Communications for Statistical Applications and Methods
2021, Vol. 28, No. 5, 1–17section.5

https://doi.org/10.29220/CSAM.2021.28.5.001
Print ISSN 2287-7843 / Online ISSN 2383-4757

Bootstrap Methods for MANOVA and GPA of Socio-Economic Factors in District Cluster of Russia

Andros Natalia[a], Yong-Seok Choi[1,a]

[a]Department of Statistics, Pusan National University, Korea

Abstract

This paper presents a comprehensive analysis of socio-economic and demographic variables across four clusters formed on the basis of the Federal and Military Districts of the Russian Federation. The objective is to identify regional disparities and economic imbalances through multivariate assessment and statistical testing. The paper begins with an overview of Russia's Federal and Military Districts, followed by the calculation of descriptive statistics for each cluster. Socio-economic data were obtained from authoritative sources such as ROSSTAT and EMISS, covering the period from 2010 to 2021. With these data, both MANOVA and GPA (generalized profile analysis) can be applied to our analysis to compare clusters. Moreover, MANOVA and GPA require the assumptions of multivariate normality and homogeneity of covariance matrices. To account for violations of multivariate normality and homogeneity of covariance matrices assumptions, we propose bootstrap methods of MANOVA and GPA. This approach ensured a more robust and reliable examination of significant cluster differences.

Keywords: Bootstrap, MANOVA, Generalized Profile Analysis, socio-economic factors.

1. Introduction

This paper investigates regional socio-economic and demographic patterns in Russia through multivariate statistical analysis. The study focuses on clusters formed from the Federal and Military Districts of the Russian Federation. The primary objective is to analyze and compare socio-economic and demographic indicators across regional clusters in Russia through multivariate assessment and statistical testing. The study is motivated by persistent disparities in economic development, labor market conditions, and demographic trends, which have important implications for national policy.

The paper begins with a brief overview of the Federal and Military Districts of Russia. Section 2 defines the socio-economic and demographic variables used for the analysis, obtained from two authoritative data sources: ROSSTAT and EMISS. The data include unemployment rates, employment rates, consumer basket, inflation, number of births, deaths, marriages, divorces, and active population. These quarterly values were collected from 2010 to 2021, providing a foundation for subsequent statistical analysis. Descriptive statistics are computed for each cluster, followed by visualizations such as boxplots and biplots (Jolliffe, 2002). Section 3.1 explains the rationale for applying MANOVA and GPA, as well as their respective algorithms (Choi, 2019). This section discusses the need to test the data for multivariate normality and homogeneity of covariance matrices, which were assessed using

Footnote for research fund.

[1] Corresponding author: Department of Statistics, Pusan National University, 2, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Korea. E-mail: yschoi@pusan.ac.kr

Published 30 September 2021 / journal homepage: http://csam.or.kr

Mardia's and Box's M tests, respectively (Srivastava, 2002). Since diagnostic tests indicate violations of these assumptions, bootstrap methods are introduced as a more reliable alternative. Section 3.2 focuses on the algorithms of the Bootstrap MANOVA and GPA. The bootstrap method for MANOVA was applied to test for significant differences across the clusters, using the multivariate bootstrap test (Choi, 2021). At the same time, Bootstrap GPA was developed to assess whether the clusters exhibited similar or distinct socio-economic profiles. Section 4 presents the results of the Bootstrap MANOVA and GPA, providing a detailed interpretation of the differences identified across the clusters. Particular attention is paid to how these variables differ across the four clusters, highlighting economic imbalances and demographic trends that may influence regional policy. The results are supported by confidence intervals and visualizations, which enhance the interpretability and robustness of the findings.

Finally, Section 5 concludes the study by summarizing the main findings and discussing their policy implications, with a particular focus on regional disparities and the need for targeted socio-economic strategies.

2. Data

The analysis begins with a brief overview of the administrative division of Russia, focusing on the Federal Districts — macro-regions that group the subjects of the Federation for efficient governance and regional coordination.

Figure 1: *The Federal Districts of the Russian Federation*

- Central District: The economic and geographical position can be assessed as beneficial for economic development. First of all, as the capital, Moscow is the country's largest center of science, culture, management, and education;
- Northwestern District: The country's second most important center of science and higher education. The district is favorable for the development of most sectors of the economy. A significant drawback is the harsh natural conditions in the northeastern part of the district;
- Southern District: The district occupies an advantageous geographical position, with ice-free access to the Black Sea and proximity to developed European and resource-rich Central Asian countries. It is strategically located with access to the ice-free Black Sea, but faces industrial and trade limitations;

- Caucasus District: Natural conditions are favorable for human life and agriculture. Economic opportunities are much lower than the national average, with low productivity and wages. High unemployment, lack of land, and water resources for agriculture form the basis for social conflicts;
- Volga District: The district has a favorable economic and geographical position. A major industrial hub with strong automotive and aerospace industries. The district is well-connected to other areas of the country and Central Asia countries;
- Ural District: In terms of self-sufficiency in resources, technical means, and technologies, the Ural Federal District is the most complex region and has the potential to become a leader in the country's economy. The main disadvantages are harsh natural conditions;
- Siberian District: It has the least favorable natural conditions for human habitation. On the positive side, the district has an abundance of natural resources and acts as a transit route between Europe and Asia;
- Far Eastern District: The district has an unfavorable economic and geographical position for human habitation and the development of most sectors of the economy. The district's advantage lies in its wealth of natural resources. The main significance lies in its geopolitical position, bordering China, North Korea, America, and Japan.

In addition to the administrative division, Russia's territory is also structured into Military Districts. Military Districts are critical components of the country's defense system and play a vital role in ensuring national security. These serve as the basis for the four clusters examined in this paper, reflecting both security and economic dimensions.

