Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663
www.czasopisma.uni.lodz.pl/foe/
6(332) 2017

DOI: http://dx.doi.org/10.18778/0208-6018.332.07

Małgorzata Misztal
University of Łódź, Faculty of Economics and Sociology, Department of Statistical Methods,
mmisztal@uni.lodz.pl

On the Use of Redundancy Analysis to Study the Property Crime in Poland

Abstract: Redundancy analysis (RDA) is a canonical form of principal components analysis (PCA) and is one of, so-called, linear ordination techniques. The goal of ordination is to represent objects and response variables relationships as faithfully as possible in a low-dimensional space. Redundancy analysis is also a technique of exploratory data analysis. Graphical presentation of the results using the ordination biplots or triplots can facilitate the analysis of the relationship between the variation in the set of the response variables and the variation of the explanatory variables. In the paper, redundancy analysis was applied to assess the relationships between the selected socio-economic factors and the intensity of the crime against property in Poland.

Keywords: redundancy analysis, ordination diagram, crime against property

JEL: C38, K42

1. Introduction

Redundancy analysis (RDA; Rao, 1964; van den Wollenberg, 1977) is a canonical form of principal components analysis (PCA) and is one of, so-called, linear ordination techniques.

In multivariate statistics, ordination is "the process of reducing the dimensionality (i.e. the number of variables) of multivariate data by deriving a small number of new variables that contain much of the information in the original data. The reduced data set is often more useful for investigating possible structure in the observations" (Everitt, Skrondal, 2010: 312).

Redundancy analysis and other ordination techniques are very popular in ecological research but almost completely unknown in, for example, socio-economic research. The advantages of these methods include, among others, the possibility of graphical presentation of the analysis results in two-dimensional space using ordination plots (biplots and triplots). Graphical presentation plays an important role in the interpretation of the results.

The goal of the paper is to analyze the relationships between some socio-economic factors and the intensity of the crime against property in Poland with the use of redundancy analysis. The so-called variation partitioning procedure, proposed by Borcard et al. (1992) and Peres-Neto et al. (2006), was also applied to apportion the variation of the property crime data among the subsets of the analyzed explanatory variables.

All the calculations were performed using CANOCO for Windows software.

2. Material and methods

2.1. Data sets

Intensity of six types of property crime in Poland in 2014 (by voivodships, per hundred thousand inhabitants) were analyzed. These were: robbery, theft, embezzlement, burglary, car theft and criminal damage (see: *Raport o stanie bezpieczeństwa w Polsce w 2014 roku*).

Many different factors that influence crime rate are described in the literature. These are, among others, age and sex of the offender, educational level, unemployment, poverty, income inequality, population density, urbanization rates, living conditions, economic growth, alcohol consumption and crime detection rates (see e.g. Sztaudynger, Sztaudynger, 2003; Szczepaniec, 2012; Bieniek et al., 2012; Kądziołka, 2014; 2015).

For the purposes of the study almost 30 different socio-economic factors, which influence the crime against property rate, were initially analyzed. The para-

metric Hellwig's method and the method of inverse correlation matrix were applied to select a set of explanatory variables. Finally, four variables were used in further analysis: the percentage of urban population in total population, the percentage of unemployed persons with at most lower secondary education, unemployed persons seeking work for 13 months or more (in % of total unemployed; see: *Rocznik Statystyczny Województw*, 2015) and alcohol consumption (1 – low/2 – medium/3 – high; see: *Spożycie alkoholu w Polsce w 2012 r. Raport z badania*, 2013). Relationships between response variables (types of property crime) and explanatory variables were linear.

To examine the interrelationships between two sets of variables, **Y** (*m* response variables) and **X** (*p* explanatory variables), redundancy analysis (RDA) can be used.

2.2. Redundancy analysis

Redundancy analysis is a method combining regression with principal component analysis (PCA) and can be described as a direct extension of regression analysis to model multivariate response data (Borcard et al., 2011).

RDA consists of two steps (Legendre, Legendre, 2012). Step 1 is a multivariate regression of **Y** on **X** leading to a matrix of fitted values $\hat{\mathbf{Y}}$ through the linear equation:

$$\hat{\mathbf{Y}} = \mathbf{X}\left[\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}. \tag{1}$$

Step 2 is a principal component analysis of $\hat{\mathbf{Y}}$ Both – the fitted values of the multivariate linear regression and the canonical axes – are linear combinations of all explanatory variables in **X**. In other words (Borcard et al., 2011: 155), the RDA algorithm produces, "in successive order, a series of linear combinations of the explanatory variables that best explain the variation of the response matrix". The canonical axes are orthogonal to one another.

