Multivariate statistical analysis of environmental data

Justyna Brzezińska[1], Aneta Rybicka[2], Marcin Pełka[3]

Abstract

One of the characteristics of environmental data is that many of them are mostly described by complex and large number of variables. To understand this phenomena it is necessary to analyze the relationship and association between them. In this paper we apply multivariate statistical methods for the analysis of environmental problems. The main aim of the paper is to present an application of the linear ordering with multidimensional scaling for results visualization in the environmental data (green growth) analysis. The main contribution of this paper is the empirical part of this paper will that presents the application of linear ordering several multivariate methods and graphical presentation using modern and advanced visualizing tools based on datasets and reports from the Organization for Economic Cooperation and Development (OECD). Presented analysis may be used in all types of environmental practice and real life solutions. All calculations will be conducted done in R software using.

Keywords: *environmental data, multivariate statistical analysis, R software*
JEL Classification: C01, C38, C39
AMS Classification: 62H30, 62H86
DOI: 10.14659/SEMF.2018.01.04

1 Introduction

Most environmental data involve a large degree of complexity and uncertainty. Environmental Data Analysis is created to provide modern quantitative tools and techniques designed specifically to meet the needs of environmental sciences and related fields. Statistics is an indispensable means of environmental research. It is used to analyze and to interpret the increasing flood of vast data from environmental areas, which are often of heterogeneous nature and show high variability. Many important results and statements concerning the environment are based on statistical investigations, such as changes of the ozone layer, climate changes, global warming, air quality etc. But also less spectacular results about the influence of various human activities on certain environmental parameters which are not

[1] University of Economics in Katowice, Faculty of Management, Department of Economic and Financial Analysis e-mail: justyna.brzezinska@ue.katowice.pl.
[2] Wroclaw University of Economics, Faculty of Economics, Management and Tourism, Department of Department of Econometrics and Computer Science e-mail: aneta.rybicka@ue.wroc.pl.
[3] Wroclaw University of Economics, Faculty of Economics, Management and Tourism, Department of Department of Econometrics and Computer Science e-mail: marcin.pelka@ue.wroc.pl

obvious at first glance and superimposed by considerable random variations are important findings of statistical analysis.

In official statistics environmental monitoring has become a serious tool for political consulting and economic practice. For environmental research also statistical methods for the design and analysis of experiments play an important role. A specific scientific discipline of environmental statistics does not exist. The whole statistical methodology may be used in investigating environmental questions and a wide range of statistical methods can be applied. Several statistical methods may be applied for the data analysis, such as: time series analysis (see for example: Fu and Weng 2016; Chaudhary et. al. 2015; Proulx et. al. 2015), spatial analysis (see for example: Dale, Fortin 2014; May et. al. 2017) parametric and nonparametric regression analysis (see for example: Rivest et. al. 2016; Cade and Noon 2003; Xu et. al. 2016), exploratory data analysis (see for example: Seifert et. al. 2014), multivariate statistical analysis (see for example: Šmilauer and Lepš 2014).

In environmental research quite often the measurement, collection, storage, processing and analysis of data are not carried out by the same institution or researcher. Only in exceptional cases there is one person who knows about all the details of collection and analysis of the data, who knows about the scientific environmental background and at the same time also about the mathematical procedure and algorithms for the statistical evaluation and the presentation of the results. There is a great complexity in environmental statistics. Heterogeneous data from different sources and collection principles are analyzed simultaneously. There are dependencies between the measured quantities that usually do not follow fixed laws or rules, but reveal random variability. Besides this variability in the nature of the environmental problem there is variability in time and space. And measurements in time are not repeatable. Another problem lies in the abundance of data in environmental statistics. Although modern computer technique can cope with this, problems of compatibility, standardization and data harmonization arise as obstacles. Storing of data is often connected with coding. But the description of the coding principles is sometimes insufficient and this may make it difficult to join data from different sources together. There are several applications of statistical analysis of environmental data (see for example: Fu and Weng, 2016; Chaudhary et. al., 2015; Proulx et. al., 2015; Dale and Fortin, 2014; May et. al., 2017; Rivest et. al., 2016; Cade and Noon, 2003; Xu et. al., 2016; Seifert et. al., 2014; Šmilauer and Lepš, 2014; Crawford et. al., 2018; Ver Hoef, 2018).

