EKONOMETRIA ECONOMETRICS 1(43) • 2014

ISSN 1507-3866

**Agnieszka Stanimir**
Uniwersytet Ekonomiczny we Wrocławiu

# CORRESPONDENCE ANALYSIS OF TIME SERIES DATA OF RESULTS OF GYMNASIUM EXAMS IN POLAND

**Summary:** The aim of this article is to present the possibilities of using correspondence analysis in the study of changes in the gymnasium exam results in Poland. The study used exam results from 2003 and 2010 of the Regional Examination Board, which consist of two regions: Lower Silesia and Opolskie. The usefulness of correspondence analysis using multiway contingency tables and correspondence analysis of supplementary points is shown. From the examinations' differences between the levels of exam results in both years, gender and place of exam are indicated.

**Keywords:** correspondence analysis, multiway contingency tables, supplementary points, gymnasium exam.

DOI: 10.15611/ekt.2014.1.03

## 1. Introduction

The data available in analysis concerned the exam results. To present the results of the analysis it is necessary to present the research problem: what are the changes in exam results for the period 2003-2010 taking into account gender and the place of exam[1].

The aim of this article is to present the possibilities of using correspondence analysis in the study of changes in gymnasium exam in Poland. This exam is conducted in Poland since 2002. It was introduced following a change in the education system. In the current Polish system of education gymnasium is lower secondary school. Due to the importance of conducting comparisons of changes in gymnasium exam results, it is important to present the method of analysis. Correspondence analysis is a method that is appropriate here, both in terms of the variables measurement scale and clarity graphical interpretation.

[1] Database of the Lower Silesia Regional Examination Board.

## 2. Correspondence analysis

Correspondence analysis is a multivariate technique of exploratory data analysis. This method is suitable for the analysis of nominal variables. A simple correspondence analysis is based on recording the observed frequencies of the categories of the analyzed variables in the contingency table. A detailed description of the algorithm of correspondence analysis can be found in many works, e.g. K. Backhaus et al. [2003], J. Blasius (1998), S.E. Clasusen [1998], M. Greenacre [1984; 1993; 2006], A. Stanimir [2005].

The procedure in simple correspondence analysis, as well as the techniques of many nominal variables, is based on the singular value decomposition (SVD) of the selected matrix (in the case of symmetric matrices, eigenvalues decomposition is used). On this basis it is possible to obtain the coordinate of projection in full space for categories of rows and columns.

The results of the correspondence analysis are mostly presented in low dimension space. For this reason, it is necessary to define the criteria for assessing the quality of the presentation. For the analysis of contingency tables the $\chi^2$-test of independence is very popular. This test determines whether the variables are dependent. The $\chi^2$ statistic is then used to assess the quality of representation in low-dimensional space. During SVD singular values are calculated, and their square values are the eigenvalues. Among the highlighted indicators there is a close relationship:

$$\frac{\chi^2}{n} = \lambda = \sum_{k=1}^{K} \gamma_k^2, \tag{1}$$

where:

- $\chi^2$ is a value of $\chi^2$ statistics for the analyzed contingency table;
- $\lambda$ is a total inertia, $\lambda = \sum_{k=1}^{K} \lambda_k$, $\lambda_k$ – eigenvalue of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ and $\mathbf{A}\mathbf{A}^{\mathrm{T}}$, where matrix $\mathbf{A}$ is a matrix in SVD;
- $\lambda_k = \gamma_k^2$ – squares singular values of $\mathbf{A}$;
- $\mathbf{\Gamma}$ $(k \times k)$ is the diagonal matrix of singular values $\gamma_k$ $(k = 1, \ldots, K)$, $\mathbf{\Lambda}$ $(k \times k)$ is the diagonal matrix of eigenvalues $\lambda_k$ $(k = 1, \ldots, K)$, wherein $\mathbf{\Lambda} = \mathbf{\Gamma}^2$;
- $K$ is the rank of matrix $\mathbf{A}$, and dimension of full space.

