www.ees.uni.opole.pl
ISSN paper version 1642-2597
ISSN electronic version 2081-8319
Economic and Environmental Studies

Vol. 18, No 2 (46/2018), 747-761, June 2018

On the application of selected exploratory data analysis methods to assess differences in the level of sustainable development in the environmental domain of voivodships in Poland

Małgorzata MISZTAL

University of Łódź, Faculty of Economics and Sociology, Department of Statistical Methods, Poland

Abstract: The purpose of the study was to assess the differences in the level of sustainable development in environmental domain of voivodships in Poland in the years 2008-2015. Seven indicators belonging to 3 areas of the environmental domain (energy, air protection, waste management) were analysed with the use of principal component analysis (PCA) and the between-class PCA. The results revealed large differences between voivodships mainly due to the level of air pollutants emissions from plants especially noxious to air purity. It has also been shown that in the years 2008-2015 a visible increase in the outlays on fixed assets serving environmental protection and development of ecological awareness of society was observed.

Keywords: sustainable development, environmental domain, multivariate exploratory data analysis, principal component analysis

JEL codes: C02, C38, R11

DOI: https://doi.org/10.25167/ees.2018.46.17

Correspondence Address: Małgorzata Misztal, Department of Statistical Methods, Institute of Statistics and Demography, Faculty of Economics and Sociology, University of Łódź. ul. Rewolucji 1905 r. 41, 90-214, Łódź, Poland. Tel.: +48426355178. E-mail: mmisztal@uni.lodz.pl

1. Introduction

Sustainable social and economic development is one of the most important challenges of the modern world. According to GUS (2011: 4): “sustainable development of the country, accepted as a Constitutional Principle of the Republic of Poland, has been defined in the Law on Environmental Protection as such a socio-economic development, in which the process of integrating political, economic and social actions occurs, taking into account maintenance of the equilibrium of nature and stability of basic natural processes, to guarantee the possibility of fulfilling basic needs of societies or citizens not only of the present generation, but for future generations as well.”

A lot of indicators have been proposed to assess the level of sustainable development in social, economic, environmental and institutional-political domains. This means that a multivariate approach is needed to analyse such data. As Balicki (2009: 15) points out: “most of the statistical data are multidimensional in nature. This means that both the objects of the statistical population as well as the examined phenomena are described by means of many different, usually dependent features (…) The use of statistical methods cannot therefore be limited to simple univariate analyses. Analyses of large and complex datasets require the use of multivariate statistical analysis methods.”

In recent years, many researchers have proposed to use a number of different multidimensional statistical analysis methods to evaluate the level of sustainable development in Poland (see e.g.: Bal-Domańska and Wilk, 2011; Bal-Domańska, Wilk and Bartniczak, 2012; Roszkowska and Karwowska, 2014; Fura, 2015; Roszkowska and Filipowicz-Chomko, 2016; Drabarczyk, 2017; Łuczak and Kurzawa, 2017). The most popular multivariate statistical methods used in analyses were: linear ordering with a common development pattern, synthetic measure of development and the TOPSIS method.

The objective of the study was to assess differences in the level of sustainable development in environmental domain of voivodships (provinces) in Poland in the years 2008-2015 with the use of selected exploratory data analysis methods.

2. Data sources and methods

For the purpose of the study the data made available by the Central Statistical Office through the Sustainable Development Indicators Application (regional module) were used. Datasets for the voivodships cover the years 2004-2016, but due to their incompleteness only time series from 2008-2015 were made use of. The environmental domain is characterized by 17 indicators belonging to 7 areas (Climate change, Energy, Air Protection, Fresh water resources, Land use, Biodiversity, Waste Management). The study finally used seven indicators belonging to 3 areas, presented in Table 1.

Table 1. Selected indicators used in the analysis

Area	Description	
Energy	**X1**	Renewable energy sources in percentage of the total production of electricity (%)
	X2	Electricity consumption per 1 million PLN GDP (GWh)
	X3	Outlays on fixed assets serving environmental protection: energy saving per capita (PLN)
Air protection	**X4**	Emissions of air pollutants (gases) from plants especially noxious to air purity (t/y)
	X5	Emissions of air pollutants (particulates) from plants especially noxious to air purity (t/y)
Waste management	**X6**	Municipal waste collected separately during the year in relation to the total municipal waste (%)
	X7	Mixed municipal waste from household collected during the year per capita (kg)

Source: Author's own elaboration.

The choice of these indicators was dictated mainly by the data availability for individual voivodships. Additionally, due to the objective of the analysis, i.e. the attempt to assess the changes over time on the level of the voivodships, some indicators that remained more or less stable during the years 2008-2015 were omitted (these were among others the indicators belonging to the land use and biodiversity areas with the coefficients of variation less than 5%).