Figure 2: *The Military Districts of the Russian Federation*

- Central cluster: The Central Military District covers the expansive areas of the Volga, Ural, and Siberian Federal Districts. Due to its central location, this district is crucial for rapidly mobilizing forces across Russia and serves as the state's strategic reserve. This area is rich in natural resources and has a strong industrial base;
- Southern cluster: The Southern Military District focuses on securing the Black Sea region and maintaining a robust defensive stance in Russia's southern territories. This district includes the

Southern and Caucasus Federal Districts, known for their favorable climate for agriculture, significant agro-industrial production, and strategic naval importance;

- Western cluster: The Western Military District includes major cities of federal significance such as Moscow and St. Petersburg, located in the Central and Northwestern Federal Districts. The Military District plays a crucial role in protecting Russia's interests in the Baltic Sea region. Further, the Central and Northwestern Federal Districts contribute significantly to the economy and politics of Russia, especially in science, education, and culture. These districts benefit from favorable economic and geographical positions;
- Eastern cluster: The Eastern Military District ensures national defense for Russia's vast eastern territories, rich in natural resources and critical for security in the Asia-Pacific region. The strategic importance of this district is underscored by its location and resource wealth, which are essential for Russia's long-term security and economic interests.

This section also provides definitions and descriptions of the variables included in the analysis. The unemployment rate is a socio-economic indicator representing the portion of the economically active population. According to the International Labor Organization methodology, an unemployed person is defined as an individual of working age (over 15 years).

The formula for calculating the unemployment rate is the following:

$$Unemployment\ rate = \frac{number\ of\ unemployed}{economically\ active\ population} \times 100\% \tag{2.1}$$

The economically active population includes citizens over 15. Now, in the Russian Federation, the retirement age for men is 65, and for women, it is 60.

$$Economically\ active\ population = number\ of\ employed + number\ of\ unemployed \tag{2.2}$$

Russia's population employment structure is analyzed using the following formula for the employment rate:

$$Employment\ rate = \frac{number\ of\ employed}{total\ population} \times 100\% \tag{2.3}$$

Inflation is an important macroeconomic indicator that represents an increase in general price levels. This paper utilizes quarterly data in Russia from 2010 to 2021.

$$Inflation = \frac{CPI_{current} - CPI_{previous}}{CPI_{previous}} \times 100\%. \tag{2.4}$$

The consumer price index (CPI) is one of the main indicators of inflation.

$$CPI = \frac{\sum_{i=1}^{n} Q_i^0 \times P_i^t}{\sum_{i=1}^{n} Q_i^0 \times P_i^0} \tag{2.5}$$

where n is the number of goods in the consumer basket; Q_i^0 is the number of goods in the basket for the base period; P_i^0 and P_i^t are base price and current price, respectively.

The consumer basket is the minimum set of food and non-food goods and services.

$$Consumer\ basket(100\%) = food(50\%) + non\text{-}food\ products(50\%) \tag{2.6}$$

Demographic indicators, such as the numbers of births and deaths, accurately reflect a country's social, economic, and political situation. Next, the variables representing marriages and divorces capture the number of officially registered unions and dissolutions, respectively.

Table 1: Variables Overview

Variable	Contents
y_1	Unemployment rate (%)
x_1	Employment rate (%)
x_2	Births (number of births)
x_3	Deaths (number of deaths)
x_4	Economically active (EA) population
x_5	Consumer basket (rubles)
x_6	Inflation (%)
x_7	Marriages (number of marriages)
x_8	Divorces (number of divorces)

As part of the ongoing study, the analysis focuses on the period from 2010 to 2021, as quarterly data for all eight Federal Districts of Russia are available for this time frame.

Figure 3 shows the unemployment rate trend in Russia from 2000 to 2021, highlighting key changes over time. The graph shows a significant decrease in unemployment from 10.6% in 2000 to 4.8% in 2021. The data capture the effects of various economic cycles, including periods of growth and recession, and the impact of global events such as the 2008 financial crisis and the COVID–19 pandemic in 2020–2021.

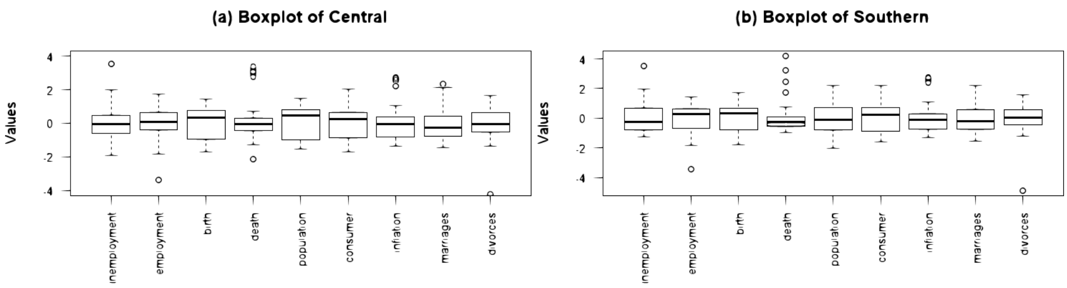
Figure 3: *Unemployment Rate Trend Graph from 2000 to 2021*

In addition to Figure 3, Table 2 provides a detailed overview of regional differences in employment, the numbers of births and deaths, inflation, and other important factors. From Table 2, the Western cluster has more favorable economic conditions with the lowest unemployment and inflation rates, while the Eastern cluster faces higher unemployment and living costs. The Southern cluster has the lowest employment rate among the four clusters. The Central cluster deals with higher unemployment and inflation challenges.

