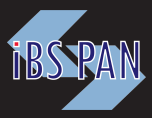

INSTYTUT BADAŃ SYSTEMOWYCH
POLSKIEJ AKADEMII NAUK

TECHNIKI INFORMACYJNE TEORIA I ZASTOSOWANIA

Wybrane problemy
Tom 4 (16)

poprzednio

ANALIZA SYSTEMOWA W FINANSACH
I ZARZĄDZANIU

Pod redakcją
Andrzeja MYŚLIŃSKIEGO

Warszawa 2014

INSTYTUT BADAŃ SYSTEMOWYCH
POLSKIEJ AKADEMII NAUK

TECHNIKI INFORMACYJNE TEORIA I ZASTOSOWANIA

Wybrane problemy
Tom 4 (16)

poprzednio

ANALIZA SYSTEMOWA W FINANSACH I ZARZĄDZANIU

Pod redakcją
Andrzeja Myślińskiego

Warszawa 2014

Wykaz opiniodawców artykułów zamieszczonych
w niniejszym tomie:

Prof. Bernard De BAETS

Dr hab. Ewa BEDNARCZUK, prof. PAN

Dr hab. inż. Wiesław KRAJEWSKI, prof. PAN

Dr hab. inż. Andrzej MYŚLIŃSKI, prof. PAN

Dr inż. Jan W. OWSIŃSKI

Dr hab. Dominik ŚLĘZAK, prof. UW

Prof. dr hab. inż. Andrzej STRASZAK

Prof. dr hab. inż. Stanisław WALUKIEWICZ

ISBN 83-894-7555-3

GRADE DATA ANALYSIS APPLIED TO THE EUROPEAN AGRICULTURE

Stanisław Lenkiewicz

Systems Research Institute, Polish Academy of Sciences,
Ph. D. Studies, Warsaw, Poland,
stan@lenkiewicz.eu

Abstract. The paper presents results of the analysis of agriculture in the European Union. Based on 15 key characteristics, member countries have been divided into groups consisting of countries with a similar condition of agriculture. As a research method the Grade Data Analysis has been applied, and to treat the data the GradeStat software has been used.
The text is composed of four parts. The first part is a brief analysis of agriculture as a sector of the economy, while the second is dedicated to the measurement of the condition of agriculture. The third part presents the research tools: cluster analysis and Grade Data Analysis, and the fourth shows the results of using these tools for the assessment of EU agriculture.
Keywords: agriculture, productivity, cluster analysis, Grade Data Analysis, overrepresentation, outlier

1 AGRICULTURE AS A SECTOR OF THE ECONOMY

Agriculture constitutes an important sector of economy in almost every country. For some countries it is an important source of revenue and wealth, while in some other ones its purely economic role may be quite limited. The latter is the case for most of the developed world. Yet, with only few exceptions, agriculture is considered to be of high importance, if not for just economic reasons, then for the social, cultural, and, most of all, political ones. Agriculture provides food, and is a producer of raw materials for many industries, but it is also the reservoir of very important resources, including land, landscape, cultural heritage, people, and their specific know-how. Thus it is in the centre of interest of countries' governments, as well as of various international communities, such as the European Union (EU) or the FAO.

Because of its importance and complexity, agriculture is also an area of many studies. However complex they may get and how many factors they take into account, they all have two main purposes. First, they try to assess the "health" of agriculture. Secondly, they aim – on the basis of assessments – at identifying the ways of improving its condition. In particular, within the EU, a major share of the Community's budget has always

been going into agriculture; first, in order to keep the European agriculture alive in the face of competition from the outside (the developing world), and from the inside (the other sectors of economy, including construction, transport, etc.), and secondly, to retain the rural economy and society, as well as landscape, as an element of the European culture and identity that has been developing over millennia.

1.1 Specific

Agriculture is "based on the ground", i.e. it uses the land as its primary resource. The land, even if in the case of highly intensive farming its area can be minimised, constitutes the sine qua non condition of agricultural production, whatever the "national specialisation" or orientation in a given country. This fact makes agriculture different from a majority of sectors of the economy.

Agriculture is dependent on many factors over which people have no influence or which can be formed by people to a limited degree. The most important of them are: climate, surface relief, soil quality, precipitation. Another set of factors is related to the farming population – its skills, traditions and attachment to farming and to land. As a result, agricultural production is always specific to a particular country and agricultural products are the subject of intense trading, especially between countries with different natural conditions and different farming cultures.

