DOI: 10.20472/SS.2016.5.1.004

SATISFACTION WITH CHOSEN ASPECTS OF LIFE IN POLAND - EVALUATION BY CANONICAL CORRELATION METHODS

ALICJA GRZESKOWIAK

Abstract:

This paper is focused on the analysis of relations between variables representing the satisfaction with various aspects of life and other socio-economic indicators. The main objective of the study is to identify associations between multivariate datasets. The research is based on the datasets containing both metric and non-metric data, which determines the use of adequate analytical techniques. Hence, canonical correlation analysis and nonlinear canonical correlation analysis are applied, respectively. Special attention is paid to the visual presentation of the outcomes of the analyses. The approach used in this paper allows for the detection of interesting relations and patterns between datasets under consideration.

Keywords:

satisfaction with life, canonical correlation analysis, nonlinear canonical correlation analysis

JEL Classification: I31, I39

Authors:

ALICJA GRZESKOWIAK, Wrocław University of Economics, Poland, Email: alicja.grzeskowiak@ue.wroc.pl

Citation:

ALICJA GRZESKOWIAK (2016). Satisfaction with chosen aspects of life in Poland - evaluation by canonical correlation methods. International Journal of Social Sciences, Vol. V(1), pp. 60-71., 10.20472/SS.2016.5.1.004

1 Introduction

This paper presents the results of the research on satisfaction with chosen aspects of life in Poland in comparison to some other variables describing overall situation in the sense of the sustainable development idea and some variables representing the socio-economic background. The most recent vast survey on the social situation in Poland called *Social Diagnosis* shows that most areas of life are more and more highly rated (Czapiński & Panek, 2015, p. 188). This study attempts to analyze the problem in another context and on the base of various data sources as EU-SILC outcomes, Eurostat data on the sustainable development and the Quality of Life Survey data.

The main objective of the study is to identify relations in the considered datasets, thus the descriptive approach is prevailing. The specific objectives are formulated as follows:

- to search for relations between the level of satisfaction with selected aspects of life and the level of sustainable development,
- to assess the position of Poland in comparison with other EU countries,
- to search for relations between the level of satisfaction with selected life aspects and the socio-economic background of the respondents in Poland.

As the goals of the paper are concentrated on the identification of the relations between two datasets, canonical correlation analysis is applied. Since the variables included in the analyses are either metric or non-metric, two methods from the canonical correlation family are used with respect to the character of the data.

2 Data sources

The analyses were based on both aggregated and individual data coming from various sources. The international comparison across European Union countries was performed according to aggregated data made available by Eurostat and organized for this study in two sets. The first one comprises information on personal satisfaction with life indicators acquired from the EU-SILC (European Union Statistics on Income and Living Conditions). The second one includes nine sustainable development indicators.

EU SILC *ad-hoc* module which was implemented in 2013 provides information on the *average rating of satisfaction by domain.* The satisfaction was evaluated on the scale from 1 to 10 and covers the following aspects [Eurostat code: ilc_pw01]:

- *satisfaction with financial situation* - denoted FIN in further considerations,
- *satisfaction with accommodation* - denoted ACC,
- *job satisfaction* - denoted JOB,
- *satisfaction with commuting time* - denoted COM,
- *satisfaction with time use* - denoted TIM,
- *satisfaction with recreational and green areas* - denoted REC,
- *satisfaction with living environment* - denoted ENV,
- *satisfaction with personal relationships* - denoted PER,

- *meaning of life* - denoted LIF.

Sustainable development indicators are gruped in ten themes (European Union, 2015, p. 9): *socioeconomic development, sustainable consumption and production, social inclusion, demographic changes, public health, climate change and energy, sustainable transport, natural resources, global partnership, good governance.* Headline indicators for themes were considered, i.e (cf. European Union, 2015):

