MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
SUMY STATE UNIVERSITY
UKRAINIAN FEDERATION OF INFORMATICS

PROCEEDINGS

OF THE VI INTERNATIONAL SCIENTIFIC CONFERENCE

ADVANCED INFORMATION SYSTEMS AND TECHNOLOGIES

AIST-2018

(Sumy, May 16–18, 2018)

SUMY
SUMY STATE UNIVERSITY
2018

Regional Sustainability Assessment Through Multivariate Statistical Analysis

Tetiana Marynych[1], Stanislav Smolenko
Sumy State University, Ukraine, [1]t.marynych@ssu.edu.ua

***Abstract* – The work assesses regional differentiation based on the extended and classical variables' selection using multivariate statistical techniques: cluster and principal component analysis.**

Keywords – Cluster Analysis, Principal Component Analysis, Sustainable Regional Development.

I. INTRODUCTION

The concept of sustainable and inclusive development (SID) has been widely discussed in recent decades and has been recognized as the priority policy areas both at the regional and international levels [1]-[2]. Their successful realization requires proper identification of the main indicators, their assessment, and monitoring. Modern researchers extend classical approach that focuses on the issues of poverty, inequality, unemployment, and growth by the variables that evaluate education, institutions, health, environmental sustainability and informality [3]-[4]. Methodology varies from aggregate indexing and causality modeling to regional, international structuring using multivariate statistical techniques, e. g. clustering, factor and principal component analysis [3]-[5].

The main objective of this paper is to make regional differentiation based on the extended and classical variables' selection using cluster analysis (CA) and principal component analysis (PCA).

II. DATA AND METHODOLOGY

Our research explores 22 economic, financial, social and environmental indicators for 24 Ukrainian regions. Regional codes used in the analysis are given in Table I.

TABLE I. DESCRIPTION OF REGIONAL CODES

Region	Code	Region	Code
Vinnytsia	VN	Mykolaiv	MK
Volyn	VO	Odessa	OD
Dnipropetrovsk	DP	Poltava	PT
Donetsk	DN	Rivne	RV
Zhytomyr	ZH	Sumy	SM
Zakarpattia	ZA	Ternopil	TP
Zaporizhia	ZP	Kharkiv	KH
Ivano-Frankivsk	IF	Kherson	KS
Kiev	KV	Khmelnytskyi	KM
Kirovohrad	KG	Cherkasy	CH
Luhansk	LG	Chernivtsi	CS
Lviv	LV	Chernihiv	CN

We calculated variables as relative indicators using average values for 2013-2016 years, based on official data of State and regions' statistics offices [6]. It should be noted that due to the data unavailability the values of GRP for 2016 were estimated on the basis of the 2015 values, multiplied by the index of industrial production in 2016.

Table II provides a description of the indicators used in this multivariate analysis.

TABLE II. DESCRIPTION OF VARIABLES

Indicators	Code
1. Social and demographic indicators	
The ratio of the regional population to the total population of Ukraine	ppl
The ratio of the population over the age of 60 to the economically active population of working age	old
Total fertility rate (per woman)	br
The share of urban population	ur
The ratio of the average life expectancy at birth to the average in Ukraine	ls
The number of offences in the region per 1000 population (crime rate)	off
The ratio of average monthly wages by region to the average monthly wage in Ukraine	sal
2. Economic and financial indicators	
Terms of trade (the ratio of exports to imports)	tr
The ratio of direct foreign investments to GRP	di
The ratio of capital investments to GRP	ci
The ratio of gross regional product (GRP) to gross domestic product (GDP) of Ukraine	grp
The ratio of local budget expenditures to GRP	exp
The ratio of local budgets' incomes to their expenditures	inc
Spread between the credit interest rates of banking institutions and the National Bank discount rate	ir
The ratio of loans to GRP	loa
The ratio of deposits to loans	dep
Consumer price index (compared to December of the previous year)	pi
3. Environmental indicators	
CO2 emissions per GRP	poll
The volume of wood harvesting per unit of area	wd
The extraction of water bioresources per 1000 population	wt
The amount of waste per capita	wst
The amount of recycled waste per capita	ut

PCA can substitute a data set with a smaller number of representative variables that explain most of the variability in the original data [8]. PCA as a dimension reduction tool is a common strategy to deal with the problem of correlation dependence between factors, as well as the problem degrees of freedom (correspondence of the number of variables to the number of observations) in small samples. Principal components represent linear combinations of the original variables weighted by their contribution to explaining the variance of the variables. A matrix of the factor loadings represents a system of the following equations [8]:

$$PC_i = W_{1i} \cdot X_1 + W_{2i} \cdot X_2 + ... + W_{3i} \cdot X_3, \qquad (1)$$

where PC_i – i-principal component; X_k – a k-feature, W_{ik} – the corresponding loading (weight).