The main aim of the paper is to present an application of the linear ordering with multidimensional scaling for results visualization in the environmental data (green growth)

analysis. The main contribution of this paper is the empirical part of this paper that presents the application of linear ordering and graphical presentation using modern and advanced visualizing tools based on datasets and reports from the Organization for Economic Cooperation and Development (OECD). Presented analysis may be used in all types of environmental practice and real life solutions. All calculations will be done in R software.

2 Environmental data analysis

In environmental statistics most data are the result of a measurement process, where the measuring instruments have a certain degree of precision and a limited range of scale. Both have to be taken into account in the analysis of the data, as well as their liability. Unintentional failure of measurement instruments and disturbances of a transmission channel may lead to false or missing values. If the deviations are big enough, the corresponding data are detected as outliers. There are statistical procedures to perform this. Supplementary etiology may even lead to a correction of the values.

Talking about environmental analysis it is crucial to talk about one of the most important area which is green growth. Due to the OECD green growth is a subset of sustainable development. It is narrower in scope, entailing an operational policy agenda that can help achieve concrete, measurable progress at the interface of the economy and the environment. It fosters the necessary conditions for innovation, investment and competition that can give rise to new sources of economic growth that are consistent with resilient ecosystems. Green growth strategies need to pay specific attention to many of the social issues and equity concerns that can arise as a direct result of greening the economy both at the national and international level. This is essential for the successful implementation of green growth policies. Strategies should be implemented in parallel with initiatives focusing on the broader social pillar of sustainable development. Green growth means fostering economic growth and development, while ensuring that natural assets continue to provide the resources and environmental services on which our well-being relies. To do this, it must catalyse investment and innovation which will underpin sustained growth and give rise to new economic opportunities. We need green growth because risks to development are rising as growth continues to erode natural capital. If left unchecked, this would mean increased water scarcity, worsening resource bottlenecks, greater pollution, climate change, and unrecoverable biodiversity loss.

Green growth is a big challenge and huge problem that deserve analysis. Multivariate statistical methods may bring solutions that may solve economic and government problems.

3 Data analysis in R software

The concept of pattern of development and the measure of development was proposed by Professor Zdzisław Hellwig in 1967 (Hellwig 1967). The general procedure in linear ordering of the data set based on pattern object (or anti-pattern object) and metric data requires to choose a complex phenomenon, that cannot be measured directly, for ordering the set of objects. The objects are described by variables. Preferential variables (stimulants, destimulants and nominants) must be identified among variables (see for example Hellwig 1981, for formal definitions of stimulants, destimulants and nominants). Nominants have to be transformed into stimulants. A pattern object and anti-pattern object are added to the data set. Data has to be normalized if the variables are measured on interval or ratio scale. The distance measure between object is calculated. Multidimensional scaling is done. The iterative procedure in the **smacof** algorithm was applied in the study (Borg and Groenen, 2005). Finally, two-dimensional space is obtained. The graphical presentation and interpretation of the results in a two-dimensional (multidimensional scaling results) and one-dimensional space (linear ordering results). In the Fig. a straight line that connects pattern and anti-pattern for the MDS results is added. This line is the so-called axis of the set. Also isoquants of development are added to the MDS plot. These isoquants are determined on the basis of the pattern object, e.g. by dividing the set of axis into four parts. The objects between isoquants present similar development level. The same level of development can be reached by objects placed in different locations on the same isoquant. Such representation expands the interpretation of the results. Finally, normalized distances d_i^+ of *i*-th object from the pattern of development are calculated as follows (Hellwig, 1981):

$$d_i^+ = \frac{\sqrt{\sum_{j=1}^{2}\left(v_{ij} - v_{+j}\right)^2}}{\sqrt{\sum_{j=1}^{2}\left(v_{i+j} - v_{-j}\right)^2}}, \quad d_i^+ \in [0;1], \tag{1}$$

where: $\sqrt{\sum_{j=1}^{2}\left(v_{ij} - v_{+j}\right)^2}$ – Euclidean distance between *i*-th object and the patter object, $\sqrt{\sum_{j=1}^{2}\left(v_{i+j} - v_{-j}\right)^2}$ – Euclidean distance between pattern and anti-pattern object.