Table 2: Basic Descriptive Statistics

Clusters	Variable	Mean	SD	Median	Min	Max
Central	y_1 = Unemployment	6.03	1.02	5.97	4.07	9.67
	x_1 = Employment	63.92	1.00	64.03	60.53	65.70
	x_2 = Births	62719.55	10222.02	66022.82	45035.66	77427.31
	x_3 = Deaths	44514.59	3980.91	44390.65	36064.33	57888.65
	x_4 = EA population	10360.84	26746.37	10542.79	9746.77	10954.30
	x_5 = Consumer	10372.11	2280.64	10950.95	6511.29	15046.38
	x_6 = Inflation	19.01	9.66	18.48	5.88	45.23
	x_7 = Marriages	37757.64	15155.60	33609.99	15596.00	73481.98
	x_8 = Divorces	22670.73	2557.80	22570.17	11877.33	26974.66
Southern	y_1 = Unemployment	6.22	0.86	6.00	5.13	9.23
	x_1 = Employment	40.58	0.91	40.87	37.13	42.03
	x_2 = Births	156827.50	18879.44	163076.99	122913.99	189082.65
	x_3 = Deaths	169353.00	19334.47	164233.32	150271.66	250069.99
	x_4 = EA population	26768.06	2320.07	394.78	26294.47	27283.27
	x_5 = Consumer	6820.55	1536.13	7154.87	4363.06	10215.20
	x_6 = Inflation	13.21	7.09	12.57	3.89	32.78
	x_7 = Marriages	94569.48	34349.69	86931.66	41168.33	170623.33
	x_8 = Divorces	54487.34	5592.71	54730.33	27254.99	63471.66
Western	y_1 = Unemployment	2.64	0.45	2.58	2.03	4.13
	x_1 = Employment	45.20	0.66	45.29	43.47	46.17
	x_2 = Births	47112.66	5102.16	47353.33	37098.99	56826.32
	x_3 = Deaths	61397.51	6802.52	59797.16	53844.99	88227.65
	x_4 = EA population	9570.04	70.80	9583.47	9413.00	9712.80
	x_5 = Consumer	7596.75	1774.50	8040.78	4725.28	11046.88
	x_6 = Inflation	13.53	7.15	12.65	4.92	34.62
	x_7 = Marriages	33128.35	12765.50	31261.99	14412.66	59962.33
	x_8 = Divorces	19240.70	2050.95	19445.17	8457.66	22947.32
Eastern	y_1 = Unemployment	6.40	0.10	6.30	4.70	10.40
	x_1 = Employment	65.09	1.34	65.20	60.70	67.40
	x_2 = Births	21129.12	1647.76	21044.50	17085.00	24323.00
	x_3 = Deaths	22027.48	4337.92	20575.99	18018.99	36196.98
	x_4 = EA population	3768.49	425.58	3456.15	3340.60	4351.00
	x_5 = Consumer	14461.00	2756.88	15630.24	9625.17	19753.87
	x_6 = Inflation	18.91	9.11	18.64	5.46	44.69
	x_7 = Marriages	13464.60	3698.77	12867.50	7710.00	21397.00
	x_8 = Divorces	8755.43	1195.59	8739.50	5177.99	11860.99

Next, boxplots were constructed to visualize the distribution of each variable across the four clusters, identifying trends such as outliers and variability (Choi, 2018). Throughout 2020–2021 each cluster shows an increase in the number of deaths, likely due to the higher number of deaths during the COVID–19 pandemic. Unemployment reached its highest levels during 2010–2011, reflecting the Great Recession (2008–2013).



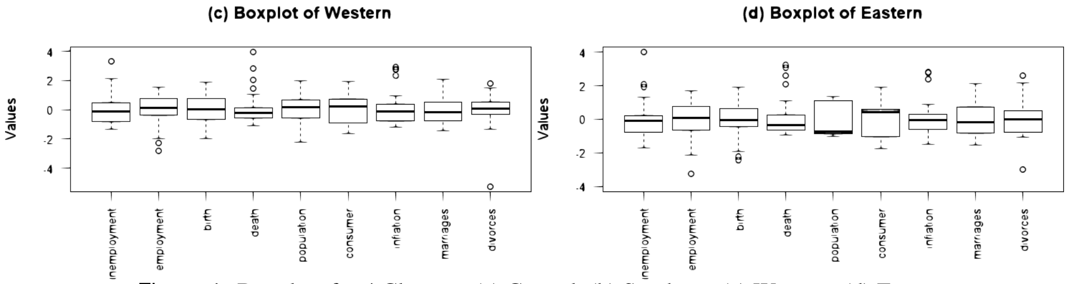


Figure 4: *Boxplots for 4 Clusters: (a) Central, (b) Southern, (c) Western, (d) Eastern*

The biplots in Figure 5 provide an overview of the relationships between the variables in each cluster (Jolliffe, 2002). Notably, unemployment and employment rates have negative correlation across all clusters. Demographic variables tend to cluster together across all clusters, indicating their interrelated nature. Economic patterns, like inflation, positively correlate with unemployment, which in turn reflects economic uncertainty in Russia.

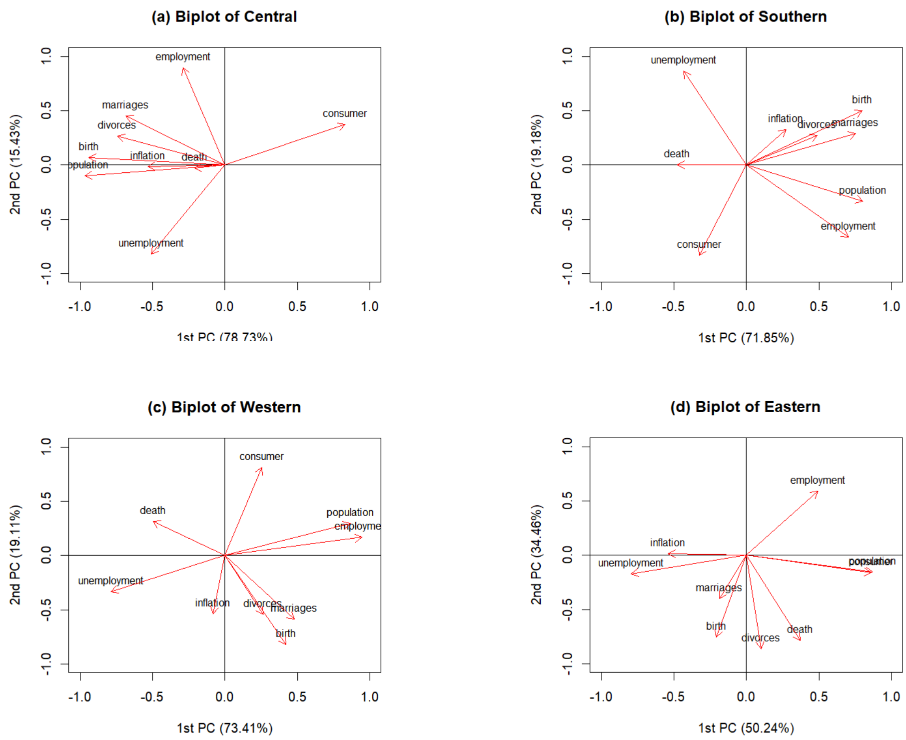


Figure 5: *Biplots for the Relationship Between Unemployment and Other Variables in 4 Clusters: (a) Central, (b) Southern, (c) Western, (d) Eastern*

3. Methods

3.1. MANOVA and GPA

We use MANOVA and GPA to test whether there are significant distinctions in socio-economic and demographic indicators among the four clusters. MANOVA and GPA assume that the data follow a multivariate normal distribution, and the covariance matrices of the groups being compared are equal. That is why it is crucial to assess multivariate normality. Additionally, Box's M test is applied for several theoretically grounded reasons (Srivastava, 2002).