Some informative statistics can be computed after the first step of RDA (see eg. Legendre, Legendre, 2012: 632–633). The canonical R^2 is a measure of the linear relationships between **Y** and **X** (the proportion of the variation of **Y** explained by a linear model of the variables in **X**):

$$R^2_{\frac{\mathbf{Y}}{\mathbf{X}}} = \frac{SS\left(\hat{\mathbf{Y}}\right)}{SS\left(\mathbf{Y}\right)} \tag{2}$$

where $SS\left(\hat{\mathbf{Y}}\right)$ is the total sum of squares of $\hat{\mathbf{Y}}$ and $SS(\mathbf{Y})$ is the total sum of squares of **Y**.

The adjusted R^2 can be computed as (Ezekiel, 1930):

$$R^2_{adj} = 1 - \frac{(n-1)}{(n-p-1)}\left(1 - R^2_{\mathbf{Y}\atop\mathbf{X}}\right), \tag{3}$$

where p is the number of explanatory variables in **X** and n – the number of observations.

Graphical presentations of the RDA results are biplots and triplots. An RDA biplot presents objects as points and either response or explanatory variables as vectors. In a triplot, objects are presented as points while both response and explanatory variables as vectors (arrows). Levels of nominal variables are plotted as points.

The interpretation of these plots depends on what type of scaling has been chosen. In general, type I scaling (focus on objects) should be considered if the distances between objects are of particular value or if most explanatory variables are binary or nominal. Type II scaling (focus on response variables) should be considered if the correlative relationships between variables are of more interest (for more details see: Legendre, Legendre, 2012; Lepš, Šmilauer, 2003).

2.3. Variation partitioning

It is often possible to identify in the set of explanatory variables **X** two or more subsets of variables representing different classes. The so-called variation partitioning procedure, proposed by Borcard et al. (1992) and improved by Peres-Neto et al. (2006), can be applied to apportion the variation of **Y** among the subsets of predictor variables **X**.

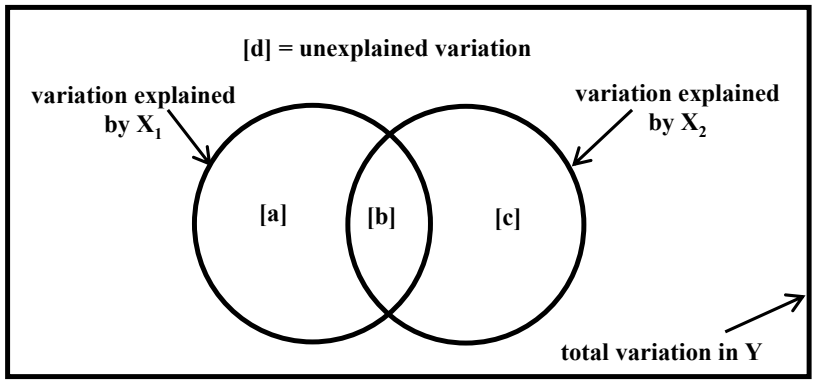

Figure 1. Venn diagram showing the variation partitioning scheme according to two sets of explanatory variables

Source: based on Legendre, Legendre, 2012: 572

For two explanatory data sets $\mathbf{X}_1$ and $\mathbf{X}_2$ the total variation of $\mathbf{Y}$ can be partitioned into four fractions as in Figure 1. Fraction [a + b + c] is based on both sets of predictor variables, fraction [a + b] based on data set $\mathbf{X}_1$, fraction [b + c] based on data set $\mathbf{X}_2$ and fraction [d] is the residual fraction not explained by $\mathbf{X}_1$ and $\mathbf{X}_2$.

The following steps are needed to partition the total variation into the four fractions [a], [b], [c], [d] (Borcard et al., 2011):

1. Perform RDA of the response matrix $\mathbf{Y}$ by:
   - $\mathbf{X}_1$ to obtain $R^2_{\mathbf{Y}/\mathbf{X}_1} = [a+b]$;
   - $\mathbf{X}_2$ to obtain $R^2_{\mathbf{Y}/\mathbf{X}_2} = [b+c]$;
   - $\mathbf{X}_1$ and $\mathbf{X}_2$ together to obtain $R^2_{\mathbf{Y}/[\mathbf{X}_1;\mathbf{X}_2]} = [a+b+c]$.
2. Compute the adjusted R^2 (Eq. 3) for the three RDAs above;
3. Compute the four fractions of adjusted variation as:
   - $[c]_{adj} = [a+b+c]_{adj} - [a+b]_{adj}$;
   - $[c]_{adj} = [a+b+c]_{adj} - [a+b]_{adj}$;
   - $[b]_{adj} = [a+b]_{adj} - [a]_{adj} = [b+c]_{adj} - [c]_{adj}$;
   - $[d]_{adj} = 1 - [a+b+c]_{adj}$.