- *real GDP chain linked volumes (2010), euro per capita* [Eurostat code: tsdec100] - denoted GDP in further considerations,
- *resource productivity (EUR/kg)* [Eurostat code:tsdpc100] - denoted RES,
- *people at risk of poverty or social exclusion, % of total population* [Eurostat code:tsdsc100] - denoted POV,
- *employment rate of older workers* [Eurostat code: tsdde100] - denoted EMP,
- *life expectancy - females, Life expectancy - males* [Eurostat code: tsdph100] - denoted LEF and LEM, respectively,
- *primary energy consumption (2005=100)* [Eurostat code:tsdcc120] - denoted ENE,
- *greenhouse gas emissions, 2012 (1990=100)* [Eurostat code: tsdcc100] - denoted GAS,
- *energy consumption of transport relative to GDP (2000=100)* [Eurostat code: tsdtr100] - denoted TRA,
- *official development assistance as share of gross national income* [Eurostat code: tsdgp100] - denoted ASS .

No headline indicator is defined for *good governance* area and there were no complete data for the *natural resources* main indicator, so these two categories are not taken into account. As the EU-SILC data on well-being concern the year 2013, the sustainable development indicators correspond to the same year to make an appropriate comparison, with an exception of the greenhouse gas emissions for which the latest available data are for the year 2012.

The second analysis was carried out on the base of individual data from the *Quality of Life Survey* [European Foundation for the Improvement of Living and Working Conditions, 2015] which took place in 2011-2012. Only the answers given by the Polish respondents were considered. Six variables depicting personal satisfaction with certain aspects of life were included in the analysis, i.e.

- satisfaction with education,
- satisfaction with standard of living,
- satisfaction with accommodation,
- satisfaction with family life,
- satisfaction with health,
- satisfaction with social life.

Each domain was evaluated from 1 to 10. The assessment of the satisfaction was compared to some categorical variables representing the socio-economic background of the respondents, namely:

- gender (male, female),
- age (categorized into five intervals: 18-24, 25-34, 35-49, 50-64, 65+),
- education level (primary or less, secondary, tertiary),
- place of residence (four categories: countryside, village or small town, medium/large town, city),
- labour market status (recoded into three categories: employed, retired, not working - comprising unemployed, disable and ill persons, homemakers, students and others).

As some missing data occurred, a few observations were removed from the primary dataset, so the final set used for the analysis included N=2206 records.

3 Canonical correlation methods as analytical techniques

Canonical correlation methods are applied to investigate relationships between multivariate datasets. The principles of the canonical correlation analysis were formulated by Hotelling (1935, 1936). The canonical correlation analysis is used mainly for two purposes, i.e. data reduction and evaluating independence of two sets of variables (Raykov & Marcoulides, 2008, p.368). The rationale for this technique is that complex multivariate data structures can be better examined when a low-dimensional projections are made (Haerdle & Simar, 2003, p. 361). New axis are determined in such a way that the correlation between new (canonical) variables is as strong as possible. The gain is that each initial set is reduced to a few linear combinations and only correlations between them are considered instead of examining numerous relationships of original variables (Sharma, 1996, p. 396). The details of the canonical correlation analysis can be found in many publications, e.g. in Bernstein et al. (1998, p. 363-375), Gittins (1985, p. 11-66), Haerdle & Simar (2003, p. 361), Sharma (1996, p. 391-418), (Raykov & Marcoulides, 2008, p.367 - 390), Rencher (2003, p. 361-379), Thompson (1984). The general idea of the procedure is as follows (Raykov & Marcoulides, 2008, p. 368):

- a linear combination Z_1 of the variables from the first set is found and a linear combination W_1 of variables from the second set is determined such that their correlation is maximum across all possible weights,
- a linear combination Z_2 of the variables from the first set is found and a linear combination W_2 of variables from the second set is determined such that their correlation is maximum under the assumption of no correlation with Z_1 and W_1,
- the process is continued in an analogous way until all possible pairs are obtained.

The interpretation of the results includes, among others, the assessment of the importance of canonical variate pairs (i.e. Z_1 and W_1, Z_2 and W_2, and consequent ones), the canonical coefficients and loadings evaluation, the computation of the redundancy measures (Sharma, 1996, p. 401-406). Canonical correlation algorithms are available in many

statistical packages as SPSS, SAS, Statistica and R. The latter one offers a wide range of visualization techniques to help with the information obtained from the analysis They can be found in packages `vegan` (Oksanen et al. 2015) or `CCA` (González et al. 2008).