The objects are ordered by the growing values of the distance measure (1). Linear ordering results are graphically presented on thure. The empirical study uses the statistical data representing green growth level of 33 OECD member countries in 2010. The data was

selected by using the convenience sampling. The evaluation of the green growth was done by using eight metric variables (measured on a ratio scale):

x_1 – renewable electricity generation (% of total energy produced),

x_2 – production-based CO_2 productivity (GDP per unit of energy-related CO_2 emissions),

x_3 – environmentally related government R&D budget (% of total government R&D),

x_4 – development of environment-related technologies (% of all technologies),

x_5 – population with access to improved sanitation (% total population),

x_6 – mortality from exposure to PM2.5,

x_7 – forests under sustainable management certification (% total forest area),

x_8 – municipal waste generated in kg per capita.

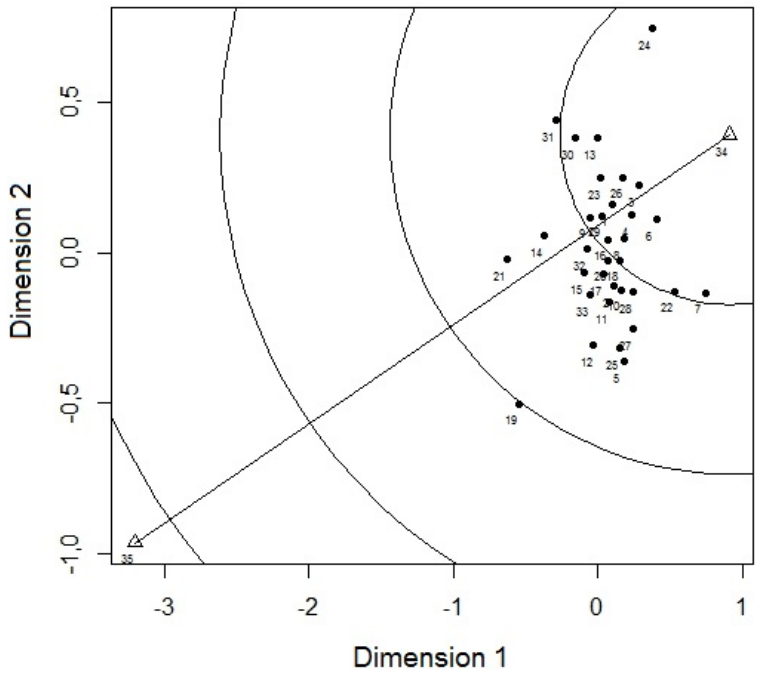

Fig. 1. Graphical presentation of multidimensional scaling results in two-dimensional space of 35 objects containing 33 countries, pattern object (34) and anti-pattern object (35) referring to the OECD countries green growth. Source: authors' compilation using R software.

Variables x_1, x_3, x_4, x_5, x_7 are stimulants, x_2, x_6 and x_8 are destimulants. The data was collected in 2010 from OECD data bank. A pattern and anti-pattern were added to the data set, so data matrix covers 35 objects described by eight variables.

Due to the fact all the variables are metric data was normalized. mdsOpt (Walesiak and Dudek, 2017) package of R software was used to find optimal multidimensional scaling procedure. The best result was obtained using positional standardization, Manhattan distance and ratio multidimensional scaling model. The Stress-1 was equal to 0.137825 and HHI spp (Hirschman-Herfindahl HHI index calculated based on stress per point) 345.2902.

The results of linear multidimensional scaling with the axis of the set, four isoquants are presented on the Fig. 1.

The ordering 33 countries from the pattern object by growing measure value (1) are shown in Table 1.