On the basis of the eigenvalue, the diagram with an elbow criterion is created and on these grounds the dimension of space for the presentation of the analysis results is chosen. In assessing the quality of representation there are also considered the degrees of explaining of the total inertia $\lambda$ by principal inertias $\lambda_k$:

$$\tau_{K^*} = \frac{\sum_{k=1}^{K^*} \lambda_k}{\sum_{k=1}^{K} \lambda_k} = \frac{\sum_{k=1}^{K^*} \lambda_k}{\lambda}, \tag{2}$$

where $K^*$-dimensional space of projection $\left(K^* \leq K\right)$.

Correspondence analysis is also used in cases where the number of variables is more than two. In such a situation it is necessary to use other tables containing the observed frequencies of the variables:

– Multivariate indicator matrix. The matrix **Z** consists of blocks (indices matrix) corresponding to each subsequent variable. The number of rows in each indices matrix (blocks) is the same and equal to the number of units in the survey (respondents or objects). The number of columns of each block is different and equal to the number of categories of variable corresponding to the block. This approach suggests a "... strict zero/one logical coding of the indicator matrix and use values between 0 and 1 as well, which is known as 'fuzzy' coding" [Greenacre 1984]. The number of ones in each block is equal to the number of objects. In each row the number of ones is equal to the number of variables, and indicates the categories to which each object (respondent) belongs. Assuming that $J_q$ ($q = 1, \ldots, Q$) is the number of categories of variable $q$, so in the row of matrix **Z** the number of ones will be equal to $Q$. The number of ones in the whole matrix **Z** will be equal to $n \times Q$, where $n$ is the number of units (objects).
– Burt matrix. In this table in rows and columns there are the same categories, so this is a symmetric block matrix. Down the diagonal lie square tables cross-tabulating each variable with itself. These are diagonal matrices of marginal frequencies of each variable. Off-diagonal submatrices of the Burt matrix are contingency tables for each pair of variables.
– Multiway contingency table. This table arises from the cross-classification of categories, for example, of three nominal variables. The table may be viewed as having rows, columns and layers. A layer may be situated in rows or in columns. It depends on which variable categories should have been broken down by categories of a third variable. If in the study are four variables, the layers may be in rows and columns simultaneously.
– Concatenated tables. This is a block-matrix composed of several contingency tables cross-tabulating one variable with several other variables, "...the cross-tables are stacked one on top each other, i.e. row-wise" [Greenacre 2006, p. 21]. The total amount of each table is similar to the sample size.

CA can be extended to analyze order data, rankings and preferences, paired comparison data and multiresponse tables.

During the interpretation of results of the correspondence analysis it is important to evaluate:

– the position of the point towards the centroid (origin); if the point is located (compared to other points position) very close to the origin, this means that its profile has a value close to the average profile, "... points far away from origo indicates a clear deviation from ... independence" [Andersen 1991, p. 372].
– the location of a point relative to other points of the categories belonging to the same variable; in this case this can be done by summing in a contingency table of the observed frequencies of these two categories into one category [Clausen 1998; Jobson 1992].

– the location of a point relative to another point describing the category of the other variable; points lying close together indicate a relationship between the categories, if the categories do not occur together, the representing points should be located on opposite sides of the origin.

## 3. Correspondence analysis of time series data

In literature can be found examples of correspondence analysis in the study where time series data are taken into account.

J.C. Dore, T. Ojasoo [2001] analyzed the following variables: 48 countries, publications in 18 disciplines in 12 years (1981-1992). In the conducted study the classical correspondence analysis for each pair of variables was used.

P. Heijden, J. Teunissen, C. Orlé [1997] analyzed using the multiway contingency table level of the Dutch educational system, the number of classes in educational level in nine school years (1985/1986-1993/1994).

A multiway contingency table is also used in work of P. Heijden, A. Mooijaart, Y. Takane [1994]. They investigated relationships between: faculties, first year students and graduate students in six analysed years.

In the same way there were analysed data of leisure activities, age, educational level in five analysed years by T. Muller-Schneider [1994].

