Iwona Bąk
West Pomeranian University of Technology

APPLICATION OF MULTIDIMENSIONAL CORRESPONDENCE ANALYSIS IN THE RESEARCH ON THE QUALITY OF NATURAL ENVIRONMENT IN POLAND IN 2007

Summary: The article attempts to answer a question of what kind of quality the natural environment was in Polish provinces in 2007 with reference to an average level of chosen variables that characterize the environment. The chosen variables describe the degree of air, water and soil pollution and inform about the environment protection expenditure incurred in the provinces and about remarkable natural virtues of the analyzed area.

The main aim of the study was to indicate which provinces diverge *in plus* or *in minus* from the average level of chosen variables that characterize the quality of natural environment in Poland. Also, what kind of connections there are between the provinces with reference to the variables analyzed. A multidimensional correspondence analysis was used as a research tool based on a built complex matrix of markers. With respect to a large number of versions of analyzed variables the Ward method was used, which allows for identifying connections between the variants of variables.

Key words: multiple correspondence analysis, natural environment, Ward method, indicator matrix.

1. Introduction

Dynamic economic development together with fast growth of population results in the excessive use and overload of the human natural environment. Today's state of Polish natural environment and its destruction in many districts of the country indicate the need to start effective, intensive and preventive activities in order to preserve natural resources. With this in view, a diagnosis of demands' scale, the identification of environment's degradation is necessary. There is a need for reliable knowledge about the state of natural environment and any changes that take place inside it. This knowledge is essential to make optimal decisions regarding, for example, a use of grounds, localisation of all kinds of plants, or in a wider aspect, industrial conversion in a region (county, district).

In the article the main aim of the research is to identify provinces that diverge *in plus* or *in minus* from the average level of chosen variables that characterize the quality of natural environment in Poland. The article also presents what kind of connections there are between the provinces with reference to the variables analyzed. The following variables were used in measuring the quality of natural environment in 2007 [Rocznik Statystyczny... 2008][1]:

X_1 – sewage treated in % of sewage required treatment,

X_2 – population using sewage treatment plant expressed in % of the whole population,

X_3 – dust-borne air pollution emission of specific burdensome factories expressed in thousands of tons per 1 km^2,

X_4 – gas-borne air pollution emission of specific burdensome factories expressed in thousands of tons per 1 km^2,

X_5 – total waste (excluding communal waste) produced during a year expressed in tons per 1 km^2,

X_6 – total communal waste produced expressed in kg per 1 dweller,

X_7 – legally protected area of remarkable natural virtues expressed in % of the total area,

X_8 – environment monuments per 1 km^2,

X_9 – general use and housing estate's green areas located in towns and countries, expressed in m^2 per 1 dweller,

X_{10} – forestation in %,

X_{11} – expenditure on assets used for environment protection expressed in mln of PLN per 1000 dwellers,

X_{12} – expenditure on assets used for water balance expressed in mln PLZ per 1000 dwellers.

Stimulants dominate in the variables set and only four variables are recognized as destimulants (X_3, X_4, X_5, X_6).

2. Research description

Correspondence analysis is a method included in a group of statistical multidimensional analysis methods. This method is used when investigated variables are measured in the nominal scale and are characterized by co-existence, i.e. it is not possible to clearly distinguish dependent variable in the whole set of variables analyzed [Gatnar, Walesiak (ed.) 2004]. A starting point in the multidimensional correspondence analysis is a proper preparation of entry data set. Numbers assigned to variants (categories) of variables may be formulated in the following way: complex markers matrix, Burt matrix, multidimensional quota table and total quota table.

[1] The choice of the diagnostic variables for the classification of objects from natural environment quality point of view was made inter alia in the study [Bąk, Sompolska-Rzechuła 2005]. Therefore these variables were taken into account in the article as they represent good object discrimination characteristics.

The multidimensional correspondence analysis with complex markers matrix were used in the article. In the matrix the number of rows was equal to the number of units analyzed (provinces) while the number of columns corresponded with double number of variables analyzed. Such a quantity of columns results from the essence of markers matrix where elements have values of 1 and 0 [Stanimir 2005]. Therefore, each variable analyzed was changed into a zero-one variable according to the following rule:

For stimulants: $$xs_i = \begin{cases} 1 & where \quad x_i \geq M \\ 0 & where \quad x_i < M \end{cases}.$$

For destimulants: $$xd_i = \begin{cases} 1 & where \quad x_i \leq M \\ 0 & where \quad x_i > M \end{cases}.$$

Taking median for a boundary value resulted from the types of analyzed variables distribution which were characterised by a very large diversity and strong asymmetry [Wawrzyniak 2000].