Table 1. The ordering of 35 objects regarding green growth.

Object no.	Name	Distance
34	Pattern	0.000000
7	Estonia	0.126559
6	Denmark	0.133721
24	Norway	0.147263
22	Netherlands	0.148296
3	Canada	0.150826
4	Chile	0.167198
26	Portugal	0.172901
8	Finland	0.186051
1	Austria	0.194838
28	Slovenia	0.195421
18	Korea	0.199848
23	New Zealand	0.208219
16	Italy	0.209917
10	Germany	0.210028
13	Iceland	0.210512
29	Spain	0.212897
27	Slovak Republic	0.215031
20	Luxembourg	0.216233
2	Belgium	0.219286
17	Japan	0.229008

11	Greece	0.230475
9	France	0.232427
25	Poland	0.240424
5	Czech Republic	0.241395
32	Turkey	0.244379
30	Sweden	0.246673
33	United Kingdom	0.254697
15	Israel	0.256432
12	Hungary	0.270643
31	Switzerland	0.27851
14	Ireland	0.306645
21	Mexico	0.367897
19	Latvia	0.395211
35	Anti-pattern	1.000000

Distances from the pattern object sorted by growing values of measure (1) are presented on the Fig. 2.

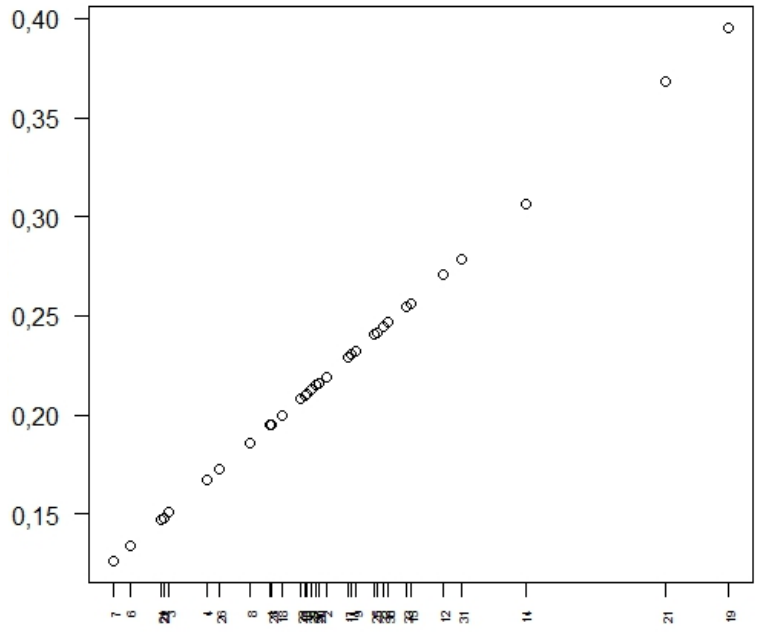

Fig. 2. Graphical presentation of linear ordering of 33 OECD countries referring to their green growth value. Source: authors' compilation using R software.

The top five countries (the best one) when considering values of the measure (1) are Estonia, Denmark, Norway and Netherlands. The five worst countries are Hungary, Switzerland, Ireland, Mexico and Latvia. Poland is the 24-th country, Czech Republic is the 25-th, Germany 15-th. When taking into consideration the location of the countries on the Fig. 1 all countries are within the first three isoquants. Countries with similar level of green growth but different positions on the Fig. 1 are: Estonia, Netherlands, Finland, Chile, Canada, Denmark, Italy, Korea (below the set of the axis), Austria, Portugal, Spain, New Zealand, Iceland, France, Sweden, Norway (above the set of the axis) that are within the first isoquant, Czech Republic, Poland, Slovak Republic, Slovenia, Hungary, Greece, Germany, Belgium, Luxembourg, Japan, United Kingdom, Israel, Turkey (below the axis of the set) and Ireland, Switzerland, Mexico (above the axis of the set) within the second isoquant. In general, we can say that more countries are below the set of the axis than above it. Also in general second isoquant contains mostly post-communist, central European countries.