1.2 Internally and Externally Differentiated

There are two main activities in agriculture: the cultivation of plants and livestock husbandry. They are very diversified. Some plants are used for food production, some of them serve as animal feed, while others become industrial raw materials (e.g. for the production of medicines, fabrics, or biofuels). Animals also can be bred to produce food (meat, milk, eggs, honey, etc.), or to provide raw materials (e.g. wool, leather, or silk).

There are, also in Europe, "traditional" farms, dealing with both plant cultivation and animal husbandry, as well as the specialized ones: focused on one particular activity (e.g, dairy farms). There are highly mechanized farms, as well as those that use mechanization to a lesser extent. Some farms use large quantities of chemicals, while others – small or not at all. Little use of machinery and chemicals does not necessarily result from backwardness; on the contrary, may be the result of a decision to start organic farming activity, which is increasingly popular in highly developed countries.

Thus, agriculture is very diversified even within one country. Furthermore, it is generally highly diverse among various countries, especially when they have different natural conditions for farming and different farming traditions.

1.3 Critically Important

It is hard to believe that in the 21st century, many people still suffer from hunger. Though the number of undernourished persons dropped from 15% at the beginning of the century to 12% now, it is still unacceptable.[1] Even a short-term food shortage is a serious problem and can be a source of social unrest. This is an explanation as to why agriculture is in the focus of national governments, as well as of international communities,[2] one of which is the EU.

Although one of characteristics of the developed countries is some kind of market economy, agriculture – as its specific and particularly important sector – is subject to special rules, which are usually presented in the form of so-called "agricultural policy". One of its elements are various forms of support for farms, among others: subsidies. Such an "exception to the rule", i.e. exerting an explicit influence on one branch of the economy through administrative decisions, is the subject of dispute among politicians, as well as among researchers dealing with the problems of agriculture.

1.4 Hard to Assess

It is very hard to assess the condition of agriculture. First of all, one should explain what is meant by the notion of "condition". Is it just a global value of agricultural goods produced by a certain country or region? Or should it be presented as a set of values describing various areas of agricultural activity? How to take into account objective conditions such as climate or soil quality? How to deal with the level of mechanization and fertilizer use?

Some countries have greater capacity of agricultural production than their demand for agricultural products. Many of them introduced various limits, forcing or motivating farms to produce below their capacities. In such a case can an evaluation be fair?

[1] Source: FAO's Hunger Portal (http://www.fao.org/hunger/en/, access: June 27th, 2014).

[2] There are many studies of the state of agriculture in developing countries which are conducted by scientists from those regions (e.g. [9], [11]), as well as numerous analyses of agriculture in those regions done by world-renowned scholars from Europe and the US (e.g. [13]). This is not surprising.

For many, the efficiency – i.e. the ratio of product (value) to costs (resources used) – is the codeword. Yet, again, both natural conditions and the (local) specificity of product mix and technologies may make the respective comparisons highly doubtful.

These are the most important questions that must be answered for an assessment to have any value. We will deal with them in more details in the following section.

2 CONDITION OF AGRICULTURE. CAN IT BE MEASURED?

Let us put it clearly: **any assessment is more or less subjective**. It is always a particular person (or a group of people) who prepares rules for the evaluation. We all follow our own perceptions of the world, thus: our individual hierarchies of values, so the assessments we make are not objective ones; they are **our assessments**. Someone who claims to be objective, is naive – or simply is lying.

It is difficult to be objective (or, strictly speaking: not too subjective) even in assessing simple subjects. It is very hard to be so when dealing with complex problems of great importance – and the condition of agriculture is one of them. So we are not trying to pretend that we strive to make a fully objective assessment of agriculture; we only aim at an evaluation the least marked by our personal convictions. Indeed, we can deal away with at least an important part of the "subjectivity" approach by making explicit the criteria of evaluation, and/or the characteristics that are taken into account ("the assessment is made from THIS point of view"). Such an explicit specification of the elements of assessment is particularly proper for the case of EU agricultural policy, where the goals and criteria have been changing from period to period. In the recent periods, care was taken to avoid overproduction that might have resulted from application of subsidies, while preserving the rural farming activities throughout Europe and leaving room for competition in terms of production costs and technology-and-product mix.

2.1 An Assessment – What Is It?

To assess something, one should collect data and construct a method to process them. To do this, he or she should:

- Prepare a list of characteristics that will be covered by the assessment.
- Prioritise these characteristics, i.e. decide whether they should be equally weighted or their weights should be different.