In order to assess the differences in the level of sustainable development in environmental domain taking into account the time (8 years: 2008-2015) and space (16 voivodships), an approach using exploratory data analysis methods was proposed. According to Everrit and Skrondal (2010: 157), exploratory data analysis can be defined as "an approach to data analysis that emphasizes the use of informal graphical procedures not based on prior assumptions about the structure of the data or on formal models for the data." The objective of such an analysis is to detect structures and general patterns in relations between different variables (e.g. socio-economic indicators), as well as the description and classification of objects characterized by these indicators in multidimensional spaces, with minimal use of formal mathematics or statistical methods.

One of the most common multivariate exploratory statistical methods used in socio-economic studies is principal component analysis (PCA; Pearson, 1901; Hotelling, 1933). Principal component analysis can be described as a procedure transforming the original variables into new ones that are uncorrelated and account for decreasing proportions of the variance in the data. These new variables, called the principal components, are defined as linear combinations of the original variables. PCA emphasized correlation structures between variables and provides an ordination of objects that underlies this correlation structure (see: Thioulouse et al. 2015: 85). Taking into account (in the presented study) the years 2008-2015, it is possible to apply 8 separate PCAs, one for every year or one PCA after concataining all datasets.

Taking the existence of groups of samples in a data table into consideration, Thioulouse et al. (2015: 141) suggest using a particular type of analysis, called between-class analysis, which models the differences between groups by computing the group means and analyses the resulting table. The between-class analysis aims at visually checking the existence of groups and describing the main characteristics of the differences between the groups. In the study, the between-class analysis based on principal component analysis was applied.

Generally speaking, the basis for the between-class analysis is the table of group means. In the presented research, two analyses may be performed: (1) groups correspond to individual years (the values of each variable are averaged across the 16 voivodships) and (2) groups correspond to individual voivodships (the values of each variable are averaged across the 8 years). A more formal description of these methods is presented by Thioulouse et al. (2015).

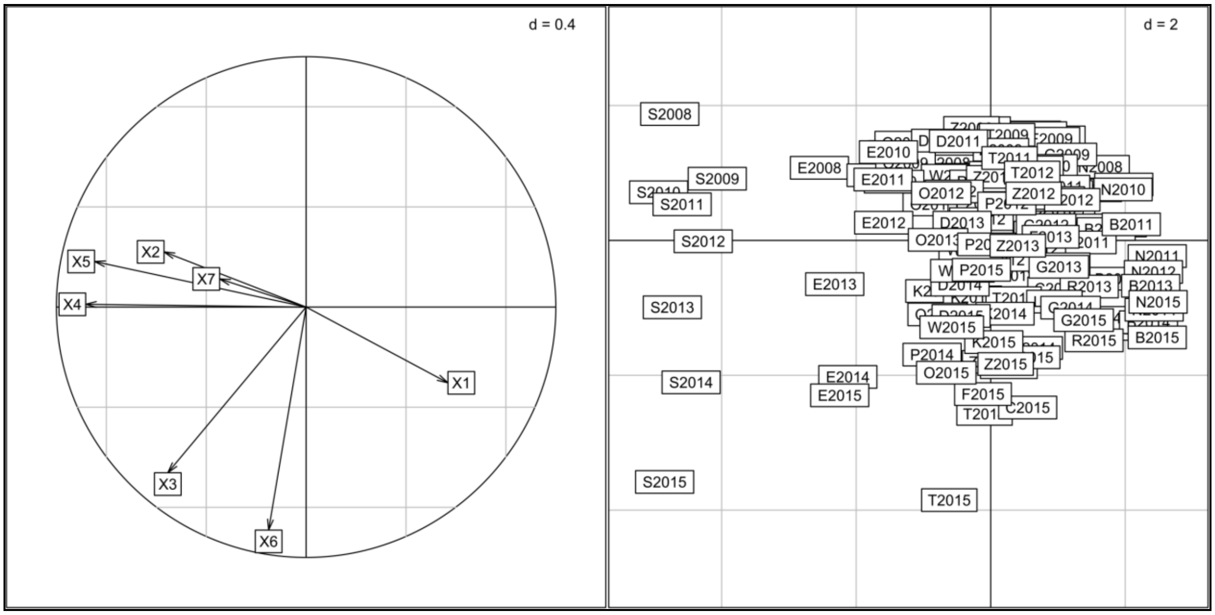
The advantages of these methods include, among others, the possibility of graphical presentation of the analysis results in two-dimensional space using scatterplots and *biplots*. The term "biplot" was introduced by Gabriel (1971) and the simplest definition of this term was given by Gower et al. (2015: 42): "A biplot is exactly what it says. It is a plot of two kinds of information displayed together. The 'bi' in biplot refers to the two kinds of information and not to the usual, but not necessary, use of two dimensions.” Graphical presentation using scatterplots and biplots plays an important role in the interpretation of PCA and between-class PCA results.