A very interesting paper was written by V. Thiessen, H. Rohlinger, J. Blasius [1994]. They used CA of contingency table with supplementary variables (additional points). The study was conducted on two groups of data: reference (active) data (tasks in households, wife-husband responsibility in 1977) and supplementary variables (years from 1978 and 1980). This way of analysis gives possibilities of the projection of categories of additional (supplementary, non-active) variables made on an already existing configuration for reference data. Active variables of the analysis determine the solution space. C coordinates of supplementary points are calculated on the basis of the analysis of active points.

To analyse changes in secondary school exam results a combination of two techniques: multiway contingency table and supplementary variables was used.

## 4. Gymnasium exam and the analyzed data

The aim of the analysis is to check the applicability of the correspondence analysis in the study of changes in the results of the gymnasium exam. For this reason it was necessary to gather the results achieved by students in different years.

Figure 1 shows the levels of education in Poland which are in force since 2002. The level covered by the research is the lower secondary school – gymnasium (ISCED2). In Poland, education in a gymnasium begins after six years of primary school. Education at this level takes three years and is not specialized, it means that only general education is conducted. Gymnasium is a compulsory school. After gra-

duation at this level, students must select a school at the next level. Schools at the upper secondary level are specialized.

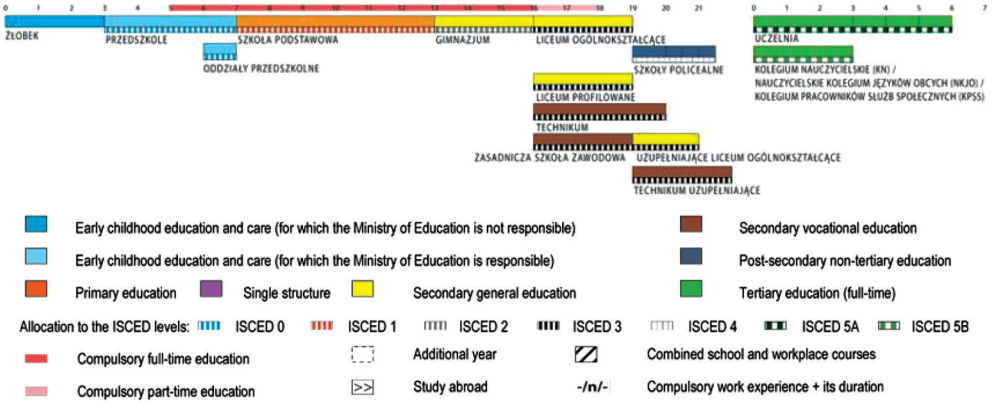


**Fig. 1.** Educational system in Poland

Source: Eurydice.

The gymnasium exam is an external exam with standardized procedures, so a nationwide comparability of results is possible. The results of the exam are not relevant for school graduation, but sometimes are relevant for further enrolment in the next level of the educational system.

The gymnasium exam is a cross-subject exam consisting of two parts: humanities (Polish language, history, civic education, arts) and science (mathematics, physics, astronomy, chemistry, biology, geography). The first part of the exam tests abilities and knowledge in two areas: reading and interpretation of texts and creation of own text. Each is scored with 25 points. The second part – science – consists of four areas: finding and using information (15 points), application of terms and procedures (12), identifying and describing facts, relationships and dependences (15), application of integrated knowledge and skills to solve problems (8). Since 2011, students also take a third part of the exam – foreign languages. Due to the scope of analysis, the overall results of students obtained from the first and second part of the exam will be taken into account. The sum of points obtained by the student is then presented on the stanine-scale (standard nine-scale): R1 – 1st stanine (the lowest result), R2 – 2nd stanine (very low), R3 – low result, R4 – low medium result, R5 – medium result, R6 – high medium result, R7 – high result, R8 – very high result, R9 – the highest result (a description of the structure of the stanine scale can be found on the web site of the Regional Examination Board includes both the analyzed regions [*Skala staninowa...* 2013].

The analyses carried out included the examination results of students in the Lower Silesia and Opolskie Region. The following variables were analyzed:

– Results of the exam: R1, R2, …, R9;
– Gender: G – girls, B – boys;
– Region: LS – Lower Silesia, OP – Opolskie Region;
– Type of Communes: UC – urban, URC – urban-rural, RC – rural;
– Big Cities: W – Wrocław, L – Legnica, JG – Jelenia Góra, O – Opole.
– Parts of the Exam: HUM – humanities, SC – science.