– If necessary: normalize weighted characteristics.
– Choose the way to treat pre-processed characteristics to obtain a measure (or measures) of the phenomenon of his or her interest.

These elements determine the entire study, highly influencing its outcome. Describing these "pillars" means: presenting the method of research.

2.2 Condition of Agriculture

As stated in Section 1.4, the term "condition" is abstract. Before we get to the methods of its measurement, we have to be more specific. For many researchers it is a synonym of the agricultural **productivity**. Productivity is understood as the **ratio of outputs to inputs** in production; thus agricultural productivity is the ratio of agricultural outputs to agricultural inputs.[3] There are also many publications where the word "productivity" has been replaced by "**efficiency**".[4] Sometimes these terms are used interchangeably.[5]

There is a mess in terminology; however, in each analysis of agriculture we may find the same basic element: it examines how well outlays (inputs) are processed in the results (outputs), and tries to present it in numerical form.

Most researchers list as inputs of "agricultural production system" three elements: **land, labour, and capital**. Some of them perform for them separate analysis, highlighting productivity of land, productivity of labour, and productivity of capital.

Analysis for certain regions (e.g. for EU or NAFTA) become a basis for comparison of agricultural productivity in the countries of these regions. They permit to identify strengths and weaknesses of each country and to suggest ways for improvement.[6] Assessments made for agriculture may also serve for comparison of its productivity and productivities observed in other sectors.[7] The results may be surprising.[8]

[3] See [9].
[4] See [30].
[5] See [7].
[6] See [3], [25], [27], [28], [33].
[7] See [12], [13].
[8] As Gollin, Lagakos and Waugh observed in [12], *according to national accounts data, value added per worker is much higher in the non-agricultural sector than in agriculture in the typical country, and particularly so in developing countries. Taken at face value, this "agricultural productivity gap" suggests that labour is greatly misallocated across sectors. [. . .] even after considering sector differences in hours worked and human capital per worker, as well as alternative measures of sector output constructed from household survey data, a puzzlingly large gap remains.*

2.3 Measuring Agricultural Productivity

There are a great number of methods used for agricultural productivity assessment. In this paper it is not possible neither to present all of them, nor to delve deeply into the details. Some examples may be found in [3], [7], [9], and [23].

In all methods we may find these common elements (see Section 2.1):

- Choosing characteristics (attributes) to be analysed (e.g. total crops output per hectare, fertilisers usage per hectare, total subsidies on crops).
- Pre-processing these characteristics (weighting, normalizing).
- Processing the data to obtain productivity metric(s).
- Analysis of the results. Very often: grouping the objects with similar profiles (e.g. division of the EU members into groups of countries with a similar profile of agriculture).

Such a procedure, and especially its last step, makes it natural to consider the application of data mining and cluster analysis. It is surprising that there are very few publications related to agriculture which use them.[9]

3 CLUSTER ANALYSIS AND GDA

The aim of every research is to increase knowledge. This may mean either discovering new facts, or deeper understanding of facts already known. In both cases: we want to know more about a certain subject.

We may divide all research methods into two groups: quantitative and qualitative ones. However, the boundaries between them are blurred; in fact, usually we use a mixture of both. The best example of this is data mining.

3.1 Data Mining

There is no clear definition of data mining. It is one of the stages of acquiring knowledge from the data: analysing large sets of raw values (very often having the form of databases or data warehouses), retrieving new values and interpreting them. We may say: data mining is a way to discovering patterns in large data sets, it is the search for order in chaos. As almost every study begins with an deep analysis of data, data mining is largely used in various areas of research.

Data mining is a set of methods, heuristics, and algorithms, and it is usually placed between computer science and statistics.[10]

[9] Some examples: [1], [22], and [26].

[10] See [10].

3.2 Cluster Analysis[11]

We called data mining *discovering patterns in large data sets*. Very often discovering patterns means putting objects under investigation into groups (called clusters) in such a way that those in the same group are more similar to each other than to the objects in other groups. Such an approach is called cluster analysis.

Grouping similar objects (clustering them) has two advantages. First, it makes possible to analyse data structure, to discover relations between various objects, and finally: to get information about the problem being investigated. Secondly, it can point out the direction of further study, allowing researchers to focus their attention on selected groups, thus: to reduce the search area. The analysis of discovered clusters is bidirectional. On one hand we investigate the internal structure of clusters, on the other – we try do find similarities and differences between various clusters.