All the calculations were performed with the use of the R-environment (package ade4, v. 1.7-6).

3. Results

First of all, the results of a simple PCA performed on the combined dataset are presented (Figure 1). The first three principal components explain 72.5% of the total variability. Two of them were used for further analysis and presented graphically. Variables are presented as vectors (left panel of Figure 1) and objects (voivodships) as points (the right panel of Figure 1).

Figure 1. PCA results for combined data (left panel: variables correlation circle; the right panel: objects factor map)



Legend:
variables: X1 – X7: environmental domain indicators described in Table 1;
objects: D – Dolnośląskie Voivodship; C – Kujawsko-Pomorskie Voivodship; L – Lubelskie Voivodship; F – Lubuskie Voivodship; E – Łódzkie Voivodship; K – Małopolskie Voivodship; W – Mazowieckie Voivodship; O – Opolskie Voivodship; R – Podkarpackie Voivodship; B – Podlaskie Voivodship; G – Pomorskie Voivodship; S – Śląskie Voivodship; T – Świętokrzyskie Voivodship; N – Warmińsko-Mazurskie Voivodship; P – Wielkopolskie Voivodship; Z – Zachodniopomorskie Voivodship.

Source: Author’s own elaboration.

The angles between all vectors on the PCA correlation circle reflect their linear correlations. The approximated correlation between two variables is equal to the cosine of the angle between the corresponding vectors. The perpendicular vectors indicate the lack of correlation between the variables they represent. The angle less than 90^o suggests a positive correlation between variables and the angle approaching 180^o – a strong negative correlation between variables. The direction of the vector corresponds to the direction of the highest variability of a given variable and its length is proportional to the meaning of this variable.

The following correlations can be observed, among others, on the correlation circle in Figure 1:

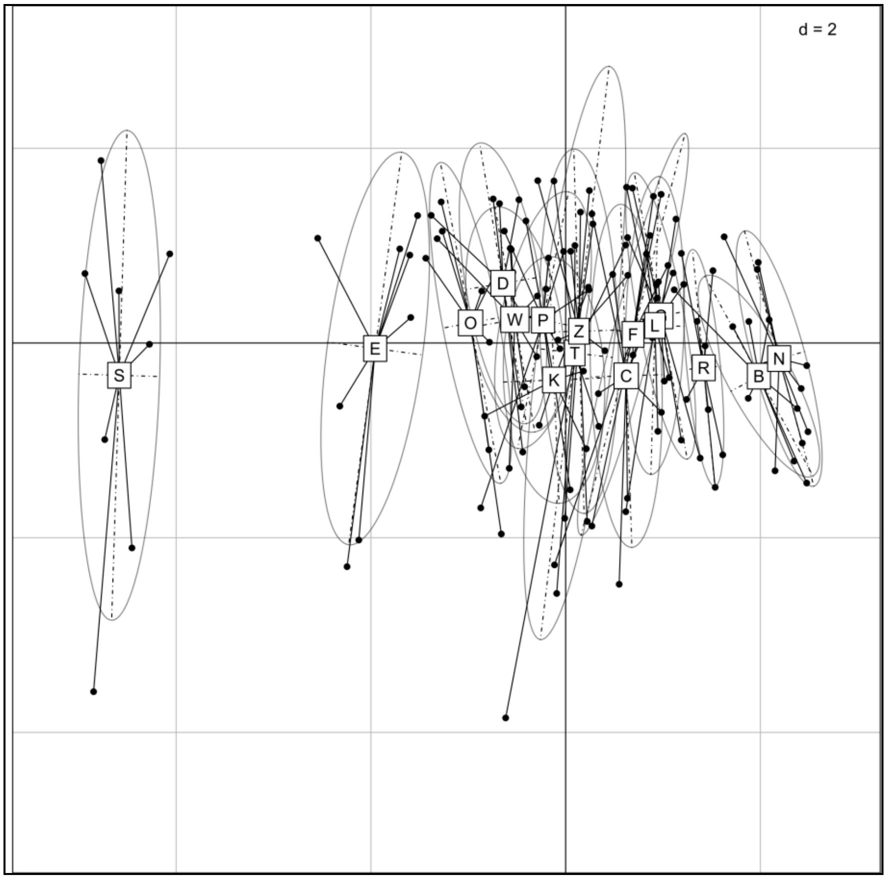
- Positive correlations between X2, X4, X5 and X7. The correlation between X4 and X5 indicates that with the increase in emission of gas pollutants (X4) the emission of particulate pollutants (X5) also increases. The higher the emission of pollutants, the greater the electricity consumption (X2). In addition, the amount of the mixed municipal waste from household collected during the year per capita (X7) increases along with the increase in air pollution and electricity consumption; such a result seems to be characteristic of highly urbanized areas rich in industrial plants.
- Strong negative correlation between X1 (renewable energy sources in percentage of the total production of electricity) and the electricity consumption (X2) and the emissions of air pollutants (X4, X5).
- Positive correlation between X3 (outlays on fixed assets serving environmental protection: energy saving per capita) and X6 (municipal waste collected separately during the year in relation to the total municipal waste).
- The lack of correlation between X6 (municipal waste collected separately during the year in relation to the total municipal waste) and the electricity consumption (X2) and the emissions of air pollutants (X4, X5).
- The lack of correlation between X3 (outlays on fixed assets serving environmental protection: energy saving per capita) and X1 (renewable energy sources in percentage of the total production of electricity).
- Weak positive correlation between X3 (outlays on fixed assets serving environmental protection: energy saving per capita) and the emissions of air pollutants (X4, X5).