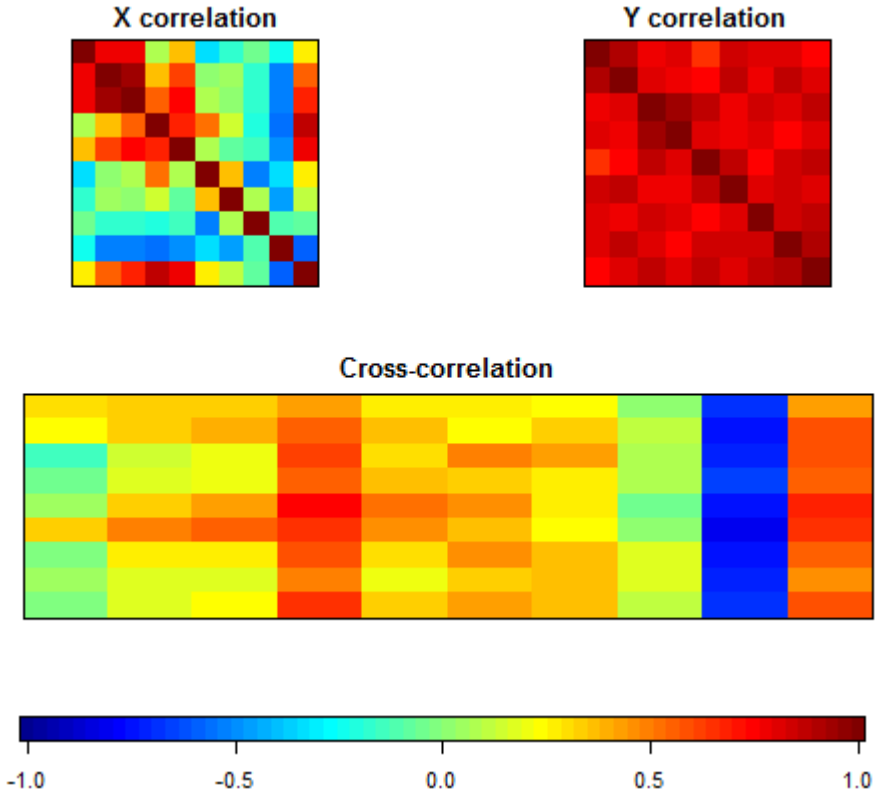
The canonical correlation analysis is carried out on the base of metric variables, and if statistical inference is one of the goals the multivariate normal distribution assumption must be held. These requirements cannot be met in many situations, especially in social sciences when survey data are analyzed. For this reason, a different procedure for sets of non-metric variables is proposed, namely nonlinear canonical correlation analysis in which the categories are converted to numerical values. This algorithm is known as "optimal quantification" or "optimal scaling" in which the variables transformed from qualitative into quantitative ones are represented by a binary indicator matrix **G** and the vector **y** comprising category quantifications (Meulman et al. 2004). It should be underlined that the expression "optimal" has a relative meaning as the solution "is always obtained with respect to the particular dataset that is analyzed and the particular criterion that is optimized" (Meulman 1998). The procedure of finding the optimal solution is an iterative process. Quantified variables are used to obtain a solution on the basis of which the quantification is updated and these steps are repeated until the final solution is found that meets the established criteria (Meulman & Heiser 2001, p.1). The nonlinear canonical correlation analysis is available in SPSS package under the OVERALS technique described in details by Van der Burg et al. (1994). A great advantage of this procedure is the possibility to represent graphically the relationships between variables belonging to the considered sets of variables. The interpretation of the results includes, among others, evaluation of the fit and the loss of the solution, the weights and the component loadings examination as well as the centroid plot presentation allowing for the assessment of associations among the categories (Meulman & Heiser 2001, p 250-271).