## 5. Analysis of gymnasium exam results

An analysis of the gymnasium exam results of 2003 and 2010 can be performed separately for each year, using the correspondence analysis twice. Then the evaluation of the location of points on both graphs should be made separately for a characterization of the results of boys and girls in both years. The full space of relationship is 8-dimensional, and the quality of the presentation in the two--dimensional space is nearly 90% of total inertia.

The most popular solution in the CA is the Burt matrix. In this analysis it is possible to study simultaneously the associations between categories of all the variables describing the population. But the research problem will not be fully realized in this way of analysis, because we get points on the graph showing the years and not the points showing the level of the results obtained by pupils in every year. In the graphical presentation of points illustrating the categories of the rows or columns, we do not get the detailed characteristics of the students. Furthermore, the full space of the relationship (with only two years) is 14-dimensional, and the quality of the presentation in the low-dimensional space may be very low.

A more appropriate table to analyze the exam results is a multiway contingency table.

In the columns there are written the categories of the exam results (R1-R9). In the rows there are layers of three variables: years, gender, and type of community (both of the Lower Silesia and Opolskie Regions). For example, G-U-03-LS means the results of girls (G) from Lower Silesian (LS) urban communes (U) in 2003 (03).

The full dimensional space of so constructed the multiway contingency table is $R^8$. A projection of variables associations on two-dimensional space describes 77.8 +12% = 89% of the total inertia. Furthermore the first axis shows almost 78% of the real relationships in multiway table.

In Figure 2 there can be observed changes in the position of the same category in 2003 and 2010. The location of these points relative to points showing the level of performance indicates the direction of change over time.

Figure 2 shows that the greatest differences in performance exist between boys and girls from big cities and regional capitals. In other communities the results of the analyzed years have similar characteristics:
– girls residing in urban, urban-rural and rural areas in both years and both regions reached the average and above average results of the gymnasium exam;

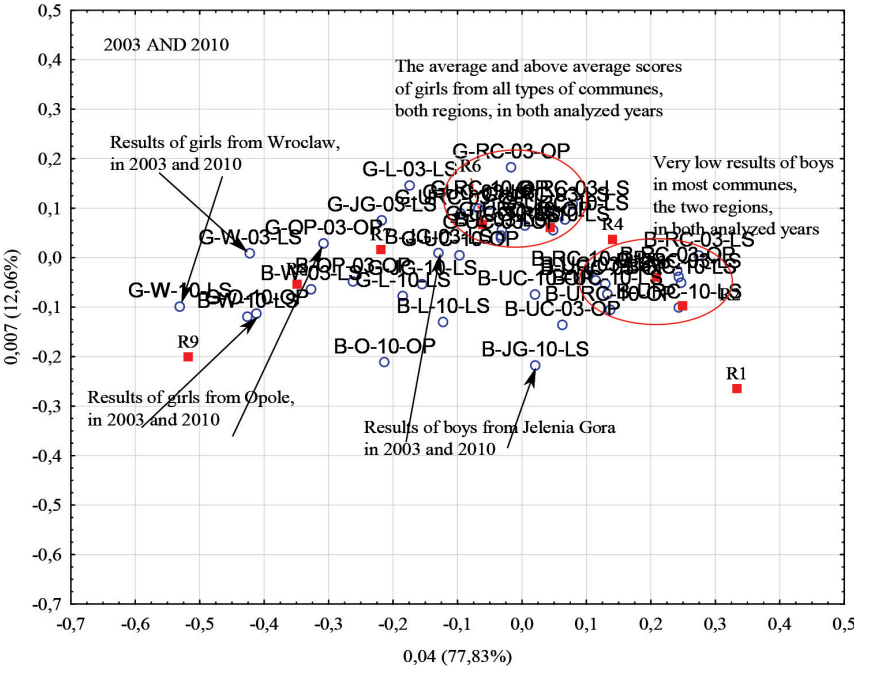


**Fig. 2.** Results the gymnasium exam in 2003 and 2010 for girls and boys from big cities and all communities

Source: own calculations.