In preparation for these analyses, min-max normalization is applied to the datasets since the data are presented at different scales. Min-max normalization scales all features between 0 and 1.

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{3.1}$$

where the normalized value is x_{scaled}, the original value is x, the minimum value is $x_{\min}$ and the maximum value in the dataset is $x_{\max}$.

To assess multivariate normality, Mardia's test is applied. Determining whether data are multivariate normally distributed is usually done by checking multivariate skewness and kurtosis. Here, we use Mardia's test with the multivariate skewness and kurtosis, and their approximate distributions:

$$\begin{aligned} Skewness &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}^3 \simeq \frac{6}{n}\chi^2_{(p(p+1)(p+2)/6)}, \\ Kurtosis &= \frac{1}{n}\sum_{i=1}^{n} m_{ii}^2 \simeq N_{(p(p+2),8p(p+2)/n)}, \end{aligned} \tag{3.2}$$

where $m_{ij} = (\mathbf{x}_i - \overline{\mathbf{x}})^T S^{-1}(\mathbf{x}_j - \overline{\mathbf{x}})$ is squared Mahalanobis distance with i^{th} and j^{th} deviations from $\overline{x} = \sum_{i=1}^{n} x_i/n$

Similarly, to assess the homogeneity of covariance matrices across groups, Box's M test is performed. Box's M test is used to determine whether the g covariance matrices are all equal. If this assumption is violated, it may affect the validity of the multivariate tests. The null hypothesis was set as follows:

$$H_0 : \Sigma_1 = \ldots = \Sigma_g = \Sigma$$

The data for each group j consists of n_j observations with p-dimensional vector and the M test statistic:

$$\begin{aligned} &\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^T, \quad i = 1, \ldots, n_j \sim N_p(\boldsymbol{\mu}_j, \Sigma_j), \quad j = 1, \ldots, g, \\ &M = -2\theta \log \Lambda = \theta \sum_{j=1}^{g} (n_j - 1) \log |S_j^{-1} S_{pooled}|, \end{aligned} \tag{3.3}$$

where Λ is test statistic, S_{pooled} is combined covariance matrix, S_j is covariance matrix, and θ is a scaling constant introduced to account for small sample bias in the Box's M statistic:

$$\Lambda = \prod_{j=1}^{g}\left(\frac{|S_j|}{|S_{pooled}|}\right), \quad S_{pooled} = \frac{\sum_{j=1}^{g}(n_j - 1)S_j}{\sum_{j=1}^{g}(n_j - 1)},$$
$$\text{and } \theta = 1 - \left[\sum_{j=1}^{g}\frac{1}{n_j - 1} - \frac{1}{\sum_{j=1}^{g}(n_j - 1)}\right]\left[\frac{2p^2 + 3p - 1}{6(p+1)(g-1)}\right]. \tag{3.4}$$

Therefore, from (3.3), if $M > \chi^2_{(p(p+1)(g-1)/2)}(\alpha)$, then H_0 is rejected, the critical value at the significance level α is $\chi^2_{(p(p+1)(g-1)/2)}(\alpha)$.

After verifying the assumptions, MANOVA can be applied to test the multivariate mean, which is the null hypothesis H_0 among the four groups (Choi, 2019).

The overall MANOVA model:

$$\mathbf{y}_{ij} = \boldsymbol{\mu}_j + \mathbf{e}_{ij}, \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, g,$$

where $\mathbf{y}_{ij}$ is the multivariate response for the i^{th} observation in the j^{th} group, $\boldsymbol{\mu}_j$ is the mean vector of the responses for the j^{th} group, and $\boldsymbol{e}_{ij} \sim N_p(\boldsymbol{\mu}_j, \Sigma)$.

The null hypothesis for MANOVA:

$$H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_g$$

Compute the Wilks' Lambda, the test statistic for evaluating the null hypothesis.

$$\Lambda = \frac{|SSE|}{|SSE + SST|} \sim U_(p, m, f), \tag{3.5}$$

where Wilks' Lambda follows a U-distribution with error degree of freedom $f = n - g$ and model degree of freedom $m = g - 1$, SSE $= \sum_{j=1}^{g}\sum_{i=1}^{n_j}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_j)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_j)^T$, and SST $= \sum_{j=1}^{g}\sum_{i=1}^{n_j}(\mathbf{y}_{ij} - \bar{\mathbf{y}})(\mathbf{y}_{ij} - \bar{\mathbf{y}})^T$.

In this paper, the Wilks' test statistic in (3.5) is Bartlett's modified test statistic that approximately follows the χ^2 distribution with $df = p \times m = p(g - 1)$:

$$\chi^2 = -\left(f - \frac{p - m + 1}{2}\right)\log\Lambda = -\left(n - g - \frac{p - (g-1) + 1}{2}\right)\log\Lambda = -\left(n - 1 - \frac{p + g}{2}\right)\log\Lambda \tag{3.6}$$

Therefore, from (3.6), if $\chi^2 > \chi^2_{(p(g-1))}(\alpha)$, then H_0 is rejected, where the critical value at the significance level α is $\chi^2_{(p(g-1))}(\alpha)$.

Following MANOVA, which assesses the overall multivariate differences across the clusters, a GPA is conducted to further explore the structure of these differences. While MANOVA tests for overall mean differences, GPA enables a more detailed examination of the patterns, levels, and interactions among multiple variables across the groups. According to Choi (2019), GPA can examine the various relationships of the average vector by p variables and g groups and follows the MANOVA model:

$$\mathbf{y}_{ij} = \boldsymbol{\mu}_j \ + \ \mathbf{e}_{ij} \tag{3.7}$$

The algorithms for parallel profiles (H_{01}), coincident profiles (H_{02}), and level profiles (H_{03}) are based on χ^2 and F-distributions. Specify null hypotheses for parallel profiles, coincident profiles, and level profiles.