Conclusions

The paper presents an application of the proposal introduced by Walesiak (2016) that allows the visualization of linear ordering results for the set of objects by applying multidimensional scaling for this task. The concept of isoquants and the path of development proposed by Hellwig (1981) allows to represent objects in two dimensions. The application of the proposed methods allows to show results for more than two variables. The proposed approach was illustrated by an empirical example where green growth data for 33 OECD member countries was used. In general, we can say that almost all OCED countries lie between two first isoquants, and so called "post-communist" countries can be found below the axis of the set. The best country, when considering green growth, is Estonia, then Denmark, Norway and Netherlands. The worst country is Latvia, then Mexico, Ireland and what can be surprising – Switzerland. The aim for the future studies should be the longitudinal analysis of the green growth and its changes during last years.

References

Borg, I. & Groenen, P.J.F. (2005). *Modern Multidimensional Scaling. Theory and Applictions. 2nd Edition*. Springer Science+Business Media, New York.

Cade, B. S. & Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, *1*(8), 412-420.

Chaudhary, S., McGregor, A., Houston, D. & Chettri, N. (2015). The evolution of ecosystem services: a time series and discourse-centered analysis. *Environmental Science & Policy, 54*, 25-34.

Crawford, F. W., Ho, L. S. T., & Suchard, M. A. (2018). Computational methods for birth-death processes. *Wiley Interdisciplinary Reviews: Computational Statistics*.

Dale, M. R. & Fortin, M. J. (2014). *Spatial analysis: a guide for ecologists*. Cambridge Unversity Press.

Fu, P. & Weng, Q. (2016). A time series analysis of urbanization induced land use and land cover change and its impact on land surface temperature with Landsat imagery. *Remote Sensing of Environment, 175*, 205-214.

Hellwig, Z. (1967). *Procedure of Evaluating High-Level Manpower Data and Typology of Countries by Means of the Taxonomic Method*, COM/WS/91, Warsaw, 9 December, 1967 (unpublished UNESCO working paper).

Hellwig, Z. (1981). Wielowymiarowa analiza porównawcza i jej zastosowanie w badaniach wielocechowych obiektów gospodarczych. In: Welfe, W. (ed.), *Metody i modele ekonometryczno-matematyczne w doskonaleniu zarządzania gospodarką socjalistyczną*, PWE, Warszawa, 46-68.

May, F., Gerstner, K., McGlinn, D. J., Xiao, X. & Chase, J. M. (2017). mobsim: An R package for the simulation and measurement of biodiversity across spatial scales. *Methods in Ecology and Evolution* (doi: 10.1111/2041-210X.12986).

Proulx, R., Parrott, L., Fahrig, L. & Currie, D. J. (2015). Long time-scale recurrences in ecology: detecting relationships between climate dynamics and biodiversity along a latitudinal gradient. In: *Recurrence Quantification Analysis* (pp. 335-347). Springer, Cham.

Rivest, L. P., Duchesne, T., Nicosia, A. & Fortin, D. (2016). A general angular regression model for the analysis of data on animal movement in ecology. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 65*(3), 445-463.

Seifert, B., Ritz, M. & Csősz, S. (2014). Application of exploratory data analyses opens a new perspective in morphology-based alpha-taxonomy of eusocial organisms. *Myrmecological News 19*, 1-15.

Šmilauer, P. & Lepš, J. (2014). *Multivariate analysis of ecological data using CANOCO 5*. Cambridge University Press.

Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M. & Fortin, M. J. (2018). Spatial autoregressive models for statistical inference from ecological data. *Ecological Monographs, 88*(1), 36-59.

Walesiak, M. (2016), Visualization of Linear Ordering Results for Metric Data with the Application of Multidimensional Scaling. *Ekonometria, 2*(52), 9-20.

Walesiak, M. & Dudek, A. (2017). *The mdsOpt package for R software*, http://r-project.org.

Xu, B., Luo, L. & Lin, B. (2016). A dynamic analysis of air pollution emissions in China: evidence from nonparametric additive regression models. *Ecological indicators, 63*, 346-358.