The angles between vectors representing the set of variables and the principal components (axes) can be used to assess the linear correlation coefficients. The correlation circle shows that the first axis corresponds to a pollution gradient (X4, X5) and the electricity consumptions (X2), with high levels of pollution toward the left and absence of pollution on the right. The second axis is negatively correlated with X3 (outlays on fixed assets serving environmental protection: energy saving per capita) and X6 (municipal waste collected separately during the year in relation to the total municipal waste). Only the first two axes are presented graphically in Figure 1, but it is also worth noting that the third axis is strongly negatively correlated with X7 (mixed municipal waste from household collected during the year per capita).

The analysis of the voivodships factor map (the right panel in Figure 1) is quite difficult since many labels are superimposed, but, as it is easy to notice, the most to the left there are points representing Śląskie Voivodship (i.e. the voivodship with the highest emission of air pollutants).

It is easier to interpret the results presented in the right panel of Figure 1 with the use of star plots with ellipses (Figure 2 and Figure 3).

Figure 2. PCA factor map with the eight years grouped for each voivodship



Legend: names of voivodships as in Figure 1

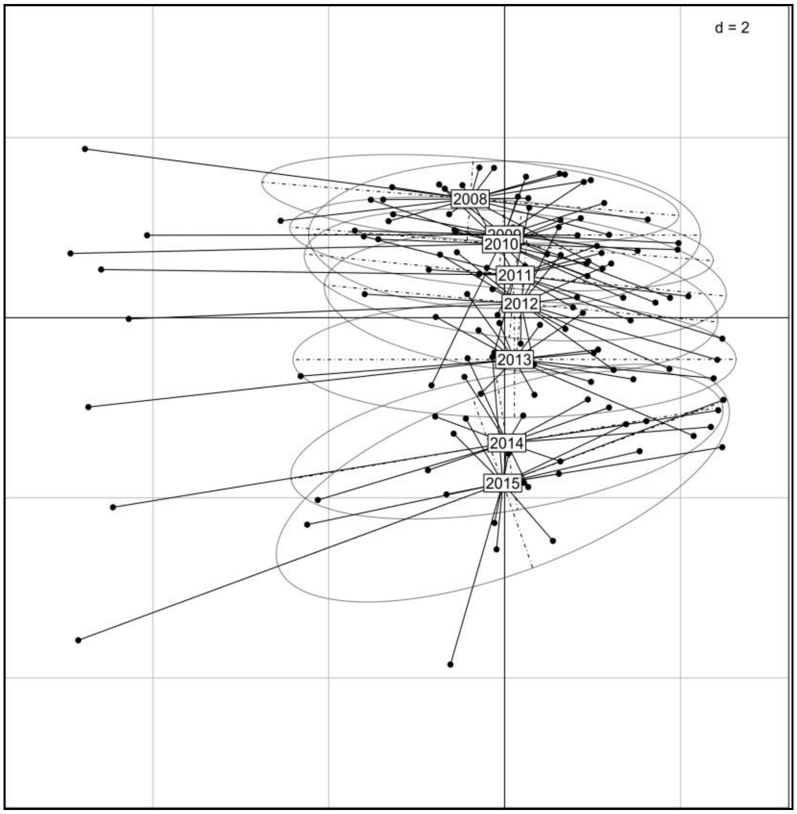
Source: Author's own elaboration.