In the analyzed set of variables apart from twelve zero-one variables a *Province* variable with 16 variants was considered. Therefore, the measurement of actual space of co-existence amounted to 27. The measurement was determined according to the formula:

$$K = \sum_{q=1}^{Q} (J_q - 1), \tag{1}$$

where: J_q – the number of categories of a quality q (q = 1, 2, …, Q), Q – the number of variables.

Next, it was checked to what degree the eigenvalues of the lower dimension area explained the total inertia (λ = 2,0769)[2]. Greenacre's criteria was used in this respect. According to Greenacre's criteria inertias larger than $\frac{1}{Q} = \frac{1}{13} = 0{,}0769$ are important for the analysis. Table 1 indicated that these were inertias for K values of maximum 12. Values of meter τ_k[3] were analyzed for this size. It turned out that the degree of inertias explanation in a second-dimension equalled 31,044%. In order to increase the quality of mapping in the second dimension[4] a modification of eigenvalues was conducted according to Greenacre's proposal in a following way:

$$\tilde{\lambda}_k = \left(\frac{Q}{q-1}\right)^2 \cdot \left(\sqrt{\lambda_{B,k}} - \frac{1}{Q}\right)^2, \tag{2}$$

[2] Total inertia is the K sum of own values, where K is a true dimension of co-existence area.

[3] This meter measures the inertia participation of a certain dimension (λ_k) in a total inertia (λ).

[4] In order to define the dimension of mapping space a eigenvalues graph was prepared. Using the criteria of the "elbow" [Stanimir 2005] it was stated that the space should be second dimension.

where: Q – number of variables analyzed, $\lambda_{B,k}$ – k-eigenvalue (k = 1, 2, …, K), ($\sqrt{\lambda_{B,k}} = \gamma_{B,k}$), $\gamma_{B,k}$ – k-singular value of matrix B.

Table 1 presents original and modified eigenvalues together with the degree of total inertia explanation. The table omits results for $K > 12$ because for those values inertias were not larger than 0,0769 and so not vital for the research.

Table 1. Singular and eigenvalues together with the degree of total inertia explanation in the original and modified version

K	Singular values γ_k	Eigenvalues λ_k	λ_k / λ	τ_k	$\tilde{\lambda}_k$	$\tilde{\lambda}_k / \tilde{\lambda}$	$\tilde{\tau}_k$
1	0.6343	0.4024	0,1937	0,1937	0.3646	0.2552	0.2552
2	0.4923	0.2424	0,1167	0,3104	0.2025	0.1417	0.3969
3	0,4470	0,1998	0,0962	0,4066	0,1607	0,1125	0,5093
4	0,4286	0,1837	0,0885	0,4951	0,1452	0,1016	0,6109
5	0,3805	0,1447	0,0697	0,5648	0,1081	0,0757	0,6866
6	0,3505	0,1229	0,0592	0,6239	0,0878	0,0615	0,7481
7	0,3356	0,1126	0,0542	0,6782	0,0785	0,0550	0,8030
8	0,3147	0,0990	0,0477	0,7259	0,0663	0,0464	0,8494
9	0,3067	0,0940	0,0453	0,7711	0,0619	0,0433	0,8928
10	0,2893	0,0837	0,0403	0,8114	0,0530	0,0371	0,9298
11	0,2852	0,0813	0,0392	0,8506	0,0509	0,0356	0,9655
12	0,2820	0,0795	0,0383	0,8889	0,0494	0,0345	1,0000
13	0,2774	0,0769	0,0370	0,9259	**1,4291**		

Source: own calculations.

The degree of total inertia explanation improved a lot as a result of conducted modification. First two eigenvalues stand for 39,69% of the total modified inertia. Therefore the graphic presentation of the results of the correspondence analysis in second-dimension was made taking into consideration the modification of eigenvalues (fig. 1). The following formula sets the new values of coordinates in the second-dimension for the variables' categories:

$$\tilde{F} = F^{*} \cdot \Gamma^{-1} \cdot \tilde{\Lambda}, \tag{3}$$

where: $\tilde{F}$ – new coordinates values for the variables' categories (40×2 dimension), F^{*} – original coordinates values matrix for variable's categories (40×2 dimension), Γ^{-1} – diagonal opposite matrix of singular values (2×2 dimension), $\tilde{\Lambda}$ – modified diagonal matrix of eigenvalues (2×2 dimension).

The following elements were considered while interpreting the scatter of points in the two dimensional:

1. The location of the point against the centre of projection (the start of coordinates set) – the location closer to the starting point indicates that its profile

has values similar to the average profile and the points located far from the starting point indicate the dependencies between the investigated variables.