4 Satisfaction with chosen aspects of life and sustainable development indicators - evaluation by canonical correlation analysis

The first analysis is based on aggregated data from the EU-SILC. The average opinions on nine aspects of life given by the respondents from 28 European Union countries are compared to a set of sustainable development indicators. As all the data are metric, canonical correlation analysis was applied to identify the relations between two sets of variables. Pearson correlation coefficients between the variables in the sets as well as cross-correlations were calculated at the preliminary stage. The results are presented in a visual way in Figure 1. Such a presentation offered in CCA R package (González et al. 2008) is an attractive way of summarizing the interdependences in the datasets. Moreover, it indicates whether the application of the canonical correlation method is justified. As this technique is oriented towards highliting correlations between two data sets, there is no sense to use it if the calculated correalations are very low, which would be expressed by dominant green colour in the matrices (González et al. 2008). The patterns which can be

observed in Figure 1 indicate that relatively high correlations can be found both within the sets of variables (X and Y) and between sets. This gives the rationale to use canonical correlation analysis in order to explain the associacions. It is worth mentioning that all satisfaction with life variables turn out to be highly positvely correlated. On the contrary, they are significantly negativly correlated with the poverty and social exclusion indicator.

Figure 1. Visualizations of correlations



Note: X correlation matrix concerns the sustainable development indicators (in the following order: GAS, LEF, LEM, ASS, RES, EMP, ENE, TRA, POV, GDP), Y correlation matrix concerns the variables representing the degree of satisfaction from various aspects of life (in the following order: PER, LIF, REC, ENV, FIN, ACC, COM, JOB, TIM), Cross-correlation - the indicators are represented in columns and the satisfaction variables are given in rows.
Source: own elaboration based on Eurostat data with R package `CCA`

Substantial results of the canonical correlation analysis are presented in Table 1.

Table 1 Canonical correlation analysis results

Measure	Satisfaction set	Sustainable development indicators set
Variance extracted	100%	96,28%
Total redundancy	68,88%	57,68%
Overall results		
Canonical correlation	0,95 (p-value = 0,01759)	

Source: own elaboration based on Eurostat data with Statistica ver. 10 programme

The overall canonical correlation referring to the first and the most important pair of canonical variates is equal to 0,95 and significant (with p-value close to 0,02). Although the

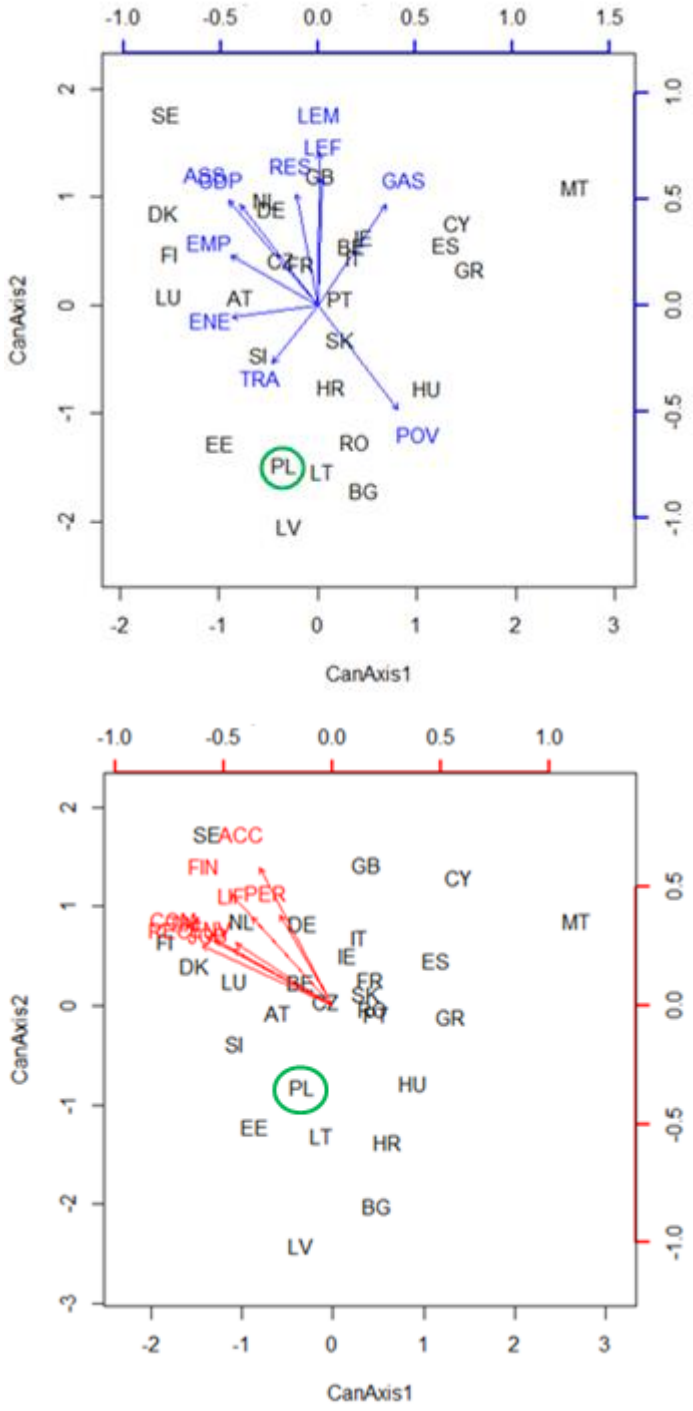
canonical correlation analysis treats the two sets equally, the sustainable development indicators can be conceptually recognized as the explanatory variables in this case. Hence, 68,88% of the variance in the satisfaction set can be captured by the known sustainable development indicators. This indicates rather strong association between the variables in the two sets.

