Control and Cybernetics

vol. **38** (2009) No. 3

Grade analysis of data from the European Economic Survey 2005 on Economic Climate in Polish Servicing Sector*†

by

Grażyna Grabowska[1] **and Marek Wiech**[2]

[1]Systems Research Institute, Polish Academy of Science
Newelska 6, 01-447 Warszawa, Poland
[2]Institute of Computer Science, Polish Academy of Sciences,
J.K. Ordona 21, 01-237 Warszawa, Poland

Abstract: The anonymous data from 1352 companies concerning the economic climate in Polish servicing sector from the European Economic Survey 2005 was obtained by courtesy of The Polish Chamber of Commerce. The Grade Correspondence Analysis (GCA) with posterior clustering (GCCA) is introduced and applied to this data. The main task of this analysis is to create the first view of data and to reveal their latent structure. This provides an insight into the economic factors and enables making conclusions about business conditions in Poland.

Keywords: clustering, data visualization, grade correspondence analysis, overrepresentation, rank correlation.

1. Introduction

Knowledge about the main economic indicators and parameters derived from business data enables forecasting future trends – in optimistic, pessimistic or neutral terms – in the country. In this article we aim to present two objectives: (a) the results of analysis of economic factors in Poland in servicing sector (b) using the algorithms of Grade Data Exploration, mainly Grade Correspondence Analysis and related clustering. The analyzed data are derived from the European Economic Survey 2005 (EES'2005) and concern the economic climate in Poland in 2004 and the forecast for 2005. The carefully chosen data, introduced in Section 1.2, were given to us by the Polish Chamber of Commerce, as we were not allowed to receive or buy comparable raw data from the Central Statistical Office; however, as we aim to focus on methodology rather than a particular

*This study was partially sponsored from the grant number 3 T11C 053 28, awarded by the Ministry of Education and Science of Poland.

†Submitted: January 2008; Accepted: January 2009.

application, we hope that clarity of the chosen example will help in grasping intuition of grade data exploration.

In our analysis we wanted to examine the interdependence between economic factors and describe business climate in Poland in 2004 and in 2005 using grade data analysis. Native grade data visualizations: overrepresentation maps of raw and ordered data, aggregated clusters of rows and rank correlation tables are given to interpret the results and support their understanding. Grade methodology itself is introduced in Section 1.1 and explained in greater detail in Section 2. Section 3 presents in short particular stages of the analysis: data preparation, visualization and interpretation of results, and also psychometric remarks on the EES Survey, which may be helpful in deeper understanding of the results. The last Section 4 concludes the analyzes and proposes how grade data analysis supplements traditional methods

All analyses and figures were made in program GradeStat. Further information concerning GradeStat, grade infrastructure and grade methods is available at http://gradestat.ipipan.waw.pl.

Brief summary of Grade Data Analysis

The main idea behind analyses of economic reports as those published at the website of Polish Central Statistical Office (http://www.stat.gov.pl) or of the National Bank of Poland (http://www.nbp.pl) is to examine economic situation by casting overall prognosis based on summary indices, which are calculated separately for particular issues. In such an approach particular firms are usually classified according to their trade, while in grade exploration the main aim is to cluster firms, accounting for the profile of *all* their answers (not only to combine them by trade and/or size). In the grade approach, raw data with full answers of each firm are needed, and the method may give us additional information, as we will see later on.

Grade data analysis is efficient on variables measured on any measurement scales (even categorical), because it bases on dissimilarity measures such as concentration curves and some precisely defined measure of monotonic dependence. Its main framework is grade transformation (proposed in Szczesny, 1991). The idea is to transform any distribution of two variables into a convenient form of the so called grade distribution. This transformation leaves unchanged the order of variables, ranks, values of monotone dependence measures like Spearman's ρ^* and Kendall's τ. In case of empirical data this approach consists in analyzing the two-way table of objects/variables, preceded by proper recoding of variable values. After grade transformation we can treat contingency table as a probability table (the benefits of probability tables are explained in Section 2.1).

The main tool of grade methods is Grade Correspondence Analysis (GCA), referring to classical correspondence analysis, but going significantly beyond it, owing to grade transformation. To put it simply, GCA orders the vari-

ables/objects table in such a way that adjacent objects are more similar than those further apart, and at the same time, adjacent variables are also more similar than those further apart. When optimal ordering is found it is possible to aggregate adjacent objects and adjacent variables, and therefore to build a definite number of clusters, consisting of objects (or variables) with similar distributions.

Last but not least, it is possible to indicate main trend in data, find objects (or variables) highly departing this trend and remove these objects from further analyses. This feature distinguishes grade data analysis from classical analyses, where outlying objects are excluded, but are not then analyzed as a sub-table. In grade methods outliers are removed and analyses are applied again to the more regular sub-table and to the sub-table with outliers.

Decisions on what level of departure is proper, how many cluster should be determined or what the detected trend shows are aided by visualizations, mostly overrepresentation and correlation maps. Overrepresentation map is the chart of the probability density of grade distribution, showing which cells are over- or underrepresented in a particular dataset, while correlation maps show linear dependences between variables.

While typical analyses focus on relating the results to larger population, grade data analysis puts the stress on trends found in gathered data. The conclusions are drawn from data and give useful information even in case of bad randomization or other problems common in analyzes of real data. It should be noted at this point that grade data analysis is an exploratory method (Greenacre, 1984) based on the development of model (or models) that fits the data, rather than on rejection of hypotheses due to the lack of fit. Therefore, there are no statistical significance tests applied to the results, as the main purpose of this method is to get "deeper insight" into analyzed data.

2. Grade Data Analysis - methodology

2.1. Grade transformation and overrepresentation maps

The main tool of grade methods is Grade Correspondence Analysis (GCA), an algorithm to order variables/objects matrix in such a way that adjacent objects are relatively more similar than those further apart, and at the same time, adjacent variables are also relatively more similar than those further apart. It is then possible to cluster adjacent objects (variables) or remove objects (variables) highly departing from the detected trend. However, the original table must be in some way normalized before ordering and one of possible methods is grade transformation.

Let us now remind that *grade of* x is a term dating from the very beginning of statistics, meaning value of *CDF* F_X (cumulative distribution function) given at point x, so that "grade of x" is equal to $F_X(x)$, where $F_X(x) = P(X \leq x)$. It describes the position of value x of variable X in the interval $[0, 1]$ against

the background of the whole probability distribution of this variable. Therefore, all methods associated with transformation of one variable by a suitably chosen *CDF* of another variable are called *grade methods*. In statistical literature the notion of *grade* is connected with Spearman *rho* index, which measures the strength of monotone dependence of a pair of random variables (X, Y). For continuous variables X and Y this index equals correlation coefficient $corr(F_X(X), F_Y(Y))$ of X and Y variables transformed by their *CDFs*, and is called *grade correlation*.

The core of analysis and interpretation of objects/variables table is comparison of object with object (record with record) and of the values of one variable values with the values of another variable. In both cases sequences of variable values are compared. A measure of differentiation between two ordered sequences is needed to measure the strength of dependence between them and departure of this pair from regularity. Such a differentiation allows for building a model in the set of objects/variables tables, referring to the model of monotone dependence of bivariate probability distributions. The framework of this model was introduced in Kowalczyk et al. (2004). This model is called *mixture of bivariate distributions with regular monotone dependence* and research concentrates on model identification and its robustness.

As mentioned, in order to introduce this model it is necessary to transform the objects/variables table so that it can be treated as a probability table of bivariate distributions. The distribution itself is just a model for the objects/variables table, or a pair (Objects, Variables), or, more universally, of (Rows, Columns), the respective terminology being appropriate for the sets of m objects and k columns. To generalize this model to an infinite distribution we can change the terminology to (Vertical, Horizontal) but we can also stay with (Rows, Columns), meaning countless number of rows and columns.

What are the conditions for treating (Rows, Columns) table as a probability table of a bivariate distribution? First of all, cells in the table must be nonnegative; when we divide each cell by the total sum of all cells we obtain the table which formally is a probability table of a certain distribution. The subsequent condition asks whether such addition and division are reasonable. It is very difficult to fulfill this condition, perhaps it would be possible to consider when such operations are allowed, basing on *measurement* theory where concept of *meaningfulness* exists. Formally, a parameter of an image of some relational structure is meaningful when it has a prototype in this structure and can be obtained from this prototype with the help of *admissible* measurement scale. To build meaningful probability table from the table objects/variables, it is necessary to describe the set of such tables with accompanying relations where prototypes of adding and dividing have sense.