2. Point location against other points characterizing the categories of the same variable – near the location of the points characterizing variants of the same variable indicate the similarities of the profiles and *ipso facto* not fundamental diversity of the set units with regards to the variants (they may be connected).

3. The point location against the point describing the categories of different variable – the closer the points are located to each other the stronger the connections between the variants.

The analysis of points scatter (fig. 1) indicates that there are few categories of variables that are located close to the starting point of the coordinates set, whereas the points presenting most of the provinces are located the farthest from the centre of projection. Such a set of points shows that the dependencies exist between the categories. It is worth noticing that the points characterizing the categories of the same variable are located at opposite sides of the axis, which testifies that their profiles are not similar. This is a result of complex markers matrix usage where zero-one variables are applied.

While analysing the location of the point against the point describing variants of other variables, it is easy to spot that there are strong connections among all: $X_{11\mathrm{p}}$, X_{7n}, X_{5n} and X_{7p}, X_{5p}, X_{11n}. As regards the aim of the research it is important to show the connections between the provinces and variants of tested variables. Unfortunately, connections of this type may be straightforward only for few provinces and may be identified based on the scatter of points. Such a regularity is visible for example for Kujawsko-Pomorskie voivodeship (X_{6p}) and Wielkopolskie voivodeship (X_{10n}, X_{2n}). Ward method[5] was used to classify all the categories of variables in order to identify a connection between the categories of analyzed variables and all thevoivodeships. The classification was made for all variables which were descried by values of two dimensions gained from correspondence analysis with consideration given to modified eigenvalues. The results of the classification by Ward method were plotted onto the results gained from second-dimension method of correspondence analysis (fig. 1). The points included in the classes resulted from Ward method were circled with solid line. Four classes were distinguished which included both voivodeships and categories of variables. This, in turn, allowed for the description of the natural environment state in Polish voivodeships in 2007. While characterising the state of natural environment in a certain group, strong variations *in plus* and *in minus* of the values analyzed from the average in Poland were considered. The lack of information about other

[5] Ward method is one of the agglomeration methods of grouping. It is used in the empiric research as regards both objects classification and characteristic classification. In this method the distance between the groups is defined as the module of difference between the sums of distance squares from the centre of the groups which contain the points [Malina 2004]. The module of concentration analysis (Agglomeration/Ward) programmed in Statistica 8.0 packet was used for calculations and graphical presentation.

variables in certain class meant that their level did not differ fundamentally from the country's average. The characteristic of individual classes was placed below:

Class I: Dolnośląskie (D), Śląskie (Śl) and Zachodniopomorskie (Z) voivodeships

The class includes those voivodeships which may be positively assessed as regards the percentage of population that uses sewage treatment plant. Also, it can be positively assessed as regards the natural attractiveness measured in the forestation and general use and housing estate green areas located in towns and countries. The expenditure on assets used for water economy were above the average in these voivodeships. The percentage of sewage treated and communal waste were assessed in minus.

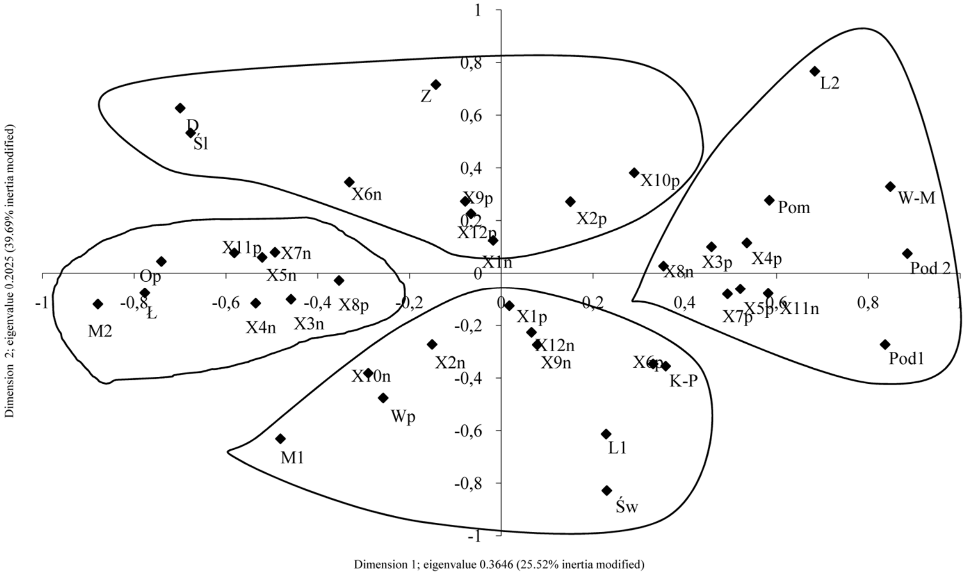

Figure 1. The presentation of the results of correspondence analysis of all categories of the variables as regards the modification of eigenvalues together with the results gained from Ward method (*p* symbol by individual categories of variables states positive level and *n* symbol – negative)

Source: author's own study.