Variation partition procedure for three sets of explanatory variables was presented by Anderson and Gribble (1998) and Cushman and McGarigal (2002). Permutation tests of the significance of individual components in explaining the variation in the response data can be used (Legendre, Legendre, 2012).

3. Results

All five canonical axes explain 78.1% of the total variability. Two of them were used for further analysis. The first canonical axis explains 81.9% of variability in the canonical space (64% of the total variability) and the second axis – 9.5% and 7.4%, respectively. However, only the first canonical axis turned out to be statistically significant (based on the permutation tests described in detail by Legendre et al., 2011 – $p = 0.001$ for the first and $p = 0.298$ for the second canonical axis).

The RDA triplot (type II scaling) for property crime data is presented in Figure 2. Objects (voivodships) are ordinated as black points, response and quantitative explanatory variables as arrows (solid black and dashed grey respectively)

and nominal explanatory variables (recoded on a set of dummy variables) as grey triangles.

The angles between all vectors on the RDA triplot reflect their linear correlation. The approximated correlation between two variables is equal to the cosine of the angle between the corresponding vectors. Perpendicular vectors indicate the lack of correlation between the variables they represent. The angle less than 90° suggests positive correlation between variables and the angle approaching 180° – strong negative correlation between variables.

Projection of the centroids of dummy explanatory variables onto the response variable arrow gives the approximation of the average values of this response variable in the individual classes of nominal predictor variable.

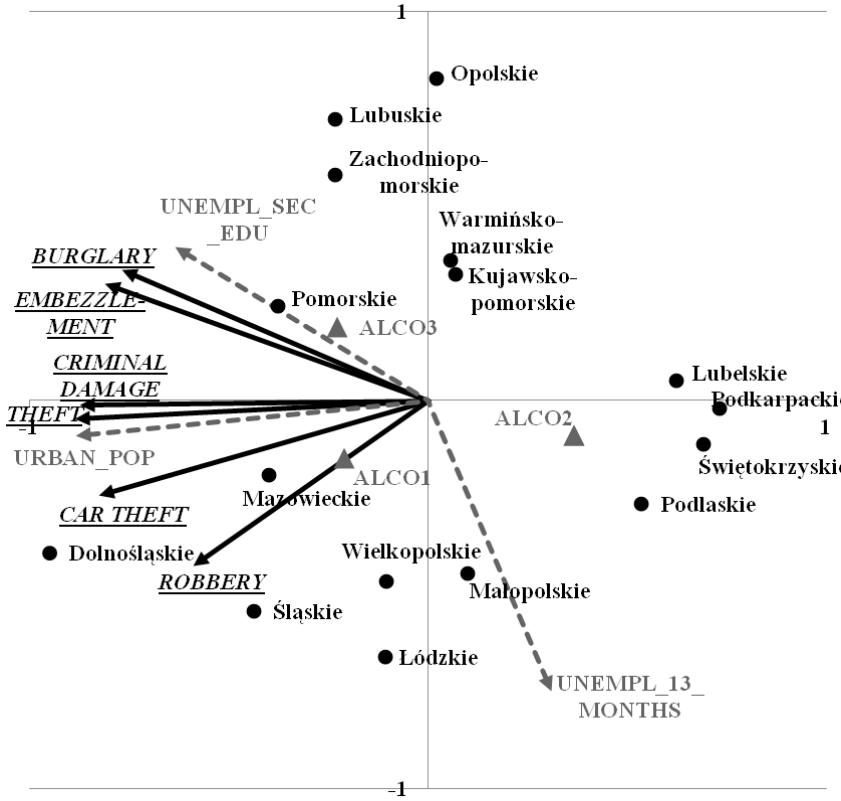

Figure 2. RDA ordination triplot (type II scaling) of the property crime data URBAN_POP – the percentage of urban population in total population; UNEMPL_SEC_EDU – the percentage of unemployed persons with at most lower secondary education; UNEMPL_13_MONTHS – the percentage of unemployed persons seeking work for 13 months or more; ALCO – alcohol consumption