Principal components' calculation applies a singular value decomposition of the centered and scaled data matrix Centering and scaling are performed using the following formula:

$$x_{ij} = \frac{x_{ij} - \mathrm{M}(X_i)}{\sigma(X_i)}. \qquad (2)$$

Here X_i – is a vector of corresponding features; x_{ij} – a particular value of i-feature of j-element; $\mathrm{M}(X_i)$ – is a mean of X_i; $\sigma(X_i)$ – a standard deviation of X_i.

Generated principal components and a matrix of the estimated factor loadings are used further in the regional structuring based on the clustering techniques. Our research employs the K-means algorithm based on the recommendations given in a "Typology of the regions of Ukraine", prepared within the framework of the EU project on supporting regional development policy in Ukraine [2]. This method implies partitioning observations into k groups by minimizing the sum of squares from the points to the assigned cluster centers [8]:

$$\sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 \to \min. \qquad (3)$$

Here S_i represents i-cluster; x_j – j-element coordinates; μ_i – coordinates of the i-cluster's centroid (mean of the coordinates of the i-cluster's elements).

III. ANALYSIS

In this section we present the main findings of our research. Fig. 1 shows numerous correlation dependences between variables.

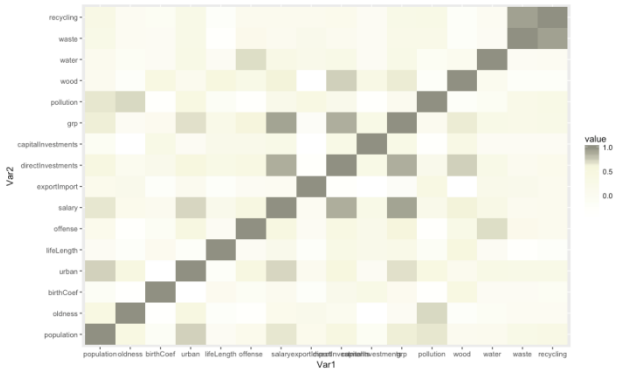

Figure 1. Correlation matrix
Source: Own processing based on official statistics [6]

Table III summarizes the importance of the principal components and demonstrates six most disperse components, which proportion of variance is greater than 0.05, total cumulative proportion of variance 0.885.

TABLE III. PRINCIPAL COMPONENT ANALYSIS SUMMARY

Region	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.197	1.934	1.412	1.326	0.961	0.952
Proportion of variance	0.302	0.234	0.125	0.11	0.058	0.057
Cumulative proportion	0.302	0.536	0.66	0.77	0.828	0.885

Fig. 2 exhibits associations between the diversity of the regions of Ukraine and the first two principal components displayed as red vectors' projections according to the weight and direction of the factor loadings. This graph does not show the factors "*ppl*", "*ut*", "*off*", as these vectors overlap and blind the visual perception.

The first principal component reflects economic and financial development of the regions. It is mostly affected by the following economic indicators: the ratio of local budgets' revenues to their expenditures ("*inc*"); the average salary in the region ("*sal*"); the share of GRP in GDP of Ukraine ("*grp*"); the ratio of loans to GRP ("*loa*"); the ratio of local budget expenditures to GRP ("*exp*") (negatively).

The second principal component characterizes human, natural and capital resources of the regions. Main contributions are made by the following factors: the ratio of capital investments to GRP ("*ci*"); the ratio of people aged 60+ to the economically active population ("*old*") (negatively); the birth rate ("*br*"); the difference between the credit interest rates of banking institutions and the National Bank discount rate ("*ir*") (negatively); the volume of wood harvesting per unit of area ("*wd*"). The absolute weights of these loadings are higher than 0.3.