Class II: Warmińsko-mazurskie (W-M), Lubuskie (L2), Pomorskie (Pom), Podlaskie (Pod2), Podkarpackie (Pod1) voivodeships

Voivodeships in this class may be assessed in a positive way as regards the gas-borne and dust-borne air pollution emission and the waste (excluding communal waste) produced during the year. The area of remarkable natural virtues was also above the average. However, the number of natural monuments was different *in minus* from the country average. The expenditure on the assets used for environment protection was also assessed in a negative way.

Class III: Mazowieckie (M2), Łódzkie (Ł), Opolskie (Op) voivodeship

These provinces strongly varied *in minus* from the average of the country because of the dust-borne and gas-borne pollution emission, waste produced (excluding communal) and the area of legally protected remarkable natural virtues. However, the number of natural monuments was formed above the average.

Class IV: Małopolskie (M1), Wielkopolskie (Wp), Świętokrzyskie (Św), Lubelskie (L1), Kujawsko-Pomorskie (K-P)

This class has a positive percentage of treated sewage and produced communal waste. The percentage of people using the sewage treatment plant was below the country's average. The natural attractiveness measured with the forestation and general use and housing estate green areas located in towns and countries and the expenditure on the assets used for water economy were assessed in a negative way.

3. Conclusions

The conducted research testifies that the voivodeships in Poland show a large diversification as regards the quality of natural environment. The provinces grouped in class I and IV and II and III are opposite to each other from the point of variables view and differ a lot from the average level of the country. The voivodeships which differed a lot *in plus* or *in minus* from the average level of tested variables were not distinguished. The most of *in plus* variations from the average value in Poland may be observed in the voivodeships included in class I and II. These voivodeships were assessed in a positive way as regards four variables. However, objects included in class III and IV were assessed in a positive way only in case of two variables while strong variations *in minus* from the median were observed in cases of four variables.

It was possible to distinguish the classes of provinces characterised by similar variables set that strongly differed from the average level of these variables in Poland thanks to the application of both multidimensional correspondence analysis and Ward method when finally interpreting the results.

References

Bąk I., Sompolska-Rzechuła A., *Wielowymiarowa analiza porównawcza jakości środowiska naturalnego w ujęciu wojewódzkim*, „Wiadomości Statystyczne" 2005, nr 9.

Gatnar E., Walesiak M. (red.), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych,* AE, Wrocław 2004.

Malina A., *Wielowymiarowa analiza przestrzennego zróżnicowania struktury gospodarki Polski według województw*, AE, Kraków 2004.

Rocznik Statystyczny Województw, Główny Urząd Statystyczny, Warszawa 2008.

Stanimir A., *Analiza korespondencji jako narzędzie do badania zjawisk ekonomicznych*, AE, Wrocław 2005.

Wawrzyniak K., *Klasyczne i pozycyjne parametry struktury jako normy w procesie oceny działalności przedsiębiorstw*, Zeszyty Naukowe Uniwersytetu Szczecińskiego nr 269, Prace Katedry Ekonometrii i Statystyki nr 8, Uniwersytet Szczeciński, Szczecin 2000.

ZASTOSOWANIE WIELOWYMIAROWEJ ANALIZY KORESPONDENCJI DO BADANIA JAKOŚCI ŚRODOWISKA NATURALNEGO W POLSCE W ROKU 2007

Streszczenie: W artykule podjęto próbę odpowiedzi na pytanie, jaki był stan jakości środowiska naturalnego w województwach Polski w odniesieniu do przeciętnego poziomu wybranych zmiennych charakteryzujących to środowisko w roku 2007. Do analizy wybrano zmienne, które opisują stopień zanieczyszczenia powietrza, wody i gleby, oraz zmienne informujące o nakładach poniesionych przez województwa na ochronę środowiska i o szczególnych walorach przyrodniczych badanego obszaru.

Celem badania było wykazanie, które województwa odbiegają *in plus* i *in minus* od przeciętnego poziomu wybranych zmiennych charakteryzujących jakość środowiska naturalnego w Polsce oraz jakie są powiązania pomiędzy województwami z punktu widzenia badanych zmiennych. Jako narzędzie badawcze wykorzystano wielowymiarową analizę korespondencji na podstawie zbudowanej złożonej macierzy znaczników. Ze względu na dużą liczbę wariantów analizowanych zmiennych zastosowano metodę Warda, która umożliwiła wyznaczenie powiązań pomiędzy wariantami zmiennych.