1. H_{01}: Parallel Profiles

$$\begin{aligned} &H_{01} : C(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_g) = \ldots = C(\boldsymbol{\mu}_{g-1} - \boldsymbol{\mu}_g) = \mathbf{0} \\ &\Leftrightarrow H_{01} : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_g = \gamma_1 \mathbf{1}_p, \boldsymbol{\mu}_2 - \boldsymbol{\mu}_g = \gamma_2 \mathbf{1}_p, \ldots, \boldsymbol{\mu}_{g-1} - \boldsymbol{\mu}_g = \gamma_{g-1} \mathbf{1}_p \end{aligned} \tag{3.8}$$

where C is the contrast matrix, γ is level difference between i^{th} and j^{th} group, $\mathbf{1}_p = (1, \ldots, 1)^T$ is a p-vector of ones.

Compute the test statistics, such as Λ_{01} and χ^2_{01} statistics, use the following formulas:

$$\Lambda_{01} = \frac{|C(SSE)C^T|}{|C(SSE + SST)C^T|} \sim U_{(p-1, g-1, n-g)}, \ \chi^2_{01} = -\left(n - g - \frac{p + 1 - g}{2}\right) \log \Lambda_{01} \tag{3.9}$$

Make decisions about the hypotheses based on the computed test statistics, use the following formula:

$$\chi^2_{01} > \chi^2_{(p-1)(g-1)}(\alpha) \Rightarrow H_{01} \text{ is rejected.} \tag{3.10}$$

2. H_{02}: Coincident Profiles

$$H_{02} : \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{g-1})^T = 0 \ \Leftrightarrow H_{02} : \mathbf{1}_p^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_g) = \ldots = \mathbf{1}_p^T(\boldsymbol{\mu}_{g-1} - \boldsymbol{\mu}_g) = \mathbf{0} \tag{3.11}$$

Compute the test statistics, such as Λ_{02} and F_{02} statistics, and make decisions about the hypotheses based on the computed test statistics, use the following formula:

$$\begin{aligned} &\Lambda_{02} = \Lambda / \Lambda_{01} \sim U_{(1, g-1, n-g, n-g-p+1)}, \\ &F_{02} = \frac{n - g - p + 1}{g - 1}\left(\frac{1 - \Lambda_{02}}{\Lambda_{02}}\right) > F_{g-1, n-g-p+1}(\alpha) \Leftrightarrow H_{02} \text{ is rejected.} \end{aligned} \tag{3.12}$$

3. H_{03}: Level Profiles

$$H_{03} : \boldsymbol{\mu}_g = \delta \mathbf{1}_p \Rightarrow H_{03} : C\boldsymbol{\mu} = \mathbf{0}, \tag{3.13}$$

where δ is a constant vector with equal components, representing the null hypothesis that all group means are equal.

Compute the test statistic, such as F_{03} statistic, and make decisions about the hypotheses based on the computed test statistic, SSE and SST, use the following formula:

$$F_{03} = \frac{n(n - p + 1)}{p - 1} \bar{\mathbf{y}}^T C^T \left[C(SSE)C^T + C(SST)C^T\right] C\bar{\mathbf{y}} > F_{p-1, n-p+1}(\alpha) \Leftrightarrow H_{03} \text{ is rejected.} \tag{3.14}$$

3.2. Bootstrap Methods of MANOVA and GPA

In this paper, the original data do not satisfy the assumptions of multivariate normality and homogeneity of covariance matrices, as indicated by the Mardia's test and Box's M test in Section 3.1. Given that, relying on traditional MANOVA and GPA procedures would risk unreliable conclusions. Therefore, to address this issue and ensure the robustness of the inference, bootstrap methods that do not rely on strict distributional assumptions are employed. That is why bootstrap methods are applied for MANOVA and GPA.

The bootstrap method for MANOVA with more than three clusters was explained using the multivariate bootstrap test (Choi, 2021). The algorithm for Bootstrap MANOVA is provided below. The number of groups is g, the number of variables is p, n is the total number of observations, n_j is the number of observations in the j^{th} group, $\boldsymbol{x}_{ij}$ is the vector for the i^{th} observation in the j^{th} group, and $\overline{\boldsymbol{x}}_j$ is the mean vector for the j^{th} group.

[Step 1] Calculate the $p \times 1$ new vectors of observation for each group:

$$\boldsymbol{u}_{ij} = \left(\frac{n_j}{n_j - 1}\right)^{1/2} (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_j), \ i = 1, \ldots, n_j, \ j = 1, \ldots, g$$

Calculate the test statistic χ_0^2 based on the Λ_0 using (3.5) in MANOVA of original data, the form is following:

$$\chi_0^2 = -\left(n - 1 - \frac{p+g}{2}\right)\log \Lambda_0 \ \text{ for } H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_g, \ n = \sum_{i=1}^{g} n_j$$

[Step 2] For each group, generate B bootstrap samples by resampling with replacement:

$$\boldsymbol{u}^*_{kij}, \ i = 1, \ldots, n_j, \ j = 1, \ldots, g, \ k = 1, \ldots, B$$

[Step 3] For each k of B bootstrap samples, calculate the mean vectors and overall mean vector:

$$\overline{\boldsymbol{u}^*}_{1j}, \ldots, \overline{\boldsymbol{u}^*}_{Bj}, \ \overline{\boldsymbol{u}^*}_{kj} = \frac{1}{n_j}\sum_{i=1}^{n_j} \boldsymbol{u}^*_{ikj}, \ j = 1, \ldots, g, \ k = 1, \ldots, B$$

[Step 4] Calculate the test statistic $\Lambda^* = \dfrac{|SSE|}{|SSE + SST|}$ in MANOVA for each bootstrap samples:

$$\Lambda^*_{(1)}, \ldots, \Lambda^*_{(B)}$$

[Step 5] Compute the Bartlett's statistic χ_0^* for each bootstrap sample:

$$\chi^{2*}_{0(k)} = -\left(n - 1 - \frac{p+g}{2}\right)\log \Lambda^*_k, \ k = 1, \ldots, B$$