Chi-square sequential Bartlett tests (Bartlett, 1941) led to the conclusion that only the first pair of canonical variates is statistically significant. The interpretation of the outcomes of the analysis includes also the evaluation of the contribution of the variables to the indentified multivariate relationship which can be done with use of the canonical functions coefficients and canonical structure coefficients (Thompson, 1984, p.21-25). The magnitude of the structure coefficients revealed that the biggest contribution (at least 0,5) is assigned to real GDP, employment rate of older workers, primary energy consumption, people at risk of poverty or social exclusion and official development assistance in the first set and to satisfaction with financial situation, recreational and green areas, living environment, commuting time, job and time use in the second set.

The results of the canonical correlation analysis can be presented visually by a pair of biplots (cf. Borcard et al. 2011, p. 213-214). Such an option is available in the R package `vegan` (Oksanen et al. 2015). Biplot approach was proposed by Gabriel (1971) and gives an opportunity to show the relations among the variables and objects simultaneously. The elementary interpretation rules are: (1) objects are represented by points and variables as arrows, (2) the smaller angle between arrows, the stronger association of the variables, (3) the larger projection of an object on the arrow, the bigger deviance of the object from the average (cf. Kroonenberg 2008, pp.497-498).

For practical reasons and better visual perception two-dimensional biplots presentation was chosen (Figure 2) although only the first canonical root turned out to be significant. The pair of biplots reveals that the high assessment of satisfaction with aspects of life correspond to the high level of such sustainable development variables as: real GDP, employment rate of older workers and official development assistance. It is apparent, once again, that the satisfaction with life variables are negatively correlated with the percentage of the people at risk of poverty or social exclusion. The position of Poland is indicated by the green circle. The projections on the arrows representing the variables show that the evaluation of all the considered aspects of life is close to the average. This is accompanied by relatively low values of above-mentioned sustainable development measurers (real GDP, employment rate of older workers and official development assistance) and relatively high risk of poverty or social exclusion.

Figure 2. Biplots presentation of the results of the canonical correlation analysis