The family of objects/variables table which without reservation can be transformed into a family of probability tables is the family of contingency tables obtained from an $N \times 2$ table where N objects are described by two nominal or ordinal variables with, respectively, m and k values. For example in

election with N eligible voters we can have two variables: "party chosen" and "province of voter". This results are *unobservable* in the case of first variable, but a "province/party" table can be obtained from data available for the electoral commission, in which each cell is the number of voters in this province for this party. Here it is completely reasonable to add values in cells *horizontally* as well as *vertically*, and also sums of column, sums of row and total sum are meaningful.

The simplest grade correlation consists in transformation of a single interval or categorical variable. *CDF* is a basic term describing distribution of both variable and random variable. For interval random variable its transformation of X by F_X, i.e. variable $F_X(X)$, is obviously uniform.

For a categorical variable with values $x_1...x_k$ its transformation of X by F_X are points $p_1, p_1+p_2, ..., 1$ in the unit interval. To build from it a uniformly distributed random variable it is necessary to uniformly "blur" the probability p_i for every $i...k$ in the interval from $(p_1+...+p_{i-1})$ to $(p_1+...+p_i)$; in probability theory such a transformation is performed by suitably defined probability transformation function. The length of this interval is exactly equal p_i. This simplest grade transformation is used, e.g., to examine a pair of *categorical* variables with respectively m and k categories, when we transform them by "blurred" *CDF*. Let the probability table p_{ij} describe the probability of occurrence of an object of the ith category for the first variable and of the jth category for the second variable, $i = 1^m, j = 1^k$, then the resulting grade transformation will be probability density:

$$h(u, v) = \frac{p_{ij}}{p_{i+}p_{+j}}$$

where $p_{i+} = p_{i1} + ... + p_{ik}$ and in a similar way $p_{+j} = p_{1j} + ... + p_{mj}$. Density h attains its values almost everywhere in a unit square and is constant on rectangles, therefore it is possible to draw its chart (Fig. 1).

	Y1	Y2	Y3
X1	0.081	0.049	0.07
X2	0.031	0.109	0.061
X3	0.109	0.031	0.061
X4	0.064	0.064	0.072
X5	0.049	0.081	0.07

	Y1	Y2	Y3
X1	1.215	0.735	1.05
X2	0.4626	1.6268	0.9104
X3	1.6268	0.4626	0.9104
X4	0.96	0.96	1.08
X5	0.735	1.215	1.05

1.55
1.45
1.34
1.24
1.13
1.03
0.92
0.82
0.71
0.61
0.47

Figure 1. Table T1 with values of exemplary data (left) and map of densities $h(T1)$ for this data with color of the cell background corresponding to the value of density (right).

Density h is constant and equal 1 for the whole unit square when variables X and Y are independent, so its chart – called overrepresentation map – is uniformly grey for all rectangles (Fig. 2). The rectangles are filled with color corresponding to the value of overrepresentation depicted by each rectangle: grey for value equal 1, approaching white when overrepresentation value is less than 1 and approaching black when overrepresentation value is higher than 1.

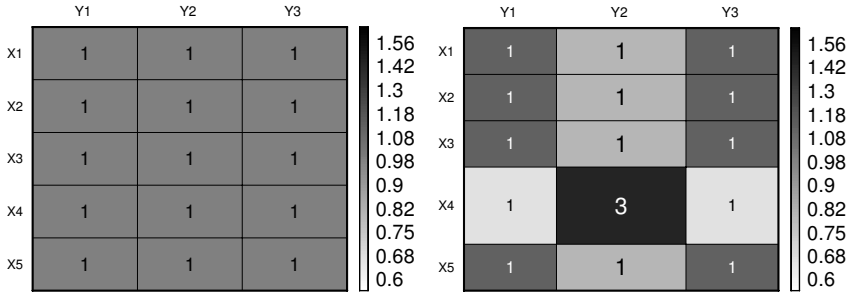

Figure 2. Overrepresentation maps (map of densities): table with constant density (left); similar table but the density of cell (4,2) is higher than of the rest (right).

The width of columns corresponds to the share of each column in the whole table, the same applies to the width of rows. Respectively the size of p_{ij} rectangle corresponds to the share of this cell in the whole table. Overrepresentation maps help in taking an overview on distributions of rows and columns. It is also possible to draw lines showing the centres of mass of columns and of rows, as shown in the next section.

Basic grade concepts may seem elderly, and indeed they are. The foundations of grade exploration are over one century old. The most important concepts were created by Gini and Lorenz at the beginning of 20th century. However, Lorenz curve and Gini index continue to be the basis for numerous statistical analyses, and for considerations oriented at decision making.

Originally, Lorenz curve described differentiation between two variables, with the first consisting of earnings of selected group of people and the second containing number of people in each group. The data made table with two columns $[(x_i, n_i), i = 1, ..., m]$, where m is the number of groups. Both variables were then normalized by creating vectors:

$$q = (q_i,\ i = 1, \ldots, m) = \left(\frac{x_i}{\sum_{j=1}^{m} x_j},\ i = 1, \ldots, m \right),$$

$$p = (p_i,\ i = 1, \ldots, m) = \left(\frac{n_i}{\sum_{j=1}^{m} n_j}, i = 1,\ \ldots, m \right).$$

Vectors q and p are called probability vectors, because their components have values in interval $[0, 1]$ and sum to 1. With every selected ordering of groups it is possible to sum cumulatively components of every vector and then compare them both and therefore create the sequence of points in unit square: $\{(S_i = \sum_{j=1}^{i} p_j, T_i = \sum_{j=1}^{i} q_j); i = 1, \ldots, m\}$.

If the sequence is preceded by the point (0,0) and consecutive points are connected by lines, then broken line in the unit square might be drawn connecting points (0,0) and (1,1). If everybody's earnings are equal, then vectors p and q are equal, and curve lies on the diagonal connecting points (0,0) and (1,1); otherwise the curve lies under or above the diagonal, or crosses it.

There are as many different curves as there are possible orderings of groups. Lorenz curve bases on the ordering when quotient q_i/p_i is non-increasing, therefore curve is convex and lies under the diagonal (or on the diagonal when vectors p_i and q_i are equal), as seen in Fig. 3. The curves based on any other ordering are drawn between the Lorenz curve and its counterpart drawn symmetrically on the other side of diagonal, called upper Lorenz curve and based on the ordering of groups exactly opposite to the one for Lorenz curve.

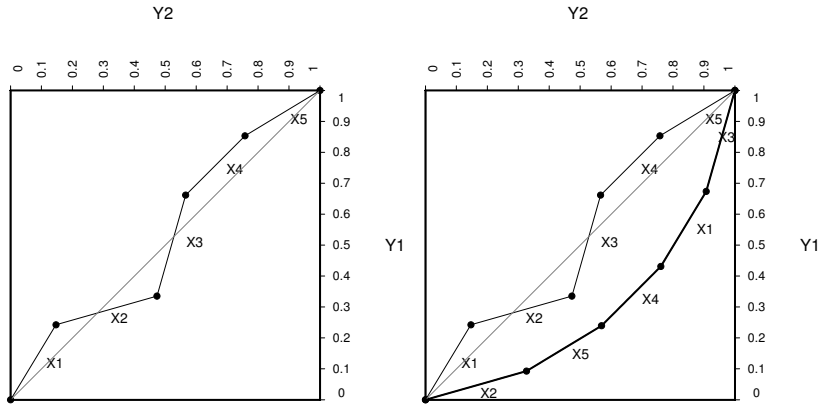

Figure 3. Concentration curves for variables from the example $T1$: left – $C(Y1 : Y2)$, right – with added maximal concentration curve $C_{\max}(Y1 : Y2)$.

Area between diagonal and a curve is expressed by an integral: $\int_0^1 (x - C(x))dx$. The integral is equal to the difference between $1/2$ (average value of diagonal) and average value of curve $C(x)$. In case of two extreme possibilities curve $C(x)$ lies on the side of square: bottom horizontal/vertical right (and its average is equal 0), or vertical left / upper horizontal (and its average is equal 1). So, the area between diagonal and curve has values in interval $[-1/2, 1/2]$. To get a conventionally normalized index with values in the interval $[0, 1]$ the difference of diagonal average and curve average is doubled, and the normalized index of curve C is denoted $ar : ar(C) = 2\int_0^1 (x - C(x))dx$.