– boys from the same (as above) communes obtained the worse results of the exams.

Figure 2 shows worse results of the gymnasium exam for boys from Jelenia Góra in 2010 than in 2003. In 2003 their results were above average and high, and in 2010 – low medium.

In the exam scores of girls from Wrocław, it can be noted that in 2010 they are characterized by the highest results and in 2003 by very high results. Girls from other big cities and all communes never reached such high results.

Similarly, very good results are also characteristic for girls from Opole in 2010 and 2003.

A deterioration of results in 2010 compared to 2003 was for boys from Jelenia Góra and Opole and for girls from urban-rural communes in the Opolskie Region. The situation observed for boys from all types of communes in both regions is stable, but bad – they obtained very low to medium low results.

If the research was not aimed at comparing the performances of the two regions, it could be used to analyze another variant of the correspondence analysis – an analysis with additional points. For this purpose, the first step of the correspondence

analysis was for a multidimensional contingency table containing the results of 2003 with layers in rows (gender/region). In the next step, based on the results for active points, there were calculated coordinates for supplementary points of the same variables but for 2010.

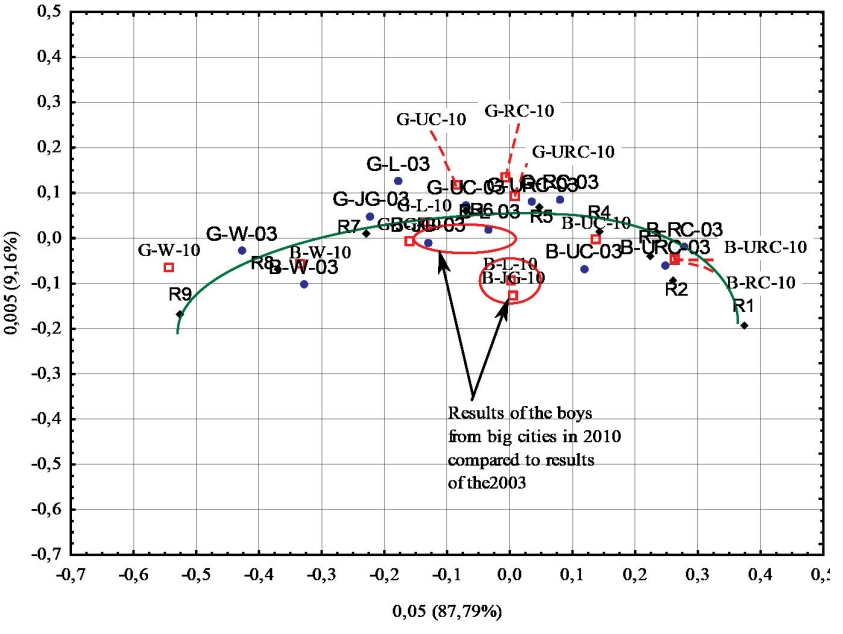


**Fig. 3.** Results of the gymnasium exam in 2003 and 2010 for girls and boys from the Lower Silesia Region (supplementary points marked red dot)

Source: own calculations.

The results of the analysis presented in Figure 3 represent 96% of the relationship between the categories of all variables. The interpretation of the position of the points for Lower Silesia is compatible with the interpretation of the results presented in Figure 3[2].

Selecting the number of layers in a multiway contingency table also allows a more specific interpretation.

The secondary school exam results can be analyzed not only due to the resulting total score of the exam but also because of the results achieved for each part of the exam. However, in this case the number of levels increases significantly resulting in a very large number of points depicting the categories in biplot. This solution is then difficult to interpret.

[2] The method of analysis depends on the researcher and on the technical possibilities of software used in the study.

For example, when trying to analyze the results of girls and boys from both parts of the exam from different cities and communities, two solutions may be considered:

– multidimensional contingency table: layers are both in rows (gender/region (city + communities)) as well as in columns (part of exam/results), changes of results during the time would be the next layer added in rows;
– analysis of additional points: a multiway contingency table is constructed of the observed frequencies of active variables. Additional points for the next analyzed year are plotted on the grounds of results of active points analysis.