Source: own elaboration based on Eurostat data with R package `vegan`

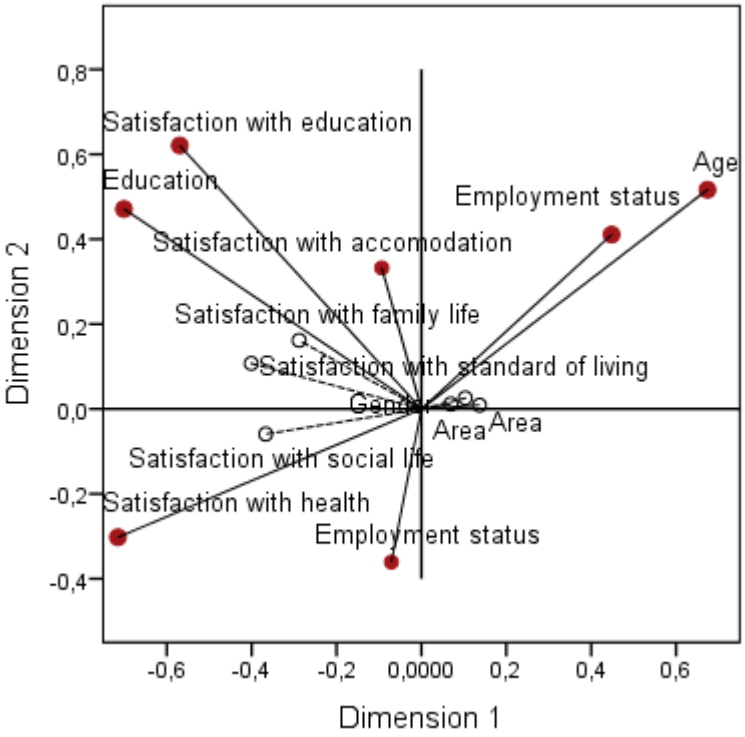
5 Satisfaction with chosen aspects of life and socio-economic background - evaluation by nonlinear canonical correlation analysis

The second analysis is based on individual data from the *Quality of Life Survey.* The opinions on six aspects of life given by the Polish respondents are compared to five

variables describing their socio-economic background. As all the data are non-metric, nonlinear canonical correlation analysis was applied to identify the relations between two sets of variables and two-dimensional solution was found.

The fit measures obtained during the analysis show that the first coordinate represents approximately 51,4 % of the captured relationship and the second coordinate covers about 48,6%. The total fit measure is equal to 1,48 and compared with the maximum fit value equal to the number of dimensions (i.e. 2) indicates that the actual fit is rather good (74%), so the results can be treated as quite informative. The partition of the fit shows that the best discriminatory power is observed for the following variables: satisfaction with education, satisfaction with accommodation, satisfaction with health, age group, education level and employment status. This is confirmed by the values of the component loadings equal to the correlation coefficients calculated between the quantified variables and the object scores (Meulman & Heiser 2001, p. 254) illustrated in Figure 3.