Index *ar* is equal to twice the absolute difference between the diagonal and the curve. Index *ar* calculated for Lorenz curve is called *Gini index*. In bivariate model the plot of *CDF* is called concentration surface (in analogy to concentration curve). Grade model is thoroughly described in Kowalczyk et al. (2004) and Książyk et al. (2005), where it is shown that for the convex curve index *ar* is maximal and called $ar_{\max}$. Index *ar* is a very important concept used in other parts of grade data exploration (especially in outlier detection), hence it was introduced here.

2.2. Grade Correspondence Analysis (GCA)

GCA is the main tool used in grade data exploration. The purpose of the algorithm is to reorder the rows and columns of a table to maximize a certain measure of dependence between variables and objects, namely the Spearman's Rho (ρ^*). It proceeds by alternating permutations of rows and columns. The rule for choosing the next permutation guarantees that ρ^* is increased at each step, as shown in Ciok et al. (1995). The algorithm stops when the rule cannot produce further improvements. GCA is a Monte Carlo method, and so the termination of the process does not mean that the largest possible ρ^* has been reached, but practice proves that repeating the algorithm over $10 \sim 100$ times produces orderings with values of ρ^* very close to the highest possible, as shown in Matyja (2002). When the table is small it is also possible to check all permutations or rows and columns and find the maximum ρ^*.

Spearman ρ^* originally was defined for continuous distributions, however it may be defined also as Pearson's correlation applied to distribution after the grade transformation. The grade distribution may be defined for discrete distribution too, and it is possible to calculate Spearman ρ^* for probability table P with m rows and k columns, where p_{is} is the frequency ("probability") of ith row in sth column:

$$\rho^*(P) = 3\sum_{i=1}^{m}\sum_{s=1}^{k}(p_{is}(2S_{row}(i)-1)(2S_{col}(s)-1))$$

where:

$$S_{row}(i) = \left(\sum_{j=1}^{i-1} p_{j+}\right) + \frac{1}{2}p_{i+}$$

$$S_{col}(s) = \left(\sum_{t=1}^{s-1} p_{+t}\right) + \frac{1}{2}p_{+s}$$

and p_{j+} and p_{+t} are marginal sums defined as:

$$p_{j+} = \sum_{s=1}^{k} p_{js}, \quad p_{+t} = \sum_{t=1}^{m} p_{ts}.$$

GCA tends to maximize ρ^* by ordering rows and columns according to their so called grade regression, which is "the centre of mass" of each row or of each column. Grade regression for columns is defined as:

$$Regr_{col}(s) = \frac{\sum_{i=1}^{m}(p_{is}\, S_{row}(i))}{p_{+s}}$$

and for rows as:

$$Regr_{row}(i) = \frac{\sum_{s=1}^{k}(p_{is}\, S_{col}(s))}{p_{i+}}.$$

Grade regressions calculated for previously presented example are shown as horizontal and vertical lines in Fig. 4. This figure illustrates also how GCA works. If we calculate the grade regression for columns and sort the columns by its values the regression for columns will increase, but regression for rows will change. Then, if we sorted regression for rows, regression for columns changes. Still, as proved in Ciok et al. (1995) each sorting of the grade regression increases the value of Spearman ρ^*. The number of possible states (combination of permutations of rows and columns) is finite and equal $k!m!$ and so the GCA must stop.

Each time the value of Spearman ρ^* increases, and the last ordering produces the largest ρ^*, called local maximum of Spearman ρ^*.

The output from GCA depends on the initial permutation of rows and columns, and if we order the reversed initial permutation, we achieve symmetrically reversed local maximum, therefore by local maximum we mean a pair of ρ^*: original and its reversal.

For more irregular examples than the one shown here it is necessary to try many initial permutations and choose the result with the highest $\rho*$. GCA at first randomly permutes rows and columns and reorders them to achieve a local maximum. This process is iterated as many times as needed (typically 100 iterations) and the result with the highest ρ^* is chosen. If we checked all possible start permutation the result would be the global maximum of ρ^* (the largest possible in the table).

It is important to state that calculation of grade regression requires non-zero sum of every row and column in a table, so this requirement applies also to the GCA. (More information on GCA is available in Kowalczyk, Pleszczyńska and Ruland, 2004, and in Matyja, 2002.)

2.3. Grade Correspondence Cluster Analysis (GCCA)

Cluster analysis aims at discovering structures in data, but without explaining or interpreting why these structures exist. Typically it sorts different objects into groups so that similarity between two objects is maximal if they belong to the same group and minimal otherwise. Grade Correspondence Cluster Analysis

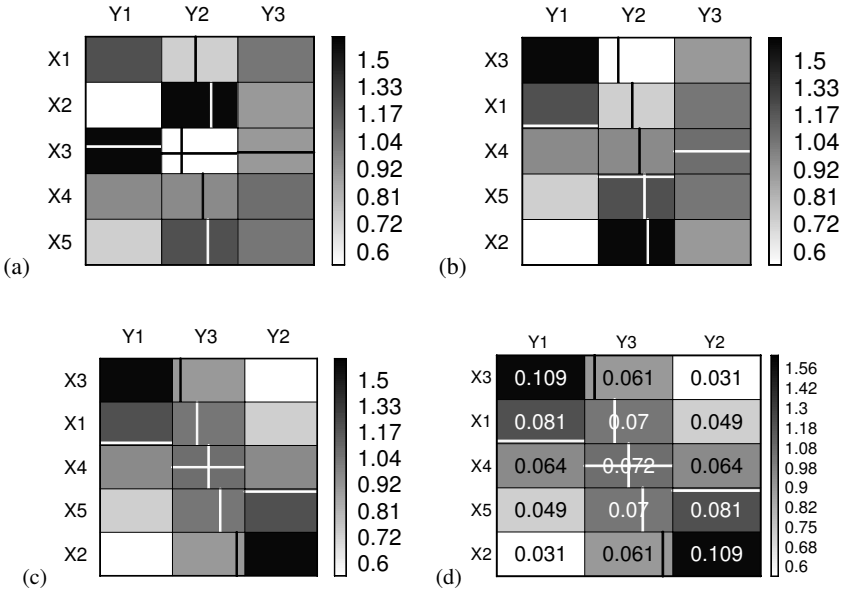

Figure 4. Overrepresentation maps showing step by step how GCA works: (a) map before ordering, horizontal lines are grade regression for columns, vertical for rows; (b) rows are ordered; (c) columns are ordered; (d) final ordering with cell values shown.

is no different – it bases on optimal permutations of data ordered earlier by GCA to aggregate the most similar rows or similar columns together. Resulting clusters can be only non-overlapping and the number of required clusters must be specified by an analyst. Grade clustering algorithm splits a series or rows into specified number of clusters trying to maximize both strength and regularity of dependence of a new table, which would be obtained by aggregation inside clusters. The respective aggregated probabilities in this table arose from sums of component probabilities in initial, optimally ordered table, and number of rows in the aggregated table equals the desired number of clusters.

Exactly the same procedure as with rows goes with columns, as they can be separately clustered. However, rows and columns in the initial table may be aggregated simultaneously, therefore GCCA is a two-way clustering method (as it allows to simultaneously cluster rows and columns). Clustering of only rows or of only columns is called single clustering, whereas clustering of both rows and columns is called double clustering. Comparison of both methods can be found in Ciok (2000), and because double clustering bases on the sequence of two single clusterings, cluster determination is common to both methods.

As Fig. 5 shows, GCCA forms clusters only of adjacent rows (columns) in the GCA optimal permutations (Ciok, 2004). GCA permutes the rows and columns of data table in such a way that they are ordered according to the values of respective grade regression functions. When optimal ordering is found by GCA it is possible to aggregate adjacent objects and adjacent variables (GCCA forms clusters by further discretization of grade regression functions) and therefore to set *a priori* number of clusters consisting of objects (or variables) with distributions as similar as possible.

