would help distinguish structural features from those that are policy-responsive or temporally volatile. Additionally, more granular regional or sub-national clustering could reveal patterns obscured at the global level, while the inclusion of additional variables, such as digital infrastructure, labour laws, education quality or political stability, could enrich the analysis. Finally, exploring how global trends like digitalisation, automation and climate migration intersect with the identified clusters would help ensure that this typology remains relevant in an evolving global labour landscape.

In conclusion, this study advances theoretical understanding and methodological precision in the analysis of global labour market variability. The identification of interpretable structural dimensions and their mapping into cohesive country clusters represents a significant step toward a unified framework for labour market comparison. These clusters are not only analytically robust but also practically valuable, providing policymakers with a framework for benchmarking and designing context-sensitive labour market reforms.

Conflict of interests

The author declares there is no conflict of interest.

REFERENCES

Adam, A., Garas, A., Katsaiti, M. S., & Lapatinas, A. (2021). Economic complexity and jobs: An empirical analysis. *Economics of Innovation and New Technology, 31*(1), 25–52. https://doi.org/10.1080/10438599.2020.1859751

Bartlett, M. S. (1951). The effect of standardization on a χ2 approximation in factor analysis. *Biometrika, 38*(3/4), 337-344. https://doi.org/10.2307/2332580

Ben-Porath, Y. (1973). Labor-force participation rates and the supply of labor. *Journal of Political Economy, 81*(3), 697-704. https://doi.org/10.1086/260065

Blanchard, O., & Katz, L. (1992). Regional evolutions. *Brookings Papers on Economic Activity, 23*(1), 1-76.

Boeri, T., & Van Ours, J. (2021). *The economics of imperfect labor markets* (3rd ed.). Princeton University Press.

Borjas, G. (2020). *Labor economics* (8th ed.). McGraw-Hill Education, New York.

Bracha, A., Gneezy, U., & Loewenstein, G. (2015). Relative pay and labor supply. *Journal of Labor Economics, 33*(2), 297-315. https://doi.org/10.1086/678494

Cattell, R. (1978). *The scientific use of factor analysis in behavioral and life sciences*. Plenum Press, New York.

CIA. (2024). *The World Factbook*. https://www.cia.gov/the-world-factbook/

Gorsuch, R. (2014). *Factor analysis* (2nd ed.). Routledge, London.

Greenacre, M., Groenen, P., Hastie, T., D'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers, 2*(100), 1-10. https://doi.org/10.1038/s43586-022-00184-w

Hair, J., Black, W., Babin, B., & Anderson, R. (2018). *Multivariate data analysis* (8th ed.). Cengage Learning EMEA, Boston.

Harvey, D. T., & Hanson, B. A. (2024). *A comparison of functions for PCA*. https://cran.r-project.org/web/packages/LearnPCA/vignettes/Vig_07_Functions_PCA.pdf

Harvey, D. T., & Hanson, B. A. (2024). *The math behind PCA*. https://cran.r-project.org/web/packages/LearnPCA/vignettes/Vig_06_Math_Behind_PCA.pdf

Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(1), 93–103. https://doi.org/10.1002/wcs.1203

International Labor Organization. (2024). *ILOSTAT.* https://ilostat.ilo.org/data/

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263-292. https://doi.org/10.2307/1914185

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*(10), 141-151. https://doi.org/10.1177/001316446002000116

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31–36. https://doi.org/10.1007/BF02291575

Kherif, F., & Latypova, A. (2020). Principal component analysis. In A. Mechelli & S. Vieira (Eds.), *Machine Learning* (pp. 209–225). Academic Press. https://doi.org/10.1016/B978-0-12-815739-8.00012-2

King, J. E. (1990). *Labour economics* (2nd ed.). Red Globe Press, London.

Lloret-Climent, M., Nescolarde-Selva, J.-A., Mora-Mora, H., Alonso-Stenberg, K., & Mollá-Sirvent, R. (2020). Modeling complex social systems: A new network point of view in labour markets. *IEEE Access, 8*, 92110-92119.

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology, 110*, 63-73. https://doi.org/10.1016/j.jclinepi.2019.02.016

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill, new York.