Source: based on own calculations using CANOCO software

The following relationships between response and explanatory variables can be observed on the RDA triplot in Figure 2:

1) strong positive correlations between the percentage of urban population in total population and the crime intensity for all types of property crime;
2) positive correlations between the percentage of unemployed persons with at most lower secondary education and the crime intensity for all types of property crime (the strongest with burglary, the weakest with robbery);
3) weak positive correlation between the percentage of unemployed persons seeking work for 13 months or more and robbery and negative correlations between the percentage of unemployed persons seeking work at least 13 months and the crime intensity for the other types of property crime (the strongest correlations with burglary and embezzlement);
4) in the voivodships with alcohol consumption at an average level (ALCO2) crimes against property are less frequent compared to other classes of alcohol intake.

Much more detailed interpretation of the RDA results presented in Figure 2 can be found in Misztal (2017).

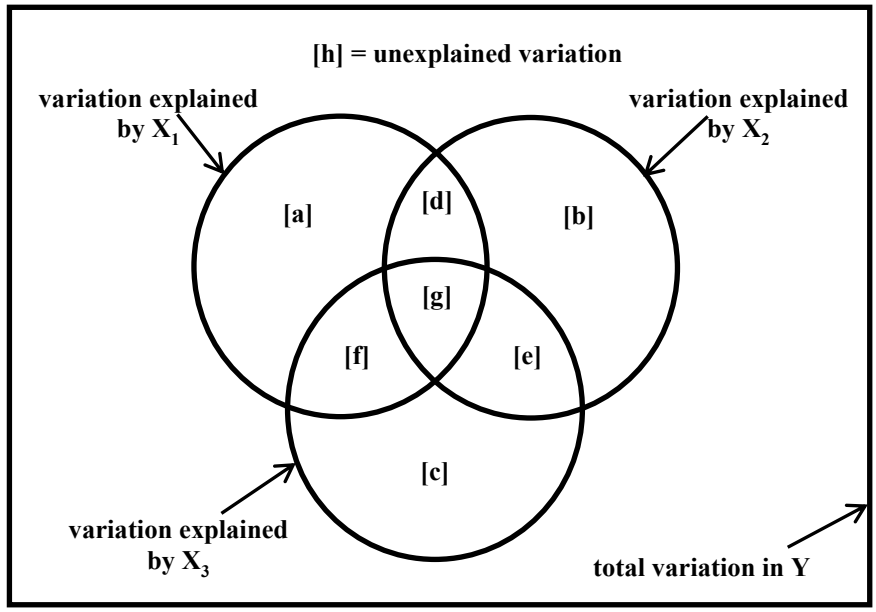

Figure 3. Venn diagram showing the variation partitioning scheme according to three sets of explanatory variables

Source: own elaboration based on Anderson and Gribble (1998)

Three subsets of explanatory variables can be distinguished in the set of explanatory variables **X**: $\mathbf{X}_1$ – containing the percentage of urban population in total population, $\mathbf{X}_2$ – containing both unemployment characteristics: the percentage of unemployed persons with at most lower secondary education and the percentage of unemployed persons seeking work for 13 months or more and $\mathbf{X}_3$ – containing the volume of alcohol consumption. The variation partitioning procedure (Borcard et al., 1992; Peres-Neto et al., 2006) can be applied to apportion the variation of **Y** (types of property crimes) among the selected subsets of predictor variables. For three explanatory data sets $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ the total variation of **Y** can be partitioned into eight fractions as in Figure 3.

Seven RDAs are needed to decompose the total variation into eight fractions. The RDA results according to required combinations of explanatory sets are presented in Table 1. The computed eight fractions of adjusted variation are shown in Table 2.

Table 1. The RDA results for different combinations of explanatory variables

Explanatory set	Part of variation	Explained variation (%)		p-value
		R^2	R^2_{adj}	
$\mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3$	[a] + [b] + [c] + [d] + [e] + [f] + [g]	78.1	67.2	0.0001
$\mathbf{X}_1 + \mathbf{X}_2$	[a] + [b] + [d] + [e] + [f] + [g]	62.3	53.0	0.0004
$\mathbf{X}_1 + \mathbf{X}_3$	[a] + [c] + [d] + [e] + [f] + [g]	69.3	61.5	0.0001
$\mathbf{X}_2 + \mathbf{X}_3$	[b] + [c] + [d] + [e] + [f] + [g]	53.0	35.9	0.0225
$\mathbf{X}_1$	[a] + [d] + [f] + [g]	50.6	47.1	0.0003
$\mathbf{X}_2$	[b] + [d] + [g] + [e]	34.6	24.7	0.0285
$\mathbf{X}_3$	[c] + [e] + [f] + [g]	36.0	26.1	0.0252