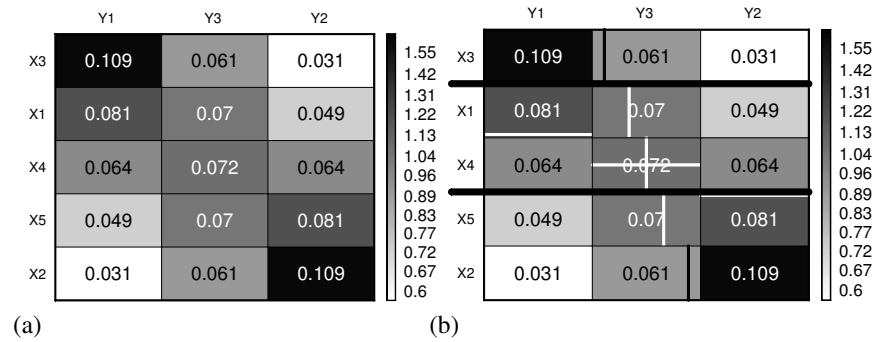

Figure 5. Overrepresentation maps: (a) ordered by GCA; (b) rows divided into 3 clusters separated one from another by two black horizontal lines.

GCCA maximizes the grade correlation coefficient of the aggregated table (therefore according to terminology used in Ciok, 2004, clustering by GCCA is of the optimization type). It can be characterized as a nonhierarchical method generating non-overlapping clusters. It slightly resembles the k-means method, because this coefficient expresses within-cluster diversity as well as differences among clusters; however, the diversity measure being used is the main difference between both methods. How optimal clusters are set is fully described in Ciok et al (1995), Kowalczyk, Pleszczyńska and Ruland (2004) and Ciok (2004).

Clustering methods may be also categorized according to what kind of input data they need. Many of them are designed for one sort of input data, but GCCA is applicable to any kind, as long as they can be transformed into probability table; however in case of values of nominal data they should be previously turned into respective separate "dummy" variables and, as in case of GCA, zero sums of rows or columns are not allowed.

Let us stress that the number of clusters is set *a priori* by the analyst and there is no obvious way on deciding on what number of clusters should be applied. In case of empirical data it depends on the goal of aggregation.

2.4. Detection and analysis of outliers in data

Last but not least, grade data analysis enables finding objects or variables highly departing from main trend in data, and to separate these objects (or variables) into different sub-tables. Then both the sub-table following main trend and the outlying sub-table are analyzed (Szczesny 1999, 2000). This feature distinguishes grade data analysis from classical analyses in supposing that outlying objects could be treated as another table, which may be governed by a different trend.

The main concept of regularity in grade methods is that when table belongs to the family of tables called TP2 (totally positively dependent of order two) it guarantees highly regular monotone trend between row variable and column variable. Such a highly regular table is, for example, a discretized binormal table. In case of empirical data, if ordering of any probability table by GCA increases its ρ^*, then this table at the same time becomes closer to TP2 table. Outlier detection consists in looking for rows (or columns) which do not follow the monotone trend introduced by GCA, and then moving these rows (or columns) to a separate sub-table called OUT (outliers sub-table). Not excluded rows form a sub-table called FIT, which has equal or higher ρ^* than the original table, and usually is more regular.

During outlier detection (performed after ordering the table with GCA) rows and columns are looked for in the same way, basing on concentration index $ar \in [-1, 1]$ and on maximal concentration index $ar_{\max} \in [0, 1]$. Index ar measures departure between two univariate distributions, while $ar_{\max}$ measures upper-bound departure between them. They are calculated for all pairs of rows and/or all pairs of columns. To obtain $ar_{\max}$ for a given pair of rows (i, j), columns are permuted until the respective ar reaches its maximum. Thus $ar(i, j) \leq ar_{\max}(i, j)$ with equality holding for all pairs (i, j) if and only if the table belongs to *TP2*. The closer a table is to a *TP2*, the more similar is the set of $ar(j : i)$ to the set of $ar_{\max}(j : i)$. Therefore, the distance of a table P from *TP2* can be evaluated from two scatterplots of points, for rows: $\{(ar(j : i, row(P)), ar_{\max}(j : i, row(P))), i = 1, \ldots, m, j = i+1, \ldots, m\}$, and for columns: $\{(ar(j : i, col(P)), ar_{\max}(j : i, col(P))), i = 1, \ldots, k, j = i + 1, \ldots, k\}$.

The measure is based on a subset of points in a scatterplot, which correspond to a particular row of table P with m rows and k columns. A subset corresponding to a row contains $m - 1$ points. Euclidean distance from the line $ar = ar_{\max}$ is calculated for every point. The expectation of all distances in a particular subset is denoted $AvgDist_{row}$:

$$AvgDist_{row}(i; P) = \sum_{s=1}^{i-1} \frac{ar_{\max}(i : s; row(P)) - ar(i : s; row(P))}{(m-1)\sqrt{2}} + \sum_{s=i+1}^{m} \frac{ar_{\max}(s : i; row(P)) - ar(s : i; row(P))}{(m-1)\sqrt{2}},\ i = 1, \ldots, m\ .$$

For columns, the respective measure is called $AvgDist_{col}$:

$$AvgDist_{col}(j;P) = \sum_{t=1}^{j} \frac{ar_{\max}(j:t;col(P)) - ar(j:t;col(P))}{(k-1)\sqrt{2}} + \sum_{t=j+1}^{j} \frac{ar_{\max}(t:j;col(P)) - ar(t:j;col(P))}{(k-1)\sqrt{2}},\ j = 1,\ldots,k$$

Rows with highest values of *AvgDist* (and therefore departing from the main trend) are excluded to sub-table OUT, while the rest of rows remain in sub-table FIT. There is no easy way to deduct how many rows should be treated as outliers and an analyst has to subjectively decide what threshold to set. Fig. 6 shows a simple example with nine objects – here rows "2161" and "1558" have the highest *AvgDist* values and will be considered outliers.

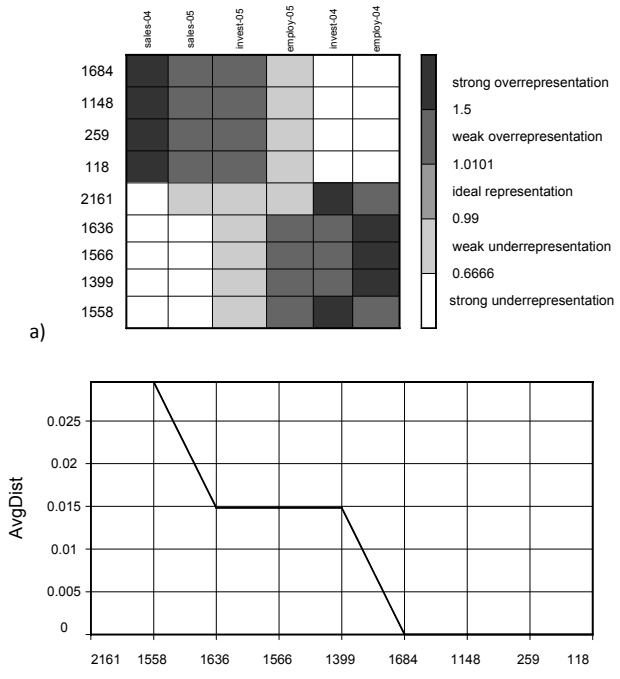

Figure 6. (a) An exemplary overrepresentation map of data ordered by GCA; (b) plot of $AvgDist_{row}$ values, the leftmost rows entitled "2161" and "1558" have the highest values and are outliers, while rows "1148", "259" and "118" do not diverge from the main trend.

2.5. Summary

A typical grade exploration analysis consists of all points mentioned above. First, we transform input data into a probability table. Then, GCA is applied to the table and optimal permutations of rows and columns are found. The pair of permutations corresponding to the highest ρ^* is then chosen. Third, GCCA is applied to ordered table to find the desired number of row (and/or column) clusters.

Afterwards outliers are detected and moved to the sub-table called OUT, while the remaining objects form the sub-table called FIT. Both sub-tables are then independently analyzed by following the procedure applied to the original table (i.e. GCA and GCCA are applied) and results are interpreted. Usually interpretation of FIT sub-table confirms main conclusions from the whole table, while interpretation of OUT sub-table brings new and important information.

In the next section we will illustrate grade exploration concepts by analyzing data from a real economic survey. Besides, it is worth to note that in years 2000-2007 grade data analysis was applied to real datasets from many fields of sciences: analysis of NMR medical images (Grzegorek, 2005, 2007); text mining and word clustering (Jarochowska and Ciesielski, 2006); assessing quality of e-learning materials (Stasiecka, Płodzień and Stemposz, 2006); questionnaire data (Pleszczyńska, 2007).