The second procedure is more complicated. For this reason, Figure 4 shows the results obtained from a multiway contingency table with a triple layer in rows and a double layer in columns.

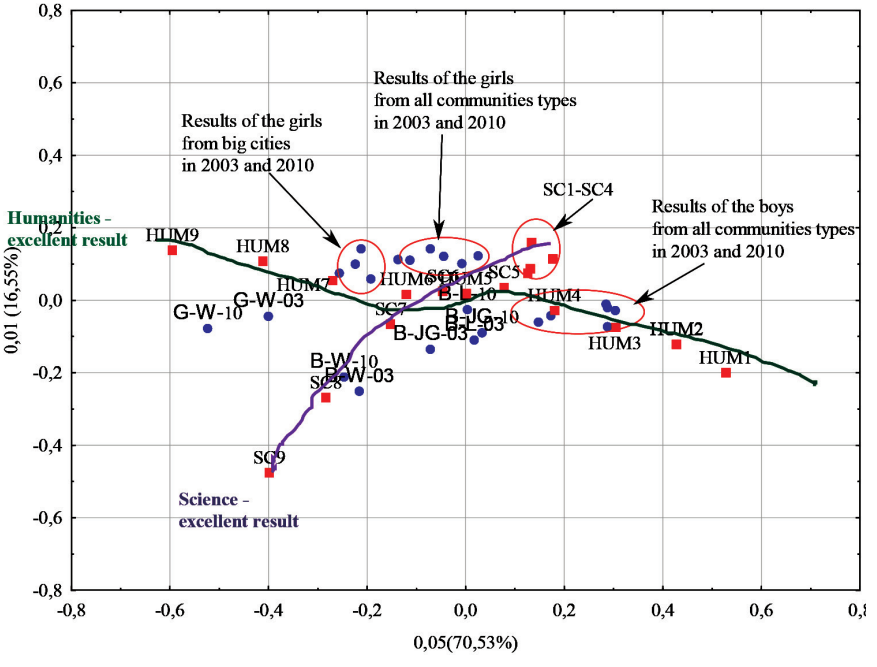


**Fig. 4.** Results of two parts of the gymnasium exam in 2003 and 2010 for girls and boys from Lower Silesia (multiway contingency table with layers in rows and in columns)

Source: own calculations.

Although the full dimensional space of so constructed multiway contingency table is $R^{17}$, a projection of variables associations on two-dimensional space describes 70.5 +16,6% = 87,1% of the total inertia. The first axis shows almost 71% of the real relationships in a multiway table.

In Figure 4, the location of the points for girls from Wroclaw is interesting. In both the analyzed periods they have achieved the highest scores in the humanities.

After the screening, on first axis it can be noted that for them there are also characteristic the highest scores for science.

The results for boys from Wrocław are very high in 2003 and 2010 in science and high in the humanities.

The results for girls from other big cities of Lower Silesia are in 2003 and 2010 high in the humanities, and the results for boys from big cities are medium in science.

Girls from all communes had medium results in the humanities in both 2003 and 2010.

Also this analysis showed that the worst results are achieved by boys in all types of communes in both analysed periods.

## 6. Final remarks

In order to summarize the conducted research, it is necessary to focus on aspects of the results obtained in different techniques of correspondence analysis and on the analysis of changes over time of the gymnasium results.

In the studies using correspondence analysis to find the relationship of several variables, the most commonly used approach is based on the Burt matrix. The use of this matrix in the analysis of time series data does not provide a full interpretation of relationships and in a limited way describes the changes of the analyzed phenomena. The use of a multiway contingency table gives better results with a higher quality of presentation. The disadvantage of this approach is the introduction of layers in the analysis and thus the recognition and interpretation of the associations of category is more difficult. The correspondence analysis with supplementary points, gives the possibility of comparing categories of passive variables to reference data. The correct construction of a matrix with active variables makes it possible to maintain the high quality of low-dimensional presentations of relationships of variables.