[Step 6] Determine the test statistics:

$$\chi^{2*}_{0(1)}, \ldots, \chi^{2*}_{0(B)}$$

Ordering these Bartlett's statistics as follows:

$$\chi^{2*}_{0(1)} \leq \ldots \leq \chi^{2*}_{0(B)}$$

[Step 7] To determine the upper-tail critical value $\chi^{2*}_{0,critical}$ corresponding to the significance level α, we calculate the $(1-\alpha)$-quantile of the bootstrap distribution of $\chi^{2*}_{0(k)}$:

$$P\left(\chi^2_{0(k)} \leq \chi^{2*}_{0,critical}\right) = 1-\alpha$$

[Step 8] The results are based on whether the test statistic χ^2_0 in [Step 1] exceeds the critical values:

$$\text{If } \chi^2_0 > \chi^{2*}_{0,critical}, \text{ then reject } H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_g \text{ at the } \alpha$$

Similarly, since the multivariate normality assumption and homogeneity of covariance matrices are not satisfied, the bootstrap method for GPA was developed. The algorithm for Bootstrap GPA is provided below. The number of groups is g, the number of variables is p, n is the total number of observations, n_j is the number of observations in the j^{th} group, $\boldsymbol{x}_{ij}$ is the vector for the i^{th} observation in the j^{th} group, and $\overline{\boldsymbol{x}}_j$ is the mean vector for the j^{th} group.

[Step 1] Calculate the $p \times 1$ new vectors of observation for each group:

$$\boldsymbol{u}_{ij} = \left(\frac{n_j}{n_j-1}\right)^{1/2} (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_j), \ i = 1, \ldots, n_j, \ j = 1, \ldots, g$$

Calculate the test statistics $\chi^2_{01}, F_{02}, F_{03}$ based on the original data:

$$\chi^2_{01} = -\left(n - g - \frac{p+1-g}{2}\right)\log\Lambda \quad \text{for } H_{01} : \text{Parallel Profiles}$$

$$F_{02} = \left(\frac{n-g-p+1}{g-1}\right)\frac{1-\Lambda_2}{\Lambda_2} \quad \text{for } H_{02} : \text{Coincident Profiles}$$

$$F_{03} = \frac{n(n-p+1)}{p-1}\,\overline{\boldsymbol{u}}_j^T\ C^T\left[C(SSE)C^T + C(SST)C^T\right] C\ \overline{\boldsymbol{u}}_j \quad \text{for } H_{03} : \text{Level Profiles}$$

[Step 2] For each group, generate B bootstrap samples by resampling with replacement:

$$\boldsymbol{u}^*_{kij}, \ i = 1, \ldots, n_j, \ j = 1, \ldots, g, \ k = 1, \ldots, B$$

[Step 3] For each k of B bootstrap samples, calculate the mean vector, covariance matrix, and joint covariance matrix as follows:

$$\overline{\boldsymbol{u}^*}_{1j} \ldots \overline{\boldsymbol{u}^*}_{Bj}, \ S^*_{1j} \ldots S^*_{Bj}, \ S^*_k = \frac{1}{n-g}\sum_{j=1}^{g}(n_j-1)S^*_{kj}, \ k = 1, \ldots, B$$

[Step 4] Calculate the test statistics $\chi^{2*}_{01}, F^*_{02}, F^*_{03}$ in GPA for each bootstrap sample and order these values:

$$\chi^{2*}_{01(1)} \leq \ldots \leq \chi^{2*}_{01(B)}, \quad F^*_{02(1)} \leq \ldots \leq F^*_{02(B)}, \quad F^*_{03(1)} \leq \ldots \leq F^*_{03(B)}$$

[Step 5] To determine the upper-tail critical values $\chi^{2*}_{01,critical}, F^*_{02,critical}, F^*_{03,critical}$ corresponding to the significance level α, we calculate the $(1-\alpha)$-quantile of the bootstrap distributions of $\chi^{2*}_{01(k)}$ and $F^*_{02(k)}, F^*_{03(k)}$-distributions:

$$P\left(\chi^{2*}_{01(k)} \leq \chi^{2*}_{01,critical}\right) = 1-\alpha,\; P\left(F^{*}_{02(k)} \leq F^{*}_{02,critical}\right) = 1-\alpha,\; P\left(F^{*}_{03(k)} \leq F^{*}_{03,critical}\right) = 1-\alpha$$

[Step 6] The results are based on whether the test statistics exceed the critical values:

If $\chi^2_{01} > \chi^{2*}_{01,critical}$, then reject $H_{01} : \mu_{1i} - \mu_{1,i+1} = \ldots = \mu_{gi} - \mu_{g,i+1},\; i = 1, 2, \ldots, p-1$ at the α

If $F_{02} > F^{*}_{02,critical}$, then reject $H_{02} : \mu_{1i} = \ldots = \mu_{gi},\quad i = 1, 2, \ldots, p$ at the α

If $F_{03} > F^{*}_{03,critical}$, then reject $H_{03} : \mu_{11} = \mu_{12} = \ldots = \mu_{1p} = \mu_{21} = \ldots = \mu_{gp}$ at the α

4. Illustrations

As noted in Section 3.1, for the data to be suitable for performing MANOVA and GPA the assumption of multivariate normality must be satisfied.

From Table 3, it is evident that Mardia's skewness statistic rejects the null hypothesis of multivariate normality within every cluster. This implies that the data distribution deviates from the assumptions required for MANOVA and GPA. This finding prompts the use of bootstrap methods.

Table 3: Results of Mardia's Test

Cluster	Test	Statistic	p-value	Result
	Skewness	305.19	< 0.00	reject
Central	Kurtosis	0.16	0.88	accept
	MVN Test	-	-	reject
	Skewness	340.19	< 0.00	reject
Southern	Kurtosis	1.41	0.16	accept
	MVN Test	-	-	reject
	Skewness	328.15	< 0.00	reject
Western	Kurtosis	1.74	0.08	accept
	MVN Test	-	-	reject
	Skewness	324.19	< 0.00	reject
Eastern	Kurtosis	0.71	0.48	accept
	MVN Test	-	-	reject

Furthermore, Box's M test was performed to assess whether the covariance matrices across the four clusters are homogeneous.