Figure 3. Component loadings for the analyzed data

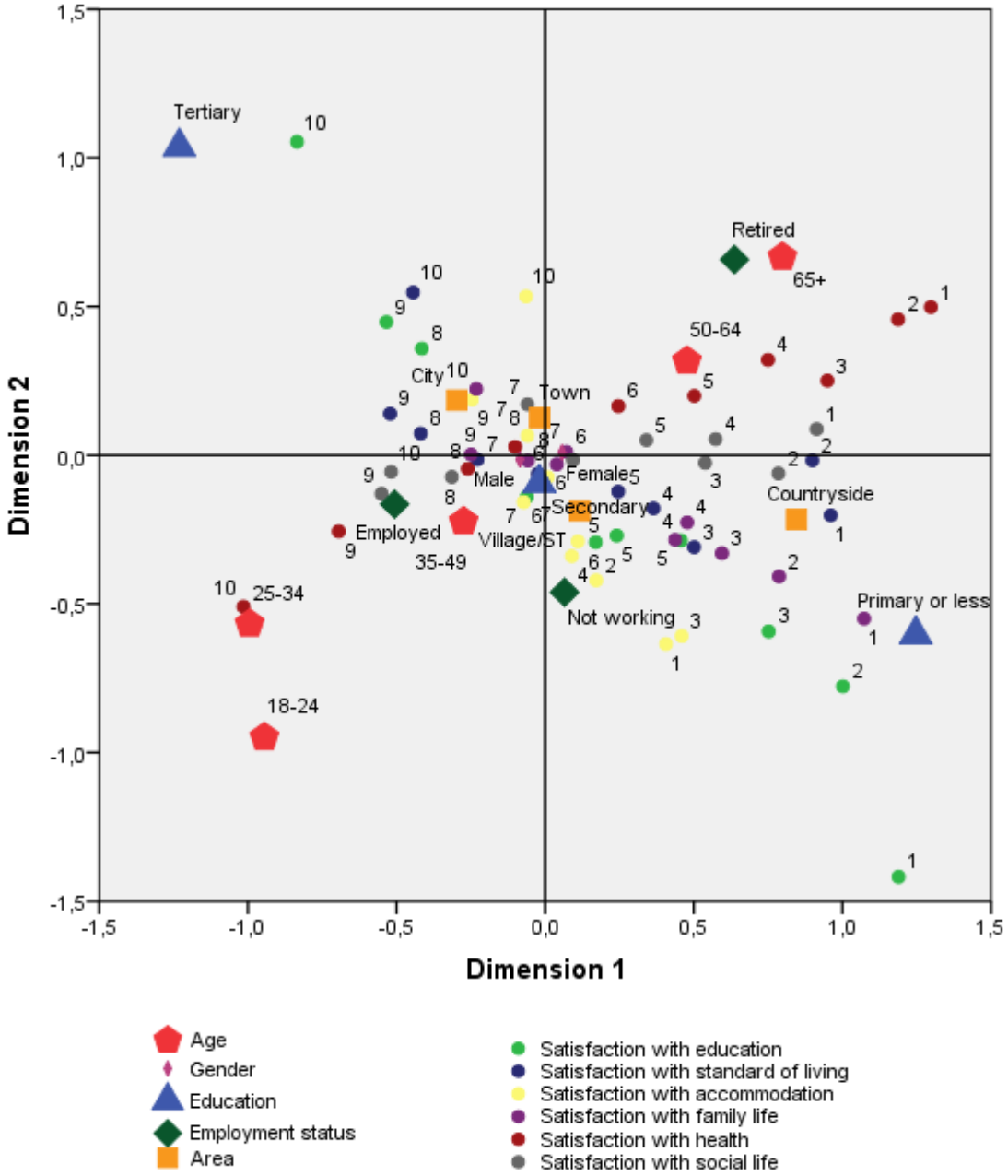


Source: own computations on the base of the Quality of Life Survey data with SPSS

The plot in Figure 3 reveals that the most important variables are the same as above-mentioned as the points representing them are located far from the origin. The employment status appears two times as it is a multiple nominal variable and the points depicting it should be considered separately. It is apparent that there are a few satisfaction variables with a relatively weaker discriminatory power, i.e. satisfaction with family life, satisfaction

with standard of living and satisfaction with social life. The detailed picture of the correspondences between categories is shown in Figure 4.

Figure 4. Graphical representation of the nonlinear canonical correlation results



Source: own computations on the base of the Quality of Life Survey data with SPSS

The configuration of points in Figure 4 allow us to identify key relationships among the degree of the satisfaction with various aspects of life and socio-economic characteristics. The evaluation of the satisfaction with education, standard of living, accommodation and family life is related to the educational level - higher levels of education are accompanied by higher satisfaction ratings. It is worth underlying the proximity of the highest satisfaction

ratings with the "tertiary" category, and the lowest score for the standard of living with the "primary" category. Another important regularity is that the health assessment is linked to the age of respondents. Living in the countryside is associated with low levels of satisfaction with standard of living and social life. In contrast, high ratings of accommodation, family life, standard of living and education seem to be specific for the residents of large cities. A very poor evaluation of accommodation conditions is typical for unemployed persons. Working respondents highly appreciate social life and standard of living. Advanced age and being retired are associated with poor health status assessment and a fairly low satisfaction with social life. There are no significant differences in the assessment of various life aspects with respect to gender of the respondents.

Conclusions

The canonical correlation analysis is a useful tool to examine and describe relations between two datasets. Various variants of the canonical correlation analysis can be applied with respect to the character of data. The application of these analytical techniques allows the researcher to identify the relations between complex and multivariate datasets. However it should be underlined that the detected relationships do not mean that there is the causality of the phenomena. The other important remark is that if canonical correlation is planned to be used for the inference, more assumptions must be met than for descriptive purposes. The attempt described in this paper allowed for combining some satisfaction with life measures with other "external" variables and some interesting associations and patterns were found. As the canonical correlation analysis encounters a lot of limitations and demands some assumptions hard to meet when considering real-life data, the outcomes should be treated as a preliminary stage for further in-depth analysis.