In the analysis of the results of the gymnasium exam conducted for the two years 2003 and 2010, it was reported that there are very large differences between the results of students from regional capitals (highest score), big cities (average scores) and communes (lower scores). Such differences were maintained over time – the results are similar for 2003 and 2010. The results of the analysis indicate that girls obtain better results than boys in all the regions and in both years.

## Literature

Andersen E.B., *The Statistical Analysis of Categorical Data*, Springer-Verlag, Berlin 1991.

Backhaus K., Erichson B., Plinke W., Weiber R., *Multivariate Analysemethoden*, Springer-Verlag, Berlin 2003.

Blasius J., *Korrespondenzanalyse*, R. Oldenbourg Verlag, München 2001.

Clausen S.E., *Applied Correspondence Analysis. An Introduction*, University Paper 121, Sage 1998.

Dore J.C., Ojasoo T., *How to analyze publication time trends by correspondence factor analysis: Analysis of publications by 48 countries in 18 disciplines over 12 years*, "Journal of The American Society for Information Science and Technology" 2001, no. 52(9), pp. 763-769.

Eurydice: *Educational System in Poland*, https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Poland:Overview (accessed 10.07.2013).

Greenacre M., *Correspondence Analysis in Practice*, Academic Press, London 1993.

Greenacre M., *Theory and Application of Correspondence Analysis*, Academic Press, London 1984.

Greenacre M., *From Simple to Multiple Correspondence Analysis*, [in:] *Multiple Correspondence Analysis and Related Methods*, ed. M. Greenacre, J. Blasius, Chapman&Hall/CRC, London 2006, pp. 41-76.

Jobson J.D., *Applied Multivariate Data Analysis*. *Vol. II: Categorical and Multivariate Methods*, Springer-Verlag, New York 1992.

Muller-Schneider T., *The Visualisation of Structural Change by Means of Correspondence Analysis*, [in:] *Correspondence Analysis in the Social Sciences*, ed. M. Greenacre, J. Blasius, Academic Press, 1994, pp. 267-280.

*Skala staninowa* ,http://www.oke.wroc.pl/images/library/File/pdfy/SP_Staniny_2011.pdf (accessed 10.07.2013).

Stanimir A., *Analiza korespondencji jako narzędzie do badania zjawisk ekonomicznych*, Wydawnictwo Akademii Ekonomicznej, Wrocław 2005.

Thiessen V., Rohlinger H., Blasius J., *The "Significance" of Minor Changes in Panel Data: a Correspondence Analysis of the Division of Household Tasks*, [in:] *Correspondence Analysis in the Social Sciences*, ed. M. Greenacre, J. Blasius, Academic Press, 1994, pp. 267-280.

Van der Heijden P.G.M., Mooijaart A., Takane Y., *Correspondence Analysis and Contingency Table Models*, [in:] *Correspondence Analysis in the Social Sciences*, ed. M. Greenacre, J. Blasius, Academic Press, 1994, pp. 79-111.

Van der Heijden P.G.M., Teunissen J., Van Orlé C., *Multiple correspondence analysis as a tool for quantification or classification of career data*, "Journal of Educational and Behavioral Statistics", December 21, 1997, no. 22, pp. 447-477.

## ANALIZA KORESPONDENCJI ZMIAN W CZASIE WYNIKÓW EGZAMINU GIMNAZJALNEGO

**Streszczenie:** Celem artykułu jest prezentacja możliwości wykorzystania analizy korespondencji w badaniu zmian w czasie wyników egzaminu gimnazjalnego w Polsce. W badaniu wykorzystano dane z 2003 i 2010 roku z Okręgowej Komisji Egzaminacyjnej obejmującej dwa województwa: dolnośląskie i opolskie. Zaprezentowano sposób wykorzystania analizy korespondencji z wykorzystaniem wielowymiarowej tablicy kontyngencji oraz wprowadzając punkty dodatkowe. Wyniki uzyskane po zastosowaniu analizy korespondencji wskazują na istnienie różnic między poziomem wyników z egzaminu gimnazjalnego w zależności od roku badania, płci oraz miejsca zdawania egzaminu.

**Słowa kluczowe:** analiza korespondencji, wielowymiarowa tablica kontyngencji, punkty dodatkowe, egzamin gimnazjalny.