In Figure 2, the eight years are grouped using the eight-pointed stars for each voivodship. The central point of each star is the centre of gravity (centroid) and the points representing the row coordinates for the 8 years are linked to the centroid.

There is a clear structure of voivodships along the first principal component corresponding to the emission of pollutants (X4, X5) and the electricity consumptions (X2). Śląskie Voivodship (S) is the absolute leader (much more polluted than the others), followed by Łódzkie voivodship (E). The most rightward are 3 voivodships (N – Warmińsko-Mazurskie Voivodship, B – Podlaskie Voivodship, R – Podkarpackie Voivodship) with the lowest levels of pollutant emissions and the highest share of renewable energy sources.

The vertical arrangement of the long axes of all ellipses along the second principal component indicates a large variation in the level of the outlays on fixed assets serving environmental protection: energy saving per capita (PLN) and separate collection of municipal waste during the years 2008-2015. This can easily be seen in Figure 3, where the voivodships are grouped using the sixteen-pointed stars for each year.

Figure 3. PCA factor map with the sixteen voivodships grouped for each year



Source: Author's own elaboration.

The centroids corresponding to the individual years are arranged along the vertical axis (the second principal component describing, in general, the level of environmental protection) and in succession from top to bottom – it can therefore be said that, from one year to another, the outlays on fixed assets serving environmental protection and the municipal waste collected separately during the year in relation to the total municipal waste have been constantly increasing (also in Śląskie Voivodship).

The classical PCA mixes both the temporal and the spatial typologies. It is possible to separate the two processes using a between-groups[1] PCA. The main goal of such an analysis is to reveal the differences between groups. The results of the between-groups PCA for groups corresponding to voivodships are presented in Figure 4.

[1] The terms: between-class PCA and between-groups PCA are treated as synonyms.

The total variability (inertia) in classical PCA equals the number of variables (7 in this research). In the presented analysis, the between-class inertia is equal to 5.0488, i.e. 72.13% of the total inertia is due to the spatial factor.

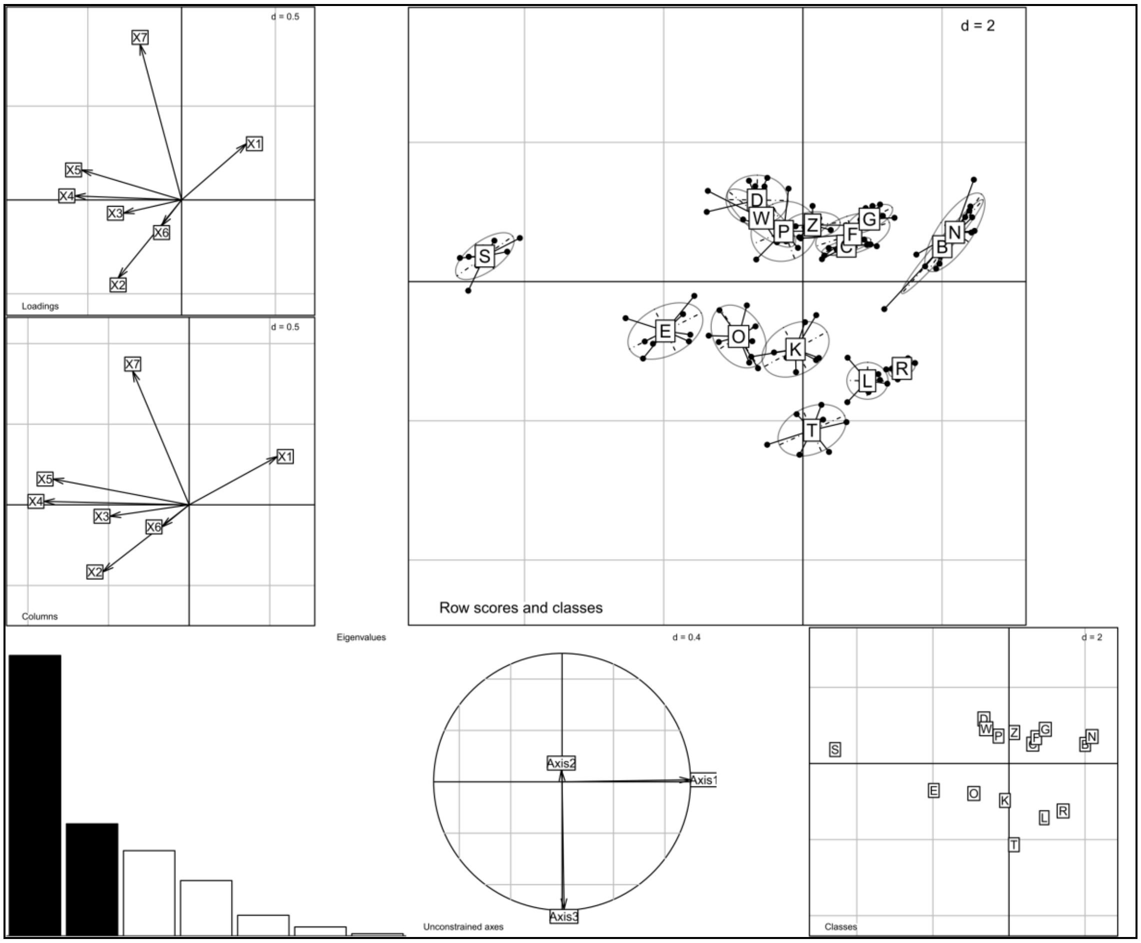
There are six elementary graphs in Figure 4. The main one (top-right: "Row scores and classes") presents the row scores for the initial data table – the eight years for each voivodship are grouped with the eight-pointed star and an ellipse. Each star and ellipse is labelled with the letter identifying the voivodship, located at the centroid of the star. In addition, the bottom-right graph ("Classes") shows the row scores of the between-class analysis for the 16 voivodships.