Table 4: Results of Box's M Test

χ^2	df	p-value
1807.80	135	< 0.00

From Table 4, the result of Box's M test shows a test statistic of 1807.8 with a p-value significantly lower than the $\alpha = 0.05$ level of significance. Therefore, the null hypothesis $H_0 : \Sigma_c = \Sigma_s = \Sigma_w = \Sigma_e = \Sigma$ is rejected.

In summary, the results from both Mardia's test and Box's M test indicate that the data do not satisfy the assumptions of multivariate normality and homogeneity of covariance matrices, which are essential for conducting MANOVA and GPA.

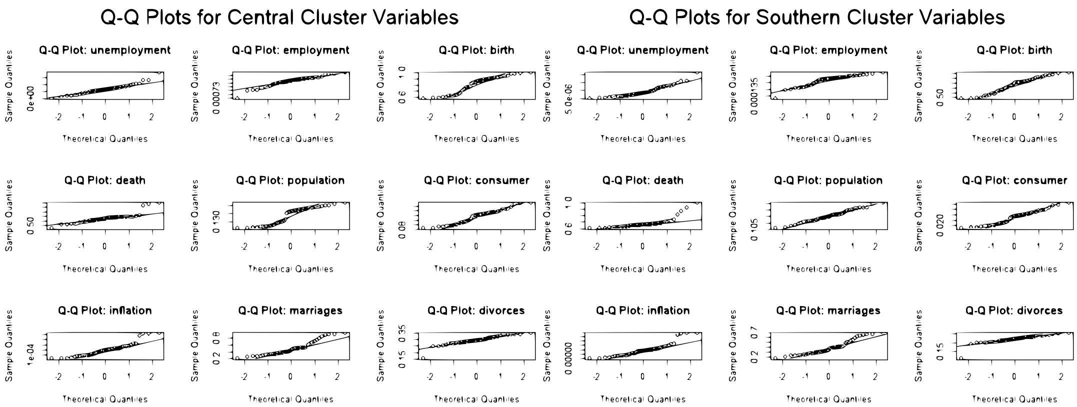
Table 5: Shapiro–Wilk Test p-values for Each Variable by Regional Cluster

Variable	Central	Southern	Western	Eastern
Unemployment	0.03	0.00	0.00	0.00
Employment	0.02	0.00	0.00	0.16
Births	0.00	0.06	0.54	0.44
Deaths	0.00	0.00	0.00	0.00
Population	0.00	0.86	0.72	0.00
Consumer	0.08	0.05	0.02	0.01
Inflation	0.00	0.00	0.00	0.00
Marriages	0.00	0.00	0.00	0.02
Divorces	0.00	0.00	0.00	0.03

To gain a more nuanced understanding of the distributional properties of every variable within each regional cluster, we applied the Shapiro–Wilk test to each variable individually.

The test revealed that the majority of variables significantly deviate from normality across all clusters. These results support the earlier findings from Mardia's test, demonstrating that the assumption of multivariate normality is violated in the data.

In addition to Mardia's and Shapiro–Wilk tests, Figure 6 presents the Q–Q plots for each variable in the clusters. Across all clusters, most variables exhibit substantial deviations from normality.



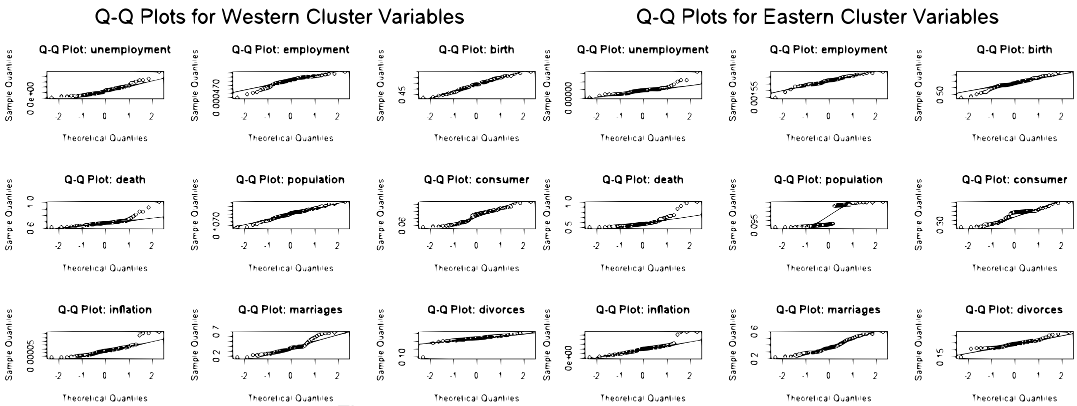


Figure 6: *Q–Q Plots for Clusters Variables*

The results of the following standard MANOVA tests indicate highly significant differences across the clusters.

Table 6: MANOVA Results Based on Pillai's Trace and Hotelling–Lawley Test

Test	Df	Statistic	Approx. F	Num Df	Den Df	p-value
Pillai's Trace	3	2.4634	92.824	27	546	$< 2.2 \times 10^{-16}$
Hotelling-Lawley Trace	3	3517.9	23279	27	536	$< 2.2 \times 10^{-16}$

The validity of these multivariate tests relies on two key assumptions: multivariate normality and homogeneity of covariance matrices. As demonstrated earlier by Mardia's test, Shapiro–Wilk tests, Q–Q plots, and Box's M test, both assumptions are violated. Consequently, a Bootstrap MANOVA was applied as a more robust alternative to the traditional MANOVA when its assumptions were not met.