3. GCA applied to selected data from the European Economic Survey 2005

3.1. European Economic Survey 2005 (EES'2005)

In EES'2005 (*http://www.eurochambres.be*) 1352 Polish companies from the servicing sector took part. Each of them answered six *pairs* of questions about the present and future situation concerning six economic factors: *business confidence*, *total turnover*, *domestic* and *export sales*, *employment* and *investment*. Only three *pairs*: *domestic sales*, *employment* and *investment* are analyzed here, just to give a short review of the method and results. The corresponding questions are shown in Table 1.

All these questions have coded answers: for "decrease(d)" ("pessimistic") the code is "1", for "not change(d)" ("neutral") the code is "2" and for "increase(d)" ("optimistic") the code is "3". The data matrix contains values of these coded answers (columns) for particular companies (rows).

3.2. Creating data matrix and its preliminary visualization

As stated before, in the grade framework input data should be a two-dimensional table with non-negative values and positive sums of rows and columns. A part of data matrix for some companies is shown in Table 2. Six variables are taken

Table 1. Questionnaire of the European Economic Survey 2005 (6 questions out of 12).

1. Compared with 2003 revenue from *domestic sales* in 2004 has:	increased/not changed/decreased
2. We expect that revenue from *domestic sales* in 2005 will:	increase/not change/decrease
3. Compared with 2003, the size of our *workforce* in 2004 has:	increased/not changed/decreased
4. We expect that during 2005 the size of our*workforce* will:	increase/not change/decrease
5. Compared with 2003, our *level of investment* in 2004 has:	increased/not changed/decreased
6. We expect that during 2005 our *level of investment* will:	increase/ not change/decrease

Table 2. The beginning and the end of the considered part of the EES'2005 data matrix (with the preliminary ordering of companies and variables).

Company No.	**employ-04**	**employ-05**	**invest-04**	**invest-05**	**sales-04**	**sales-05**
1	3	2	1	2	2	1
2	1	3	1	1	1	3
3	3	2	1	3	3	2
4	2	1	2	2	1	1
5	2	1	2	1	2	3
:	:	:	:	:	:	:
1352	2	1	3	2	1	3

into account, corresponding to respective rows of Table 1. There are no missing data in the considered part of EES'2005 questionnaire.

After applying GCA to this preliminary data matrix we get suitable GCA permutations of rows (*companies*) and of columns (variables - *economic factors*); it means that columns and rows are simultaneously ordered to detect main trend, as illustrated in Table 3.

Table 3 shows the beginning and the end of the post-GCA data matrix, representing the opposite extreme subsets of companies. Both subsets consist of companies for which there is no balance between company's "output" (*domestic sales*) and company's "input" (*employment* and *investment*). At the beginning there are companies with optimistic actual situation and also optimistic prospects for *domestic sales*, accompanied by pessimism in *investment* and *employment* observed in 2004 (but expecting to get better in 2005). At the end, *domestic sales* 2004 and *domestic sales* 2005 are pessimistic, as contrasted with positive efforts in *investment* and *employment* realized in 2004 and expected to be realized in 2005.

Table 3. Data from questionnaire with rows and columns ordered by GCA (beginning and end of the post-GCA matrix).

GCA rank	company No.	sales -04	sales -05	invest -05	employ -05	invest -04	employ -04
1	1143	3	3	3	1	1	1
2	894	3	3	1	1	1	1
3	717	3	3	3	1	1	1
4	1304	3	3	2	2	1	1
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
1350	1200	1	1	2	3	3	3
1351	754	1	1	3	3	3	3
1352	1104	1	1	2	3	3	3

Adjacent companies have similar profiles. These profiles gradually change in the post-GCA data matrix from optimism in *domestic sales* (for 2004 and 2005) accompanied by pessimism in *employment* and *investment* in 2004, to a reverse situation, passing through intermediate stages. Similarity of adjacent rows and columns enables uniting adjacent records and adjacent variables into clusters. Thus, the completed Table 3 will show the so called main trend of monotone relationship between companies and variables.

Data from the survey are visualized by means of two *overrepresentation maps*: one with raw data ordered in the preliminary way and another one with rows and columns ordered by GCA. When we confront these two maps, we see how different orderings of rows and columns are, and that the cells are scattered more randomly in the unordered map (Fig. 7).

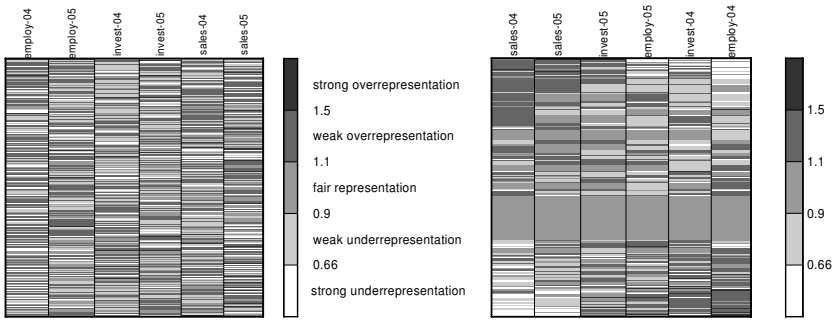

Figure 7. Overrepresentation maps for raw data (left) and data ordered by GCA (right). Note different orderings of columns in both maps. Orderings of rows are also different, what can be seen in the related data matrices.

The map with raw data is rather unreadable, whereas the map sorted by GCA gives us a more clear view, arranging together similar rows and columns. As introduced in Section 2.1, overrepresentation map consists of rectangles, colored in one of five grades ranging from white to black, that picture the overrepresentation or underrepresentation of the answer given by a particular *company* to a particular *economic factor*. The interpretation is as follows: the color for the i-th company and j-th economic factor is white or light gray when the factor's value is relatively small (*strong* or *weak underrepresentation*), middle-gray in the intermediate cases (*fair representation*), and dark gray or black when it is relatively large (*strong* or *weak overrepresentation*). "Fair representation" takes place when the coded answer in particular cell is equal (or almost equal) to the product of the means of this particular column and this particular row and finally divided by the mean of the whole matrix. Overrepresentation or underrepresentation in each cell is established with respect to the fair representation for this cell. The related scale is shown in Fig. 7 at the right-hand side of each map.

The widths of columns show the contribution of particular *variable* (*economic factor* in 2004 or in 2005) in the whole data matrix. When the column is wide, the contribution of *variable* is big (i.e. generally optimistic). Narrow columns indicate relatively small (pessimistic) contribution of the particular *variable*. For better understanding (Fig. 8), mean values of columns are shown underneath the map, useful in recognizing the smallest and the greatest contribution of variables in data matrix. In our case the most pessimistic variable with the smallest mean (2.13) is *employment* 2004, and the greatest value 2.47 occurs in *domestic sales* 2005. But the gap between minimal and maximal contribution is rather small.

Analogously, the height of each row (*company*) shows its contribution to the whole matrix – the higher is the row, the more optimistic are the *company's* answers (more precisely the sum of coded answers is larger).

The most important aspect of the overrepresentation map is the possibility of arbitrary segmentation of records into clusters by horizontal lines (visible on the map). After examining different numbers of possible clusters we found that 24 are most useful for interpretation (Fig. 8). In the middle of the map there is a strip of 207 rows with fair representation in each cell; all answers in a particular record are identical, they differ only in the height of the row (the height of the row is the smallest when there are only answers "1" and the largest when there are only answers "3"). This phenomenon will be shortly discussed in Section 3.6.

The next step of our analysis is selecting *companies* fitting the main trend of the whole data matrix (visible as a saddle surface, where the dark rectangles are lying possibly close to a decreasing curve from the upper left to the lower right corner) in contrast to *companies*, which decidedly differ from others and disturb the main structure. Elimination of the *outlying companies* from the whole data set yields a more regular matrix (better matching the main tendency) and helps us in practical interpretation of data fitting the main trend.