The graph labelled "Unconstrained axes" presents the projection of first three axes of the initial (classical) PCA into the between-class analysis and it shows the relationships between the initial PCA and the between-class PCA. The first axis of the classical PCA are equivalent to the first axis of the between-class PCA. The second axis in the between-class analysis is a combination of the second and the third axes of the initial PCA.

Eigenvalues are presented with the use of classical bar chart (in the lower-left graph), the first two eigenvalues are greater than 1.

The last two graphs on the left are labelled "Loadings" and "Columns". They both present the 7 indicators used in the analysis and they should be comparable since large differences between these two graphs indicate that the analysis results are not coherent. The "Loadings" graph gives the coefficients of the linear combination that maximize the between-class to total variance ratio and the "Columns" graph shows the scores of the variables (see Thioulouse et al. 2015: 149).

Figure 4. The between-voivodships PCA results



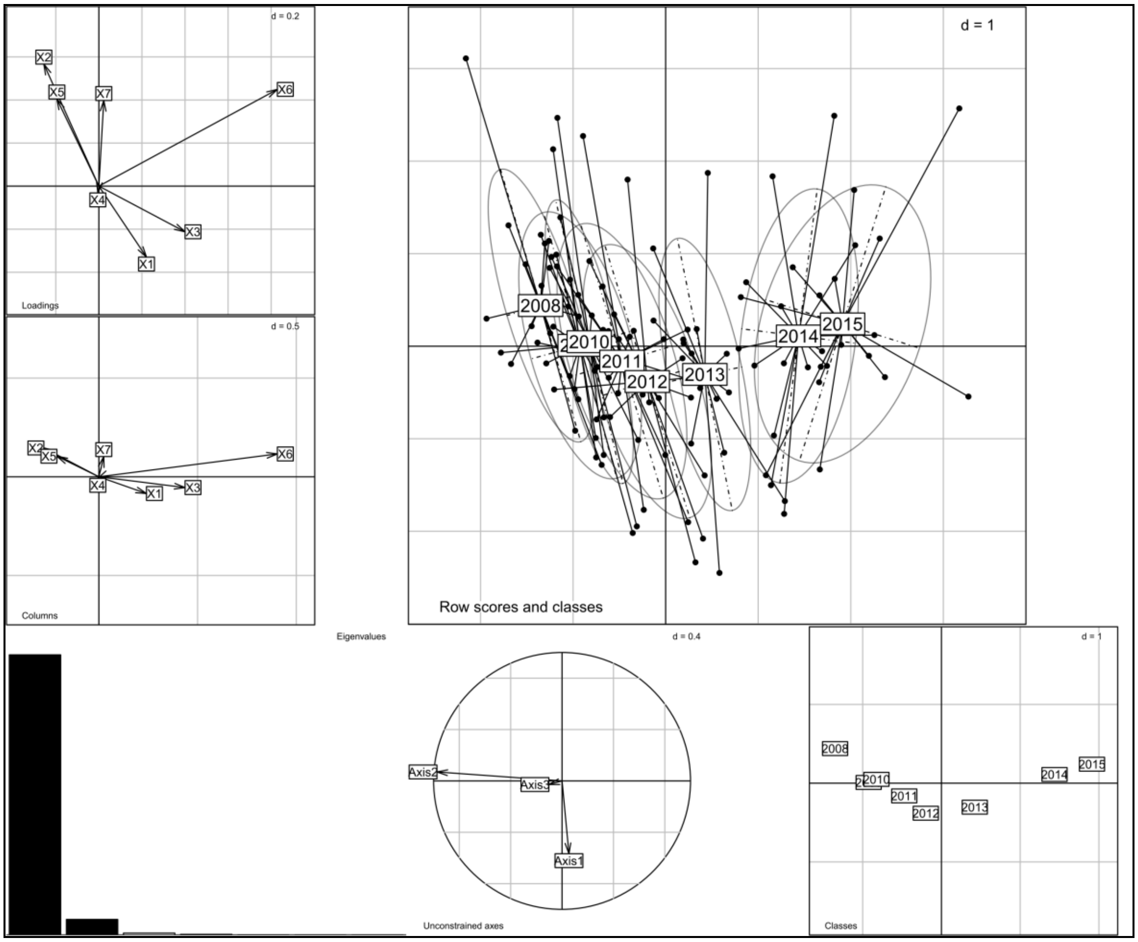
Legend: as in Figure 1

Source: Author’s own elaboration.

It can be seen that the first axis of the between-class PCA is quite similar to the firs axis of the classical PCA and corresponds mainly to the pollutant emissions. The second axis of the between-class PCA corresponds to the amount of the mixed municipal waste from household collected during the year per capita (in kg). Four groups of voivodships can be seen (the “row scores and classes” graph): (1) Śląskie Voivodship (S) which is more polluted than others; (2) Warmińsko-Mazurskie Voivodship (N) and Podlaskie Voivodship (B) with the lowest level of pollutant emissions and the highest share of renewable energy sources; (3) Łódzkie Voivodship (E), Opolskie Voivodship (O), Małopolskie Voivodship (K), Świętokrzyskie Voivodship (T), Lubelskie Voivodship (L) and Podkarpackie Voivodship (R) – 6 voivodships with a slightly higher level of outlays on fixed assets serving environmental protection: energy saving per capita and municipal waste collected separately during the year in relation to the total municipal waste and with relatively low level of mixed municipal waste from household collected during the year per capita; (4) seven other voivodships with a higher volume of the mixed municipal waste and a higher share of renewable energy sources.

The results of the between-groups PCA for groups corresponding to individual years are presented in Figure 5.

Figure 5. The between-years PCA results



Legend: as in Figure 1

Source: Author's own elaboration.

The between-class inertia is equal to 1.2680, i.e. only 18.11% of the total inertia is due to the temporal factor.