Table 7: Results of Bootstrap MANOVA

Statistic	Value
Λ_0	1.18×10^{-5}
χ_0^2	2201
χ^{2*}	36.41

From Table 7, the test statistic for the Bootstrap MANOVA with Λ_0 is calculated as follows:

$$\chi_0^2 = -\left(192 - 1 - \frac{4+9}{2}\right)\log(0.00001) = 2200.6 > \chi^{2*}_{9(4-1)}(0.05) = 36.41$$

$$\Rightarrow H_0 : \boldsymbol{\mu}_c = \boldsymbol{\mu}_s = \boldsymbol{\mu}_w = \boldsymbol{\mu}_e \text{ is rejected at the 5\% level of significance.}$$

The histogram in Figure 7(a) represents the distribution of Wilks' Lambda values (Λ) calculated from 1000 bootstrap samples. The near-symmetry suggests similarity to a theoretical χ^2 distribution. Figure 7(b) illustrates the distribution of the Bartlett's test statistic obtained from the same 1000 bootstrap resamples. Because the Bartlett's statistic is a transformation of Wilks' Lambda, its empirical shape conforms to the χ^2 distribution, consistent with theoretical expectations.

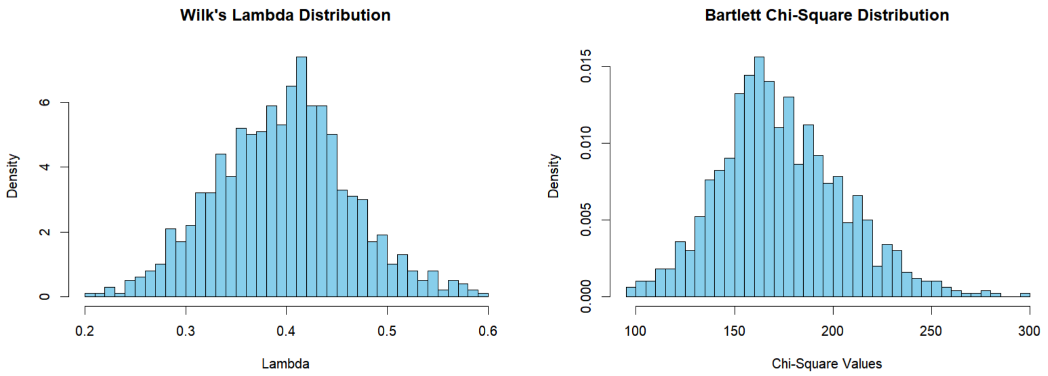


Figure 7: *(a) Lambda Distribution, (b) Bartlett's χ^2 Distribution*

To assess the robustness of the multivariate test statistic, we applied a bootstrap procedure that estimates the distribution of Wilks' Lambda from 1000 resamples. The mean bootstrap estimate was 0.3964, with a 95% confidence interval of $\Lambda \in (0.27, 0.53)$. Because these values are well below 1, they indicate substantial differences among the clusters.

Next, we conducted the bootstrap GPA, using the following contrast matrix C:

$$C = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

Three types of hypothesis tests were provided:

- H_{01} (Parallel Profiles): Tests whether the profiles are parallel across clusters;
- H_{02} (Coincident Profiles): Tests whether the profiles are consistent across clusters;
- H_{03} (Level Profiles): Tests whether the profiles have the same level across clusters.

The results of the hypothesis tests are summarized in the following table:

Table 8: Results of Bootstrap Profile Hypothesis Testing

Step	Hypothesis	Test Stat	p-value	Result
Step 1	H_{01}: parallelism	$\chi^2_{01} = 0.00009$	0.00	reject
Step 2	H_{02}: consistency	$F_{02} = 914.83$	0.00	reject
Step 3	H_{03}: levelness	$F_{03} = 5738.04$	0.00	reject

From Table 8, the null hypothesis H_{01} is rejected, indicating that the profiles are not parallel across clusters. Because parallelism fails, the profiles cannot be considered similar in shape. The null hypothesis of coincident profiles H_{02} is also rejected, showing that the cluster profiles are not coincident. Finally, $F_{03} = 5738.04$ with a p-value of 0.00 leads to rejection of the null hypothesis of equal levels H_{03}, hence, the clusters differ significantly in overall level.

5. Conclusion

This paper presents a detailed analysis of socio-economic and demographic variables across four clusters in Russia, revealing substantial regional disparities and diverse economic conditions.

It can be concluded that the data violate the assumptions of multivariate normality and homogeneity of covariance matrices, as confirmed by the results of Mardia's and Box's M tests. This necessitated the use of bootstrap methods to ensure reliable and valid inference across clusters. First, the results highlight significant socio-economic imbalances. The Western cluster demonstrates a relatively favorable economic environment and may serve as a model for balanced regional development. In contrast, the Eastern and Southern clusters face persistent challenges, including high unemployment, underutilized labor potential, and elevated living costs due to harsh natural conditions, underscoring the need for region-specific strategies. Second, as Srivastava (2002) emphasizes, verifying statistical assumptions is essential for the credibility of multivariate analysis. The application of Mardia's and Box's M tests in this study ensured methodological robustness. This statistical validation improves the accuracy of data interpretation and strengthens the credibility of the resulting policy recommendations. Since violations of these assumptions can substantially affect the validity of MANOVA and GPA results, bootstrap-based MANOVA and GPA were employed. These methods revealed statistically significant differences in socio-economic and demographic characteristics among the four clusters. The clusters are distinct in their profiles, with noticeable differences in the direction and level of socio-economic indicators. Third, high inflation and its positive correlation with unemployment – particularly in the Eastern cluster – suggest broader macroeconomic instability (Ilyashenko, 2016), again highlighting the need for targeted, regionally sensitive policy measures.

Finally, the findings emphasize the importance of tailored regional strategies to reduce disparities and promote more balanced socio-economic development. Policymakers are encouraged to pursue initiatives such as industrial diversification and year-round employment opportunities beyond the agro-industrial sector.

References

Choi YS (2018). *Multivariate Data Analysis with R*. Kyungmoon Publishers.

Choi YS (2019). *Estimation and Test for Multivariate Data Analysis with R*. Kyungmoon Publishers.

Choi YS (2021). *Shape Analysis with R*. Pusan National University Press and Culture Center.

Ilyashenko VV (2016). The Relationship between Inflation and Unemployment: Theoretical Aspects and Particularities of Manifestation in the Russian Economy. *Economy of Region*, **64**(2), 5–11.

Jolliffe IT (2002). *Principal Component Analysis* (2nd ed.). Springer-Verlag, New York.

Srivastava MS (2002). *Methods of Multivariate Statistics*. John Wiley & Sons, New York.

Received February 12, 2021; Revised April 07, 2021; Accepted May 17, 2021