Osborne, J. W., & Costello, A. B. (2004). Sample size and subject to item ratio in principal components analysis. *Practical Assessment, Research, and Evaluation, 9*, 1-9. https://doi.org/10.7275/ktzq-jq66

Pieroni, L., & Pompei, F. (2008). Labour market flexibility and innovation: Geographical and technological determinants. *International Journal of Manpower, 29*(3), 216-238. https://doi.org/10.1108/01437720810878897

R Project. (2025). *Kmeans: K-Means Clustering*. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kmeans

R Project. (2024). *Check suitability of data for Factor Analysis (FA) with Bartlett's Test of Sphericity and KMO*. https://search.r-project.org/CRAN/refmans/performance/html/check_factorstructure.html

R Project. (2025). *Find the Kaiser, Meyer, Olkin Measure of Sampling Adequacy*. https://search.r-project.org/CRAN/refmans/psych/html/KMO.html

The Global Economy. (2024). *About the site*. https://www.theglobaleconomy.com/

Tobias, S., & Carlson, J. E. (1969). Brief report: Bartlett's test of sphericity and chance findings in factor analysis. *Multivariate Behavioral Research, 4*(3), 375–377. https://doi.org/10.1207/s15327906mbr0403_8

Umargono, E., Suseno, J. E., & Gunawan, S. K. V. (2020). K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula. In *Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)* (pp. 121–129). Atlantis Press. https://doi.org/10.2991/assehr.k.201010.019

UNDP. (2024, February 2). *Human development index (HDI)*. https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

Van Buuren, S. (2012). *Flexible imputation of missing data*. Chapman & Hall/CRC Press, New York.

Van Ginkel, J. (2023). Handling missing data in principal component analysis using multiple imputation. In L. Van der Ark, W. Emons, & R. Meijer (Eds.), *Essays on contemporary psychometrics* (pp. 141-161). New York, NY: Springer. https://doi.org/10.1007/978-3-031-10370-4_8

Zamfir, A. M., Năstasă, A., Aldea, A., & Molea, R. (2021). Factors shaping labour market participation. *Postmodern Openings, 12*(1), 91-101. https://doi.org/10.18662/po/12.1/247

ОБЈАШЊЕЊЕ ВАРИЈАБИЛИТЕТА НА ГЛОБАЛНОМ ТРЖИШТУ РАДА КОРИШЋЕЊЕМ АНАЛИЗЕ ГЛАВНИХ КОМПОНЕНТИ И ВИШЕСТРУКЕ ИМПУТАЦИЈЕ

1 Милица Марић, Економски факултет Универзитета у Бањој Луци, Бања Лука, Босна и Херцеговина

САЖЕТАК

Овај чланак истражује основне факторе који доприносе варијабилности глобалног тржишта рада, примјеном анализе главних компоненти (PCA) и кластер анализе на подацима из 191 земље. Користећи широк скуп економских, демографских и институционалних показатеља, анализа има за циљ да открије примарне димензије које обликују динамику тржишта рада широм свијета. Кључни показатељи обухватају БДП по глави становника, индекс хуманог развоја (HDI), стопе незапослености и сиромаштва, индексе слободе рада и перцепције корупције, просјечне плате, као и демографске карактеристике попут структуре становништва и стопа миграције. Са циљем да се обезбиједи комплетност и поузданост података, примијењена је метода вишеструке импутације како би се попуниле недостајуће вриједности. Затим је примијењена PCA ради смањења димензионалности и идентификовања латентних структура у подацима. Добијене главне компоненте су потом коришћене у кластер анализи, што је издвојило четири кластера земаља са сличним профилима тржишта рада.

Резултати анализе потврђују да економски развој и квалитет институција у највећој мјери објашњавају варијабилитет на тржишту рада, при чему демографске варијабле попут старосне структуре и миграција такође имају значајну улогу. Ови налази подржавају хипотезу да се кластери земаља са сличним профилима тржишта рада могу идентификовати користећи економске, демографске и институционалне варијабле. Раздвајање земаља по кластерима додатно омогућава компаративну анализу и истиче регионалне изазове на тржишту рада. За креаторе политика, резултати указују на важност подстицања економске стабилности, унапређења институционалних оквира и креирања циљаних интервенција које узимају у обзир демографску стварност. Такође, будућа истраживања би могла укључити додатне социоекономске димензије и лонгитудиналне податке како би се стекла свеобухватнија слика о кретању свјетског тржишта рада.

Кључне ријечи: *тржиште рада, варијабилитет, анализа главних компоненти.*