In the between-years analysis the values of each indicator are averaged across the voivodships. This removes the spatial component in the dataset and may make the temporal structure more apparent. Unfortunately, in the presented study, the results from the between-years analysis do not differ much from the classical PCA results. Only the first axis with an eigenvalue greater than 1 should be interpreted.

The first axis of the between-years PCA is equivalent to the second axis of the classical PCA, i.e. it is correlated with X6 (municipal waste collected separately during the year in relation to the total municipal waste) and X3 (outlays on fixed assets serving environmental protection: energy saving per capita) and slightly less correlated with X1 (renewable energy sources in percentage of the total production of electricity). It can be seen that, from year to year, the situation in (generally speaking) environmental protection has been improving (the amount of the municipal waste collected separately increases, there is also an increase in the outlays on fixed assets serving environmental protection and in the share of renewable energy).

4. Conclusions

On the basis of the obtained results and their graphical representation, the variability of the distributions of indicators describing the level of sustainable development in environmental domain was analysed taking into account the changes over time (during the years 2008-2015) and in space (by voivodships). The relationships between the analysed environmental domain indicators were also determined.

The voivodships are characterized by a great diversity mainly due to the level of air pollutants emissions from plants especially noxious to air purity. The leader in that case is Śląskie Voivodship, which is also distinguished by a high level of electricity consumption per 1 million PLN GDP. At the opposite end there are Warmińsko-Mazurskie Voivodship and Podlaskie Voivodship with a low level of industrialization, which are also characterized by the largest share of renewable energy sources in total electricity production.

In the years 2008-2015, a visible increase in outlays on fixed assets serving environmental protection and development of ecological awareness of society can be observed. There is also observed a noticeable increase in the share of municipal waste collected separately in relation to the total municipal waste starting from 2012. It may be related to the amendment to the Act on maintaining cleanliness and order in communes in 2011.

The methods of data analysis used in the paper belong to the exploratory data analysis techniques, which are helpful in identifying systematic relationships between many variables when there are no *a priori* expectations as to the nature of these relationships. In such cases, the principal component analysis (PCA) and the between-class PCA may be recommended as a useful tool for preliminary and informal analysis of the data. The graphical representation of the PCA and between-class PCA results is the basis for the interpretation.