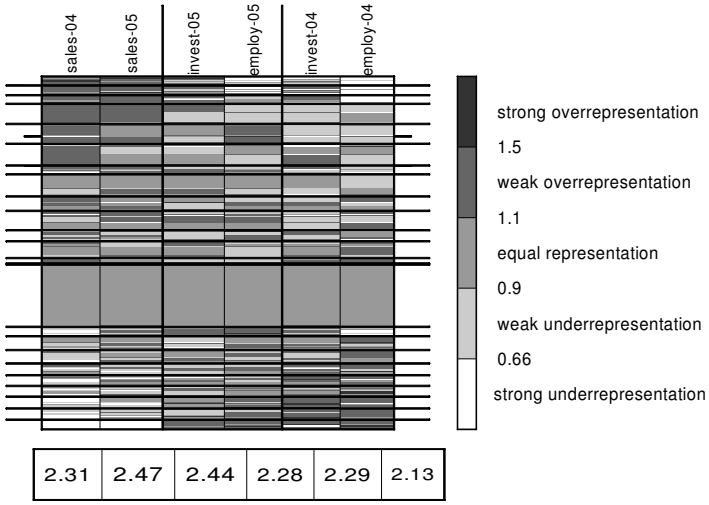

Figure 8. Overrepresentation map with 24 clusters for records and 3 clusters for variables (completed by mean values). The first records from cluster 1 and the last records from cluster 24 are shown in Table 3 and 16 records from cluster 5 (close to the mark visible as a stroke shorter than lines between clusters) are shown in Table 4.

It is interesting to see whether the main trend in Fig. 8 is sufficiently regular. A glimpse at Figs. 7 (right) and 8 shows that this matrix is not too regular. Let us start with an example indicating 16 records belonging to the 5-th cluster in Fig. 8 (from 233-rd to 248-th) by a mark (a stroke shorter than the lines between clusters).

A perfectly regular structure (trend), as shown in Fig. 4, is a rare occurrence. We usually have many records, which do not fit and disturb this structure. Six of them are presented in Table 4. They are visible even in Fig. 8, where the difference between them and the adjacent clusters, fitting regular structure, is visible. According to the regular structure, in the 5-th cluster *domestic sales* 2004 and *domestic sales* 2005 should be dark grey or black (Fig. 8), having values close to 3 (Table 4). But when we look at the map and the table we see that the order and regularity are disturbed. In cluster 5 in the *domestic sales* 2004 column a white line is visible, which is also mirrored in the Table 4 by 1's.

3.3. Dividing records into two parts: fitting the main structure and diverging from it

Cell colors in Fig. 9 present the grade of deviation from the main structure for the pairs of companies. If the rectangle is dark, the pair of companies differs

Table 4. Sixteen records from cluster 5 including six (grayed) not fitting the main trend.

GCA rank	sales-04	sales-05	invest-05	employ-05	invest-04	employ-04
233	3	3	3	3	2	2
234	3	3	3	3	2	2
235	3	3	3	3	2	2
236	1	2	2	1	1	1
237	1	2	2	1	1	1
238	1	2	2	1	1	1
239	2	2	3	2	2	1
240	2	2	3	2	2	1
241	2	2	3	2	2	1
242	2	2	3	2	2	1
243	1	3	3	3	1	1
244	1	3	3	3	1	1
245	1	2	2	3	1	2
246	3	2	1	1	2	2
247	3	3	3	1	3	2
248	3	3	2	3	2	2

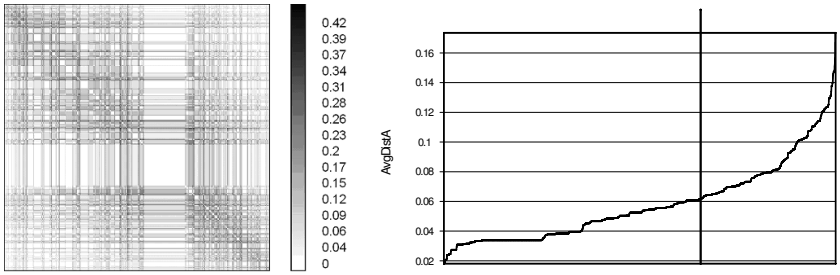

Figure 9. (left) Map indicating companies outlying from the main structure and (right) graph of AvgDist values with companies reordered according to their departure from the main structure, the higher is the AvgDist value the stronger is departure (right).

strongly from the main trend, and if the rectangle is light – pairs of companies fit the main structure.

If we want to carry out deeper analysis we have to calculate the values of a statistic *AvgDist* (Section 2.4), which evaluates departure of any row from the main structure. Thanks to this statistic we are able to segment records into two groups – the group called FIT and the group called OUT. We execute the

division by separating records with small values of *AvgDist* (FIT population on the left part of the graph – Fig. 9) and records with big values of *AvgDist* (OUT population on the right part of the graph). After such partition we can analyze each group individually. Before this we apply GCA to FIT and OUT to find new structures (trends). New overrepresentation maps are shown in Fig. 10; the first refers to FIT (885 rows) and the second to OUT (467 rows). We see that differences between FIT and OUT are large. On FIT map big separate clusters concentrating quite similar records are visible, while in OUT map the clusters are less homogeneous. The arrangement of variables in FIT and OUT is different, in OUT map the arrangement of variables was changed, while in FIT map the arrangement is the same as the original one. Beneath the maps (Fig. 10) mean values are shown, indicating the contribution of variables in FIT and OUT. In both FIT and OUT map the smallest mean is *employment* 2004 and the highest is *domestic sales* 2005, similarly like in the whole data matrix from Fig. 8.

By interpreting the mean values, shown in Fig. 10, we can learn a lot about data. For example, when we compare sums of mean values for 2004 and 2005 (the middle table in Table 5) we can say that generally situation in servicing sector in 2005 is seen more optimistically (higher values) than in 2004 (lower values). Mean values for FIT for 2004 and 2005 are respectively bigger than mean values for OUT for 2004 and 2005. This suggests that companies in FIT better manage present situation than those in OUT and have better prospects for the future.

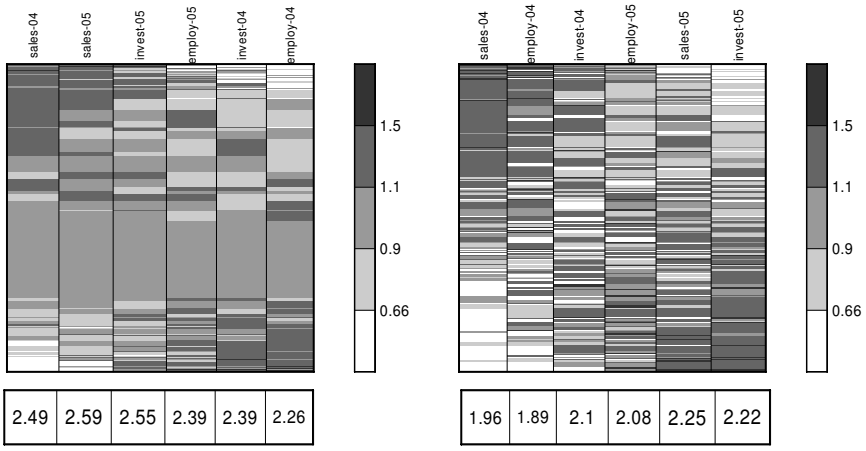

Figure 10. Overrepresentation maps: FIT (left) and OUT (right) - both after performing GCA completed by total means vectors (under the maps). Note different arrangements of variables.

Table 5. Tables of mean values for FIT and OUT. The upper table presents mean values for all variables for FIT and OUT; the middle table refers to mean values for years 2004 and 2005; the lower table shows the total mean values.

	sales-04	**employ-04**	**invest-04**	**employ-05**	**sales-05**	**invest-05**
FIT	2.49	2.26	2.39	2.39	2.59	2.55
OUT	1.96	1.89	2.10	2.08	2.25	2.22

	2004	**2005**
FIT	7.14	7.53
OUT	5.95	6.55

FIT	2.445
OUT	2.083

3.4. Set of records fitting the main trend (FIT)

As shown in Fig. 10, FIT overrepresentation map has a more consistent structure and regularity than OUT map. For a more concise analysis we have to divide the whole FIT map into clusters. We usually choose this partition, which gives visualization with possibly good interpretation. The chosen number of clusters in our case was 14, as shown in Fig. 11. The obtained clustering is still not very regular (outlying records in clusters 2, 5 and 13 are indicated in Fig. 11 by marks) but extremely homogeneous inside each cluster.