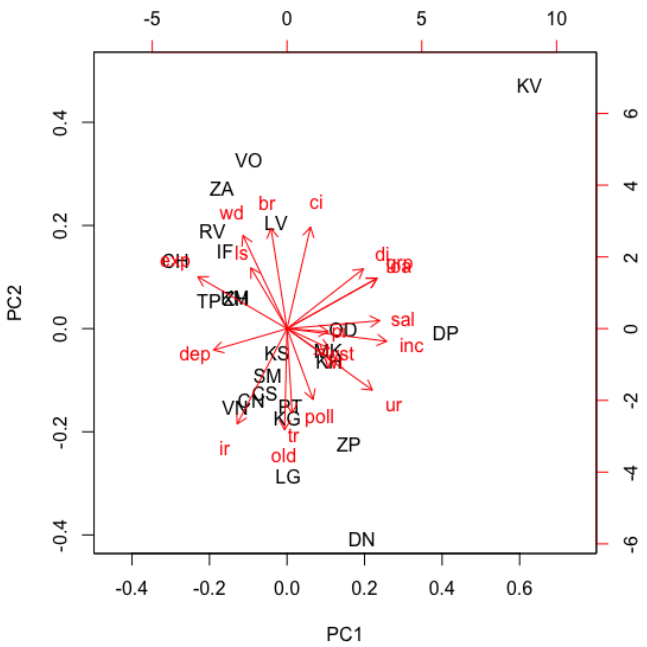

Figure 2. PCA visualization
Source: Own processing based on official statistics [6]

The third principal component accounts for ecological, monetary stability and the rule of law. This component is influenced primarily by the ratio of CO2 emissions to GRP ("*poll*") (negatively); the crime rate ("*off*"); the consumer price index ("*pi*"); the water usage ("*wt*"). Notably, the fourth principal component also applies to the sustainable development of the regions. It is mostly affected by two variables – the amount of waste per person ("*wst*") and the amount of recycled waste per person ("*ut*"). Both factors have a near 0.5-weight impact with a negative sign that declares a reverse relationship between waste-related factors and the level of region's sustainable development.

The fifth principal component is mainly determined by the human capital indicators, i.e. by the birth rate ("*br*"), the ratio of people aged 60+ to the economically active population ("*old*") (negatively), the ratio of the average life expectancy at birth to the average in Ukraine ("*ls*") (negatively). Besides, the extraction of water bioresources per 1000 population ("*wt*"), terms of trade ("*tr*"), also have absolute loading values greater than 0.35.

The sixth principal component duplicates the loadings of the fifth component ("*ls*", "*wt*", "*tr*",), adding the price index ("*pi*") and the ratio of capital investments to GRP ("*ci*"), which have positive loading values of 0.3.

The findings of the principal component analysis show that variables of human capital, natural resources usage and monetary and financial stability prevail in the choice of principal components. At the same time such factors as regional population ("*ppl*"), urban population ("*ur*"), direct foreign investments ("*di*"), and the ratio of deposits to loans ("*dep*") do not particularly affect any of the first six principal components.

For cluster analysis, we take the first six main components that summarize the main features of the original variables. As mentioned above, we apply the *K*-means clustering algorithm, setting k=5 based on the typology of the regions of Ukraine [2], findings of Ukrainian economists [5] and our own empirical assessments. Fig. 3 exhibits the results of the regional clustering within the estimated coordinates of the first two principal components that represent economic, financial, human capital and sustainable development of the regions. Here we use basic geometric structures [9] for simplicity.

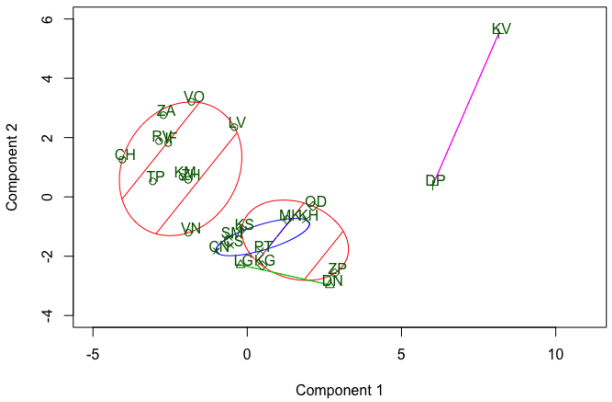

Figure 3. *K*-means clustering
Source: Own processing based on official statistics [6]

Fig 4 highlights the groups (clusters) of regions of Ukraine based on the results of the cluster analysis:

Group 1 (blue): Vinnytsia, Volyn, Zhytomyr, Zakarpattia, Ivano-Frankivsk, Lviv, Rivne, Ternopil, Khmelnytskyi and Chernivtsi regions.

Group 2 (violet): Donetsk and Luhansk regions.

Group 3 (red): Dnipropetrovsk region, Kyiv region including Kyiv.