Acknowledgments

This research was financed by Narodowe Centrum Nauki (National Science Centre) in Poland under the project entitled "*Non-metric multivariate data analysis as a tool for study of adults situation in the context of demographic changes*". Project number: 2012/05/B/HS4/02499.

This paper uses *Quality of Life Survey* data: European Foundation for the Improvement of Living and Working Conditions, 2015, *European Quality of Life Time Series, 2007 and 2011: Open Access* [data collection]. UK Data Service: SN: 7724, http://dx.doi.org/10.5255/UKDA-SN-7724-1.

References:

BARTLETT, M. S. (1941). The statistical significance of canonical correlations. *Biometrika*, Vol. 32, No. 1, pp. 29-37.

BERNSTEIN, I. H., GARBIN, C.P., TENG, G.K. (1988). *Applied multivariate analysis*. Springer-Verlag New York Inc.

BORCARD, D., GILLET, F., LEGENDRE, P. (2011). *Numerical ecology with R.* Springer Science & Business Media.

CZAPIŃSKI, J., PANEK, T. (Ed.) (2015). *Diagnoza społeczna 2015*. Rada Monitoringu Społecznego, Warszawa. Available at: http://www.diagnoza.com [retrieved 2015-12-14].

EUROPEAN FOUNDATION FOR THE IMPROVEMENT OF LIVING AND WORKING CONDITIONS. (2015). *European Quality of Life Time Series, 2007 and 2011: Open Access* [data collection]. UK Data Service: SN: 7724, http://dx.doi.org/10.5255/UKDA-SN-7724-1.

EUROPEAN UNION (2015). *Sustainable development in the European Union. 2015 monitoring report of the EU Sustainable Development Strategy*. Eurostat, Luxembourg: Publications Office of the European Union.

GABRIEL, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58, No. 3, pp. 453-467.

GITTINS, R. (1985). *Canonical analysis: a review with applications in ecology*. Springer Verlag.

GONZÁLEZ, I., DÉJEAN, S., MARTIN, P. G., BACCINI, A. (2008). CCA: An R package to extend canonical correlation analysis. *Journal of Statistical Software*, *23*(12), pp. 1-14.

HAERDLE, W., SIMAR, L. (2003). *Applied multivariate statistical analysis*. Springer, Berlin.

HOTELLING H. (1935). The most predictable criterion. *Journal of Educational Psychology*, 26, pp. 19-36.

HOTELLING, H. (1936). Relation between two sets of variables. *Biometrica*, 28, pp. 321-377.

KROONENBERG, P. M. (2008). *Applied multiway data analysis*. John Wiley & Sons.

MEULMAN, J. J. (1998). *Optimal scaling methods for multivariate categorical data analysis*. SPSS White Paper, Chicago.

MEULMAN, J. J., HEISER, W. J. (2001). *SPSS Categories 11.0*. SPSS, Chicago.

MEULMAN, J. J., VAN DER KOOIJ, A. J., HEISER, W. J. (2004). *Principal Components Analysis with Nonlinear Optimal Scaling Transformations for Ordinal and Nominal Data.* [in:] Kaplan D. (ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences*. Sage Publications, Thousand Oaks.

OKSANEN,, J., BLANCHET, F. G., KINDT, R., LEGENDRE, P., MINCHIN, P. R., O'HARA, R. B., SIMPSON, G. L., SOLYMOS, P., HENRY, M., STEVENS, H., WAGNER, H. (2015). *Package 'vegan'. Community Ecology Package* ver. 2.3-2.

RAYKOV, T., MARCOULIDES, G. A. (2008). *An introduction to applied multivariate analysis*. Routledge.

RENCHER, A. C. (2003). *Methods of multivariate analysis.* John Wiley & Sons.

SHARMA, S. (1996). Applied multivariate techniques. John Wily & Sons Inc.

THOMPSON, B. (1984). *Canonical correlation analysis.Uses and interpretation*. Sage Pubns., Beverly Hills and London.

VAN DER BURG, E., DE LEEUW, J., DIJKSTERHUIS, G. (1994). OVERALS: Nonlinear canonical correlation with k sets of variables. *Computational Statistics & Data Analysis*, *18*(1), pp. 141-163.