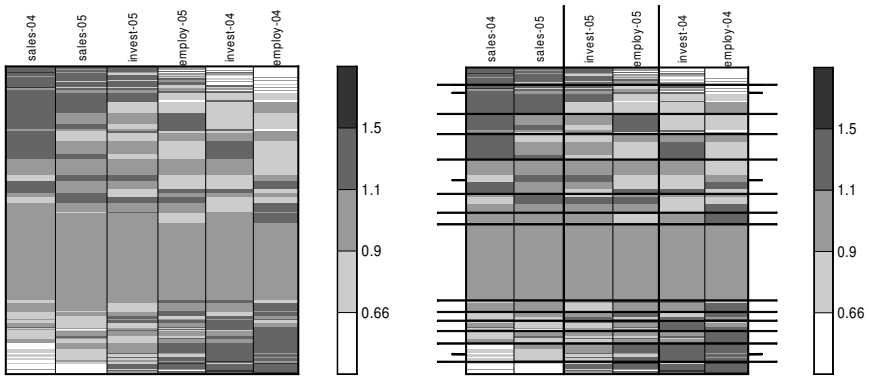

Figure 11. FIT overrepresentation maps without clustering (left) and with 14 clusters (right).

Now we calculate mean values for particular clusters, so as to get deeper understanding of the whole FIT data matrix. The mean values are shown on

the map of Fig. 12. We analyze the first and the last clusters characterized by very interesting properties.

The description of cluster 1 is: variables as *domestic sales* in 2004, *domestic sales* in 2005 and *investment* in 2005 have the highest values – each above 2 – meaning that they are estimated much more optimistically than such variables as: *employment* in 2005, *investment* in 2004 and *employment* in 2004, whose values fluctuate close to 1. We see that cluster mean values in the first cluster tend to decrease from the left (2.79) to the right (1.11), what is in accordance with the main trend in FIT.

	sales-04	sales-05	invest-05	employ-05	invest-04	employ-04
1	2.79	2.71	2.44	1.88	1.45	1.11
2	2.85	2.9	2.48	2.03	2.06	1.7
3	2.91	2.78	2.63	2.78	2	1.91
4	3	2.64	2.47	2	2.64	2
5	2.78	3	2.87	2.56	2.56	2.21
6	2.42	3	2.76	2.76	2	2.42
7	2.83	2.83	3	2	2.83	2.83
8	2.59	2.59	2.59	2.59	2.59	2.59
9	2.63	2.9	2.26	3	2.9	2.63
10	2.12	2.54	3	2.66	2.45	2.66
11	2.09	1.93	2.37	2.34	2.81	2.21
12	2	2.51	2.42	2.54	2.54	2.97
13	1.44	2.08	2.36	2.34	2.62	2.53
14	1.04	1.18	1.93	1.97	2.14	2.32

Figure 12. Clusters in FIT with mean answers inscribed into each cluster cell.

When we look at the pairs of variables in the first cluster we can compare the mean answers of particular companies. *Domestic sales* in 2004 and *domestic sales* in 2005 have the values of 2.79 and 2.71 – which we interpret as the situation in 2004 and the forecast for 2005 being practically the same. The next pair of variables: *employment* in 2004 and *employment* in 2005 makes interpretation more complex, because the respective values are 1.11 and 1.88,

meaning that businessmen might think about *employment* in the future much more optimistically than it was in 2004. The most interesting pair of variables are *investment* in 2004 and *investment* in 2005. There is a very big difference between year 2004 and forecast for 2005, the two values being, respectively, 1.45 and 2.44, which can be interpreted that companies in the first cluster have hopes for better profitable conditions in 2005 than in 2004.

Let us analyze cluster 14: the situation here is that the mean values for *domestic sales* in 2004 and *domestic sales* in 2005 are still nearly equal but with one essential difference: the values are very low, what we can interpret as a pessimistic view (a reversal of the first cluster where the values were high). Two pairs of variable values are different – *employment* and *investment*. The assessment of *employment* in 2004 is much more optimistic than the forecast for 2005 (2.32 compared to 1.97) and the same can be said about *investment* (2.14 and 1.93).

The structure between clusters 1 and 14 is nearly regular. In the matrix of aggregated clusters the post-GCA ordering of variables is the same as in FIT before aggregation and the ordering of row clusters is similar as in FIT. Yet, some cells in particular clusters disturb regularity, having high values of *AvgDist*. The most outlying clusters in FIT are numbered 7 and 9 (mean 2.83 for *employment* in 2004 for cluster 7 is much bigger and 2.26 for *investment* 2005 in cluster 9 is much smaller than expected). As we can see each cluster in FIT has its specific feature. But even when we have only basic information about variables we can easily see the main tendency and anomaly of the whole data matrix. See, e.g., cluster 8: means of all variables have all the value 2.59. All the answers were the same and this cluster can be interpreted as follows: businessmen marked answers automatically and without reflection (because it is rather rare for every question to be answered identically). However, we do not know whether these answers are based upon reliable information; so it could be safer to exclude them from further analyses.

The next step in our analysis is to construct the rank correlation table (Fig. 13), to compare outcomes of grade analysis with outcomes of a more standard method.

As we can see all correlations are positive. The rank correlations table for FIT is very regular (almost *Robinsonian*), which corresponds to the regularity of trend in FIT. Variables close in the map have stronger correlation than more distanced variables. The strongest positive dependence is between *domestic sales* in 2004 and *domestic sales* in 2005 (value 0.67), next between *investment* in 2005 and *domestic sales* in 2005 (0.59) and between *investment* in 2004 and *employment* in 2004 (0.51). So if *domestic sales* are high in 2004, the businessmen seem to believe that the *domestic sales* in 2005 will be high too, and there is a very similar case with correlation between *domestic sales* in 2005 and *investment* in 2005. And the next high correlation – *employment* in 2004 and *investment* in 2004, tells that increase of *employment* in 2004 mirrored the increase of *investments* in 2004.

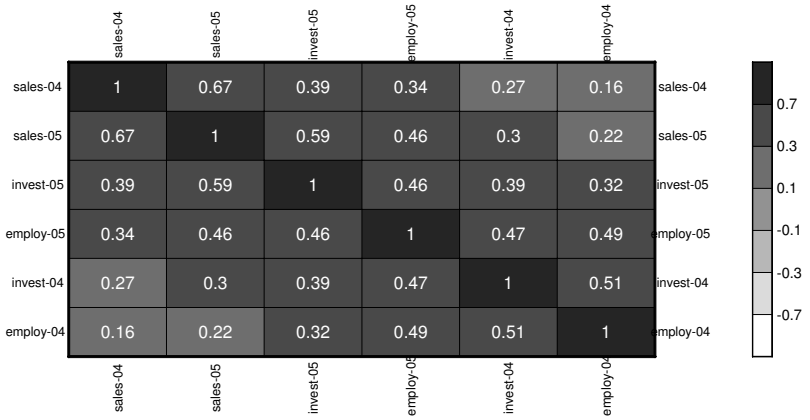

Figure 13. The rank correlation table for FIT (variables ordered according to the main trend).

The correlation matrix gives us only the most basic insight into the data structure in FIT. Without further data exploration its usefulness is very restricted. The more thorough and subtle information can be obtained only for particular clusters of companies (Fig. 12).

3.5. Data outlying from the main structure (OUT)

Analogous analysis can be performed on data outlying from the main structure. The first step is clustering. In our case we execute segmentation into 10 clusters, so there will be 24 clusters together (14 in FIT and 10 in OUT). This will make it possible to compare such *two stage clustering* with the *initial clustering* shown previously in Fig. 8.

When we look at OUT map we see that in columns the arrangement of variables has changed in comparison with FIT. On the left side are *economic factors* from 2004 and on the right side are *economic factors* from 2005. As we can see OUT group contains companies, whose answers from 2004 were significantly different from answers from 2005. By analyzing the first cluster (Fig. 14) we can say that variables of *domestic sales* in 2004, *employment* in 2004 and *investment* in 2004 reflect the optimistic view, while variables *employment* in 2005, *domestic sales* in 2005 and *investment* in 2005 tend to reflect the pessimistic view.

The analysis performed shows that the forecasts in the first cluster of OUT are rather pessimistic and mean answers for 2004 are optimistic, in contrast to the tendency in FIT population. This, though, is not surprising, as cases outlying from the main trend usually have more or less different properties than data in the whole matrix.