Group 4 (orange): Chernihiv, Sumy, Cherkasy, Poltava and Kharkiv regions.

Group 5 (green): Kirovohrad, Kherson, Mykolaiv, Zaporizhia and Odessa regions.

Obviously, the regions with the most developed cities (Kyiv and Dnipro) turn up to be in the same group. Donetsk and Luhansk as occupied territories have significant reasons to differ from other regions a lot. Surprisingly, all the other regions of Ukraine are grouped based on the first two principal components in conformity with their geographical location, creating Western, Northern, Eastern and Southern clusters.

Our results partly correspond with the typology of the regions of Ukraine [2] constructed in 2012 on the basis of geographical, socio-demographic, economic, ecological indicators, that focuses on the sustainability issues as well (Fig. 5).

Figure 4. Mapping the clustering results
Source: Own processing based on official statistics [6]

Figure 5. Mapping the clustering results
Source: [6]

The differences are explained by the choice of the original variables and the time period used in the analysis. We can see that recent changes in political and economic development of Ukraine provoked deepening of the differentiation between the western border regions and Northeastern regions, and deteriorating of the sustainability of the majority of the regions.

CONCLUSIONS

In this work, we have investigated the factors that determined regional sustainability and inclusiveness. Following up the comprehensive modern approach, we extended the traditional set of variables with the factors that influence financial stability, human capital development, and sustainability goals' achievement.

We estimated relative standardized indicators to avoid possible differences in the size of regions and scales of measurements, removing abnormal observations. Multivariate statistical tools were applied (PCA and clustering) to reveal the simplified patterns of possible relationships and data structuring.

It was proved that economic factors explain a dominant part of sustainable regional development, explaining 30% of the cumulative variance of the data. At the same time, our findings showed that human capital and ecological factors turned to be the most impactful factors that account for the rest 58% of the cumulative variance of the first six principal components of the observed data set.

We used the principal components, which describe the most important features of the data to make comparative cluster analysis of the regions of Ukraine. It was found that clusters were grouped by geographic position of the regions. In particular, empirical results revealed that Western regions had better indicators of the human capital (fewer people of the retirement age, higher life expectancy, lower urbanization and crime rates) and sustainable development (better use of water resources; lower waste volumes, but the worse situation with the deforestation). Central and Southern regions on average had better economic and financial position, while Eastern regions showed the biggest problems with human capital and environmental protection.

To conclude, this work contributes to solving the problem of the assessment and monitoring of the regional development, emphasizing on the importance of the elaboration of the multidimensional frameworks that account for a broad group of indicators that depict new challenges in terms of sustainability, inclusiveness and economic freedom. Short-term and long-term aspects of sustainable regional development and possible associations between variables can be further explored using methods of linear regression, linear discriminant analysis or artificial networks.

REFERENCES:

[1] "Draft Outcome Document of the United Nations Summit for the Adoption of the Post-2015 Development Agenda", A/69/L.85, 2015.

[2] "Regional Development and State Regional Policy in Ukraine: Analytical Report". [Online]. Available: http://surdp.eu/Analytical-Report. [Accessed April 12, 2018]. (in Ukrainian)

[3] R. M. Aguilar, "Assessment of socio-economic development through country classifications: a cluster analysis of the Latin Americaandthe Caribbean (LAC) and the European Union (EU)", *World Economy Journal*, No. 47, pp. 43-64, 2017.

[4] T. O. Marynych, "Empirical assessment of the long-term aspects of sustainable regional development", *Economic Annals XXI*, No. 166(7-8), pp. 86-90, 2017.

[5] O. A. Ryadno and Berkut O. V. "Study of structure and dynamics of differentiation of regional socio-economic development on the basis of cluster analysis", *Donbass Economic Bulletin*, No. 1(43), pp. 60-67, 2016. (in Ukrainian)

[6] "State Statistics Service of Ukraine (2013-2016)". *Statistical data*. [Online]. Available: http://ukrstat.gov.ua/. [Accessed April 12, 2018]. (in Ukrainian)

[7] "The comprehensive R Archieve Network". [Online]. Available: https://cran.r-project.org/. [Accessed April 12, 2018].

[8] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statisitcal Learning with Applications in R*. NY: Springer, 2013.

[9] Z. Gniazdowski. "New Interpretation of Principal Components Analysis", *Zeszyty Naukowe WWSI*, No 16, Vol. 11, pp. 43-65, 2017.