Beneficiaries of such analyses may include, among others, marshal offices, voivodship offices, business representatives, investors and innovators. The great diversity of voivodships from the point of view of the analysed environmental domain indicators implies the need to undertake actions in order to reduce the inequalities between the voivodships. It should be aimed at maintaining better cohesion in the voivodships' development and preventing the exclusion of the underdeveloped regions. A detailed analysis of the selected environmental domain indicators, taking into account the changes over time and in space, may be helpful in monitoring and preparing an appropriate regional policy, based on specific conditions and resources of particular voivodships.

Literature

Bal-Domańska, B., Wilk, J. (2011). Gospodarcze aspekty zrównoważonego rozwoju województw – wielowymiarowa analiza porównawcza. *Przegląd Statystyczny* R. LVIII – Zeszyt 3-4: 300-322.

Bal-Domańska, B., Wilk, J., Bartniczak, B. (2012). Pomiar postępów województw w kierunku zrównoważonego rozwoju w zakresie zdrowia publicznego. *Econometrics* 3(37): 83-92.

Balicki, A. (2009). *Statystyczna analiza wielowymiarowa i jej zastosowania społeczno-ekonomiczne*. Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk.

Drabarczyk, K. (2017). Zrównoważony rozwój województw – analiza porównawcza. *Zeszyty Naukowe Politechniki Częstochowskiej: Zarządzanie* 25(2): 23-34.

Everitt, B.S., Skrondal, A. (2010). *The Cambridge Dictionary of Statistics* (Fourth Edition). Cambridge University Press, Cambridge.

Fura, B. (2015). Zróżnicowanie poziomu rozwoju zrównoważonego województw Polski z wykorzystaniem analizy wielowymiarowej. *Nierówności Społeczne a Wzrost Gospodarczy* 44 (4): 108-117.

Gabriel, K.R. (1971). The biplot graphical display of matrices with application to principal component analysis. *Biometrika* 58 (3): 453-467.

Gower, J.C., Le Roux, N.C., Gardner-Lubbe, S. (2015). Biplots: quantitative data. *WIREs Comput Stat* no. 7: 42-62 (doi: 10.1002/wics.1338).

GUS (2011). *Sustainable Development Indicators for Poland*. Statistical Office in Katowice. Katowice.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24: 417-441, 498-520.

Łuczak, A., Kurzawa, I. (2017). Ocena poziomu zrównoważonego rozwoju powiatów w Polsce z wykorzystaniem metod taksonomicznych. *Taksonomia 29. Klasyfikacja i analiza danych – teoria i zastosowania. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu* nr 469: 109-118.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, *Philosophical Magazin*e 6(2): 559-572.

Roszkowska, E., Filipowicz-Chomko, M. (2016). Ocena poziomu rozwoju instytucjonalnego województw Polski w latach 2010-2014 w kontekście realizacji koncepcji zrównoważonego rozwoju. *Ekonomia i Środowisko* 3(58): 250-266.

Roszkowska, E., Karwowska, R. (2014). Wielowymiarowa analiza poziomu zrównoważonego rozwoju województw Polski w 2010 roku. *Economics and Management* 1: 9-37.

Thioulouse, J., Dray, S., Dufour, A.-B., Siberchicot, A., Jombart, T., Pavoine, S. (2015). *Multivariate Analysis of Ecological Data with ade4*, Springer.

O zastosowaniu wybranych metod eksploracyjnej analizy danych do oceny różnic w poziomie zrównoważonego rozwoju w zakresie ładu środowiskowego województw w Polsce

Streszczenie

Celem pracy była ocena różnic w poziomie zrównoważonego rozwoju województw w Polsce w zakresie ładu środowiskowego w latach 2008-2015. Do analiz wykorzystano 7 wskaźników ładu środowiskowego należących do 3 obszarów tematycznych (energia, ochrona powietrza, gospodarka odpadami). Zastosowano analizę głównych składowych (PCA) oraz międzygrupową analizę głównych składowych (*between-class* PCA). Na podstawie uzyskanych wyników stwierdzono występowanie różnic między województwami przede wszystkim ze względu na emisję zanieczyszczeń powietrza (gazowych i pyłowych) z zakładów szczególnie uciążliwych. Wykazano także, że w latach 2008-2015 nastąpił wyraźny wzrost nakładów na środki trwałe służące ochronie środowiska związane z oszczędzaniem energii elektrycznej a także rozwój świadomości ekologicznej społeczeństwa.

Słowa kluczowe: zrównoważony rozwój, ład środowiskowy, wielowymiarowa eksploracyjna analiza danych, analiza składowych głównych