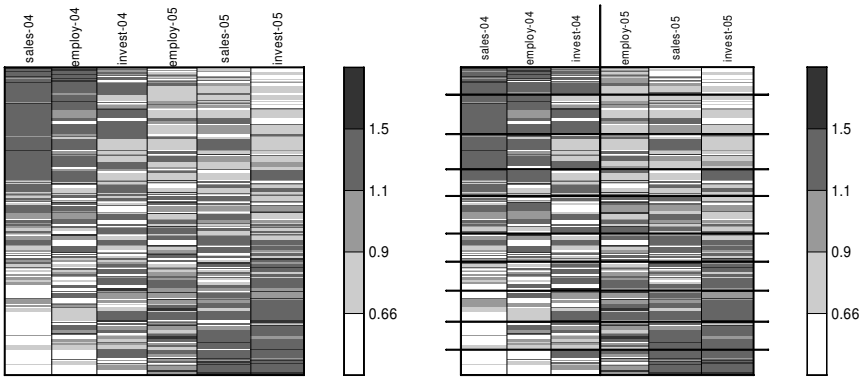

Figure 14. OUT overrepresentation maps without clustering (left) and with 10 clusters (right).

	sales-04	employ-04	invest-04	employ-05	sales-05	invest-05
1	2.89	2.56	2.56	2.07	1.89	1.38
2	2.68	2.33	2.36	1.89	2.21	1.68
3	2.91	2.47	2.39	2.19	2.34	2.17
4	2.42	2.1	2.12	1.85	1.92	2.27
5	2.12	1.94	1.94	2.28	2.3	2.16
6	1.9	1.95	1.88	2.11	2.42	2.42
7	1.48	1.51	2.21	2	2.06	2.44
8	1.2	1.4	1.83	2.07	2.05	2.35
9	1.04	1.57	1.91	2.2	2.46	2.62
10	1	1.04	1.85	2.12	2.87	2.82

Figure 15. OUT overrepresentation map with 10 clusters and mean values in each cluster.

This analysis shows that forecasts for the Polish companies in the servicing sector belonging to the first OUT cluster are not optimistic. But we also should remember that this data matrix of OUT contains just cases outlying from the main trend, which can have very different properties than whole data from the questionnaire.

The aim of each grade data analysis is to find the main tendency of data and to find variable or variables, which most strongly influence the order of other variables. In our case, both in FIT and OUT, we find the most important and the strongest variable – *domestic sales* 2004, which influences the respective orders (in FIT and in OUT) of other variables.

The further research will explore other aspects of this data matrix.

3.6. Psychometrical remarks on the survey

Interpretation of the answers to the survey would be incomplete without cautious analysis of the instrument itself. There are some points, which may play important role during interpretation of the dataset.

The first problematic point is the meaning of terms used in questions. So, "increase", "decrease" and "no change" might be understood quite differently in a small family business and in a big company employing 1000 workers. Employing additional 5 persons in big company could be interpreted as an increase, or as no change (5 workers is only 0.5% of 1000 workers, so the increase is extremely small).

The second issue is: on what data were the answers based? We do not know whether the data were sound, e.g. taken from accounting department, or were just an "impression" without detailed analysis. The survey gives no direct recommendation that the answer has to be based on actual data. Cluster filled with 207 identical answers for every question might be a sign that sometimes person was basing on imaginary data or gave "automatic" answers, without deeper involvement.

The last issue is that any company is expected to be strong, dynamic and developing. In case of a company facing decrease in *employment* and in level of *investment* there is a possibility of intentional manipulation, resulting in giving answers that hide the undesired truth. In other words some people might have given answers more optimistic than they should have provided. Presenting a company – even in an anonymous survey – as strong and developing might be a part of company's marketing strategy and may explain why there are so many optimistic forecasts. The answers are generally strongly optimistic, that is why mean answers in columns for the whole matrix are so high (over 2). However Poland in year 2004 has joined the European Union and it can also be one of possible explanations why in this year companies gave so many optimistic forecasts.

4. Conclusions

GCA with posterior clustering has helped to find some especially interesting groups of records in Survey 2005 (e.g. clusters 1 and 14 of FIT cases, and clusters 1 and 10 of OUT cases) and cluster 8 in FIT, which probably has disrupted the revealed structure of data and thus made the interpretation of results rather

difficult. There is however a psychometric reservation that the Survey itself has weaknesses.

It is very important to stress here that our aim was to explore data, not to test hypothesis or make a synthesis. We hoped to find some non-trivial trends in data by separating rows into regular FIT sub-table and outlying the trend OUT sub-table. It seems that servicing companies in FIT better manage present situation than companies in OUT and overall have better prospects for the future. Of course connecting this categorization with external information on each company (for example about its profile or size) would help to point which branches of servicing companies were doing badly or well. However, we hope to do it in future work that will provide more information on profiles of developing, collapsing and stagnant Polish companies, and also will cover remaining *economic factors* from the original matrix and additional information on company – *provinces*, *number of employers*, *lines of business.* Regrettably, to both introduce grade framework to a reader and to analyze data we have had to concentrate on the simpler version of data set.

Grade exploration of data supplements traditional methods. It is still possible to construct or use some existing indices that forecast general trend of servicing trade, but such information gives only general view on condition of the whole branch. Grade exploration allows for quickly grasping general trend in data, and then to cluster companies with similar profiles. It is thus possible to find isolated companies which, for instance, assess forecasts much worse than their actual condition. And as an exploration method grade data exploration does not test any hypothesis, but gives some hints what to test or to verify. Grade exploration allows for extracting a more regular sub-table of original table, free from disturbing objects, and to form another sub-table of outlying objects that may help in gaining insight into the structure of analyzed data. The data obtained by grade analysis might be then a basis for classical methods used, for example, by the Central Statistical Office or the National Bank.

Acknowledgement

We kindly thank the Polish Chamber of Commerce for allowing us to analyze their data.

References

Ciok, A., Kowalczyk, T., Pleszczyńska, E. and Szczesny, W. (1995) Algorithms of grade correspondence-cluster analysis. *Archiwum Informatyki Teoretycznej i Stosowanej* **7**, 5-22.

Ciok, A. (2000) Double versus optimal grade clustering. In: H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen and M. Schader, eds., *Data Analysis, Classification, and Related Methods.* Springer, 41-46.

Ciok, A. (2004) *On the number of clusters – a grade approach.* Instytut Podstaw Informatyki PAN, Warszawa.

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

Grzegorek, M. (2005) Image Decomposition by Grade Analysis - an Illustration. In: M. Kurzyński, E. Puchała, M. Woźniak and A. Żołnierek, eds., *Computer Recognition Systems, Proceedings of* IV International Conference CORES'05. Springer, Berlin-Heidelberg, 387-394.

Grzegorek, M. (2007) Towards an exploration of GCA ordered pixels. In: M. Kurzyński, E. Puchała, M. Woźniak and A. Żołnierek, eds., *Computer Recognition Systems 2. Advances in Soft Computing.* Springer, Berlin-Heidelberg, 156-163.

Jarochowska, E. and Ciesielski, K. (2006) Grade clustering and seriation of words based on their co-occurrences. In: V.P. Guerrero Bote, ed., *Current Research in Information Sciences and Technologies. Multidisciplinary approaches to global information systems, Proceedings of the I International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006*, **2**, Merida, Spain, 52-56.

Kowalczyk, T., Pleszczyńska, E. and Ruland, F., eds. (2004) *Grade Models and Methods for Data Analysis. Studies in Fuzziness and Soft Computing* **151**. Springer, Berlin-Heidelberg-New York.

Książyk, J.B., Matyja, O., Pleszczyńska, E. and Wiech, M., eds. (2005) *Analysis of medical and demographic data with the use of program GradeStat* (in Polish). Instytut Podstaw Informatyki PAN, Warszawa.

Matyja, O. (2002) Smooth Grade Correspondence Analysis and Related Computer System. PhD Thesis, Institute of Computer Science, Polish Academy of Sciences, Warszawa.

Pleszczyńska, E. (2007) Application of Grade Methods to Medical Data: New Examples. *Biocybernetics and Biomedical Engineering* **27** (3) 77-93.

Stasiecka, A., Płodzień, J. and Stemposz, E. (2006) Measures for estimating the quality of e-learning materials in the didactic aspect. In: *Proceedings of the Second International Conference on Web Information Systems and Technologies: Society, e-Business and e-Government / e-Learning,* Setúbal, Portugal.

Szczesny, W. (1991) On the performance of a discriminant function. *Journal of Classification* **8**, 201-215.

Szczesny, W. (1999) Outliers in grade correspondence analysis. In: M. Michalewicz, M. Kłopotek, eds., *Intelligent Information Systems IIS'99, Procedings of the Workshop held in Ustroń, Poland.* Instytut Podstaw Informatyki PAN, Warszawa, 332-336.

Szczesny, W. (2000) Detecting rows and columns of contingency table, which outlie from a total positivity pattern. *Control and Cybernetics* **29** (4), 1059-1073.