Source: own elaboration using CANOCO software

Table 2. The variation partitioning procedure results

Part of variation	Explained variation (%)
[a]	31.3
[b]	5.7
[c]	14.2
[d]	4.1
[e]	0.2
[f]	–2.9
[g]	14.6
total explained: [a] + [b] + [c] + [d] + [e] + [f] + [g]	67.2
unexplained variation [h]	32.8

Source: own elaboration using CANOCO software

The obtained results show that 67.2% of the total variation in property crime data can be explained by all explanatory variables. The set $\mathbf{X}_1$ explains 47.1% of the total variation and this is almost twice as much as the each of the next two sets, $\mathbf{X}_2$ and $\mathbf{X}_3$, explain (24.7% and 26.1% respectively).

The unique contribution of the percentage of urban population in total population (fraction [a] = 31.3%) is much larger than that of the volume of alcohol consumption (fraction [c] = 14.2%) and that of the unemployment (fraction [b] = 5.7%).

The variation explained jointly by the three sets (fraction [g] = 14.6%) is quite large. This indicates that variables in the three sets are intercorrelated.

Fraction [f] (the $\mathbf{X}_1$ and $\mathbf{X}_3$ shared variation) is negative. Legendre and Legendre (2012: 573) stress that it may happen because "it is not a rightful measure of variance". Negative fraction [f] indicates that two groups of variables: $\mathbf{X}_1$ and $\mathbf{X}_3$, together, explain variation of $\mathbf{Y}$ better than the sum of the individual effect of these groups. Peres-Neto et al. (2006: 2615) give two explanations of the negative value of the shared variation. It may be negative due to suppressor variables or due to two strongly correlated predictors with strong effects on $\mathbf{Y}$ of opposite signs. According to Borcard et al. (2011: 182) negative R^2_{adj} can be ignored (considered as null).

4. Final remarks

Redundancy analysis is a technique of exploratory data analysis. Graphical presentation of the RDA results, using the ordination biplots or triplots, can facilitate the analysis of the relationship between the variation in the set of the response variables and the variation of the set of the explanatory variables. The obtained results concerning the relationships between the proposed socio-economic factors and the incidence of crime against property are justified in the literature.

Unemployment is indicated in the literature as one of the main factors affecting the crime, but its impact is not clearly defined. Positive correlations between the percentage of unemployed persons with at most lower secondary education and the crime intensity for all types of property crime were observed. A similar, significant relationship between the percentage of unemployed with low educational level and the frequency of the crime against property was shown by (Kądziołka, 2014).

Strong negative correlations between the percentage of unemployed persons seeking work for at least 13 months and burglary and embezzlement were also observed. Szczepaniec (2012: 168) quotes the results of research showing, among others, that the unemployed and the poor spend more time at home out of necessity so they can protect their possessions from criminals. The long-term unemployed are likely to be women and the elderly and the statistics show that the crimes are most often committed by men at a young age (Szczepaniec, 2012: 170).

Strong positive correlations between the percentage of urban population in total population and the crime intensity for all six types of property crime were also observed. Crime against property is typical for urban areas and most of the crimes

against property are committed in the public spaces: parks, shopping centers, railway stations, etc. (Kądziołka, 2014: 17–18).

The proposed set of explanatory variables explains 67.2% of adjusted variation in the property crime data. According to the variation partitioning results, almost half of the adjusted variation (31.3%) has been explained solely by the percentage of urban population in total population.

References

Anderson M.J., Gribble N.A. (1998), *Partitioning the variation among spatial, temporal and environmental components in a multivariate data set*, "Australian Journal of Ecology", vol. 23, pp. 158–167.

Bieniek P., Cichocki S., Szczepaniec M. (2012), *Czynniki ekonomiczne a poziom przestępczości – badanie ekonometryczne*, "Zeszyty Prawnicze", vol. 12(1), pp. 147–172.

Borcard D., Gillet G., Legendre P. (2011), *Numerical Ecology with R*, Springer, New York–Dordrecht–London–Heidelberg.

Borcard D., Legendre P., Drapeau P. (1992), *Partialling out the spatial component of ecological variation*, "Ecology", vol. 73(3), pp. 1054–1055.

Cushman S.A., McGarigal K. (2002), *Hierarchical, multi-scale decomposition of species-environment relationships*, "Landscape Ecology", vol. 17, pp. 637–646.

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

Ezekiel M. (1930), *Methods of correlational analysis*, Wiley, New York.

Kądziołka K. (2014), *Wpływ wybranych czynników o charakterze społeczno-ekonomicznym na przestępczość przeciwko mieniu w Polsce*, [in:] W. Szkutnik (ed.), *Zarządzanie ryzykiem kapitałowym i ubezpieczeniowym oraz społecznymi uwarunkowaniami ryzyka rynku pracy*, "Studia Ekonomiczne", vol. 181(14), pp. 11–23.

Kądziołka K. (2015), *Bezrobocie, ubóstwo i przestępczość w Polsce. Analiza zależności na poziomie województw*, "Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach", vol. 242, pp. 71–84.

Legendre P., Legendre L. (2012), *Numerical Ecology*, Elsevier Science B.V., Amsterdam.

Legendre P., Oksanen J., ter Braak C.F.J. (2011), *Testing the significance of canonical axes in redundancy analysis*, "Methods in Ecology and Evolution", vol. 2, pp. 269–277.

Lepš J., Šmilauer P. (2003), *Multivariate Analysis of Ecological Data using CANOCO*, Cambridge University Press, Cambridge.

Misztal M. (2017), *Wizualizacja wyników liniowych technik ordynacyjnych na przykładzie analizy przestępczości przeciwko mieniu w Polsce*, "Taksonomia 28. Klasyfikacja i analiza danych – teoria i zastosowania. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu", no. 468, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.

Peres-Neto P.R., Legendre P., Dray S., Borcard D. (2006), *Variation partitioning of species data matrices: estimation and comparison of fractions*, "Ecology", vol. 87(10), pp. 2614–2625.

Rao C.R. (1964), *The Use and Interpretation of Principal Component Analysis in Applied Research*, "Sankhyā: The Indian Journal of Statistics", Series A (1961–2002), vol. 26, no. 4, pp. 329–358.

Raport o stanie bezpieczeństwa w Polsce w 2014 roku, Ministerstwo Spraw Wewnętrznych, http://isp.policja.pl/isp/aktualnosci/7789,Raport-o-stanie-bezpieczenstwa-w-Polsce-w–2014-r.html [accessed: 31.08.2016].

Rocznik Statystyczny Województw (2015), GUS, Warszawa.

Spożycie alkoholu w Polsce w 2012 r. Raport z badania (2013), TNS Polska, www.tnsglobal.pl/jakpijapolacy/pdf/raport.pdf [accessed: 31.08.2016].
Szczepaniec M. (2012), *Bezrobocie jako czynnik kształtujący wskaźniki przestępczości*, "Zeszyty Prawnicze", vol. 12(3), pp. 165–176.
Sztaudynger J.J., Sztaudynger M. (2003), *Ekonometryczne modele przestępczości*, "Ruch Prawniczy, Ekonomiczny i Socjologiczny", Rok LXV, vol. 3, pp. 127–143.
Wollenberg A.L. van den (1977), *Redundancy analysis. An alternative for canonical correlation analysis*, "Psychometrika", vol. 42, no. 2, pp. 207–219.

O zastosowaniu analizy redundancji do badania przestępczości przeciwko mieniu w Polsce

Streszczenie: Analiza redundancji (RDA) jest kanoniczną formą analizy głównych składowych (PCA) i należy do tzw. liniowych technik ordynacyjnych. Celem analiz ordynacyjnych jest przedstawienie związków między obiektami i zmiennymi objaśnianymi/objaśniającymi w przestrzeni o jak najniższym wymiarze. Analiza redundancji jest także jedną z metod eksploracyjnej analizy danych. Graficzna prezentacja wyników z wykorzystaniem diagramów ordynacyjnych (biplotów i triplotów) może ułatwić analizę powiązań między zmiennością rozkładów badanych zmiennych i czynnikami mogącymi wpływać na tę zmienność. W artykule zastosowano analizę redundancji do oceny zależności między wybranymi czynnikami społeczno-ekonomicznymi a poziomem przestępczości przeciwko mieniu w Polsce.

Słowa kluczowe: analiza redundancji, diagram ordynacyjny, przestępczość przeciwko mieniu

JEL: C38, K42

Received: 2016-12-16; verified: 2017-09-21. Accepted: 2017-11-06