To apply the cluster analysis one must decide which attributes of the investigated objects will be taken into account. Moreover, some of them may be correlated, so they should not be included in the same research. An opinion of a researcher, though important, it is not sufficient; there must be a clear rule of deciding which attributes of the objects are important, which are not and why.[12] If these attributes should be weighted the problem becomes much more complicated. Weights have a strong influence on the result of the analysis should therefore be chosen with great care.

The result of clustering the objects, i.e. determining which ones are "close" (or similar) to each other, depends on the **measure of similarity** (sometimes also called resemblance coefficient). Its choice is one of the most important steps in the whole analysis, and not an easy one. Attributes of objects may have various natures: they can be nominal (e.g. country of residence), ordinal (e.g. education level), or interval (e.g. age), sometimes they are ratios (e.g. inflation or unemployment rate).[13] They can be measured on different scales, so an initial normalisation may be necessary.

The purpose of any cluster analysis is to identify clusters. So the question: *how many should they be?* is natural. Unfortunately, there is no one clear answer. The number of clusters depends basically on the purpose of the research, as well as... on the preferences of the researcher.[14]

---

[11] See [21].

[12] This procedure is also called **choosing input variables**. [20] covers this problem in details.

[13] See [29], p. 93-177.

[14] [5] presents a detailed analysis of this issue.

We have briefly discussed only the most important problems of the cluster analysis. A detailed description of this method is beyond the scope of this article. More information may be found in [2] and [29].

3.3 Grade Data Analysis

As stated in Section 3.1, data mining is a set of various methods. In this article we will look at one of them, less known but very useful: Grade Data Analysis (GDA). It has been developed at the Institute of Computer Science – Polish Academy of Sciences.[15] Its main component is the Grade Correspondence Analysis (GCA) – an algorithm which has been implemented in GradeStat program.[16]

GDA allows for a comprehensive analysis of the data including cluster analysis and a detection of outliers. Here we present only an outline of this method; a detailed description may be found in [19], and some examples of its application: in [14], [17], and [36].

As we mentioned in Section 3.2, one of critical parameters of each cluster analysis is a measure of similarity of objects. In GDA these are **concentration indexes** which play this role. The concentration index is associated with the **concentration curve**. We will now explore both these notions.

3.3.1 Concentration Curve

Let us consider the following example. Table 1 presents the results of a survey conducted by Eurostat in 2011: self-perceived health status of responders. The study included the European Union countries, Iceland, Norway and Switzerland.[17] Each column shows the percentage of respondents who assessed their health as very bad, bad, fair, good, or very good. We may easily notice that results for Greece and Netherlands differ. But how much? To answer this question we should plot a concentration curve of Dutch results relative to Greek results.

Each value in Table 1 may be interpreted as a probability that a person randomly selected from a country indicated in the first column perceives his or her health status as very bad, bad, fair, etc. Such a probability table is

[15] Website of the Institute: http://www2.ipipan.waw.pl/index.php/en/ (access: July 5th, 2014).

[16] The CD with the program is included in [17]. To unlock some of functions it is necessary to have a registering code. It may be got via the website http://gradestat.ipipan.waw.pl/english/ (access: July 5th, 2014).

[17] [17] uses similar data to show the idea of GDA and how GradeStat application works.

Table 1. Self-perceived health in 2011 (% of total responders)

Country	Very bad	Bad	Fair	Good	Very good
Austria	1.9	7.2	21.5	38.2	31.2
Belgium	2.1	7.4	16.9	43.9	29.7
Bulgaria	2.5	9.3	21.1	50.1	17.0
Croatia	5.3	22.2	27.5	29.6	15.4
Cyprus	2.5	5.5	16.4	28.6	47.0
Czech Republic	2.4	10.1	28.0	40.5	19.0
Denmark	2.5	5.8	20.9	42.8	28.0
Estonia	2.2	14.0	32.0	44.0	7.8
Finland	1.2	6.2	23.7	47.3	21.6
France	1.2	7.6	23.6	45.0	22.6
Germany	1.5	6.7	27.0	48.2	16.6
Greece	**2.7**	**6.3**	**14.6**	**25.8**	**50.6**
Hungary	4.0	12.2	27.9	39.9	16.0
Iceland	1.6	4.7	16.0	36.2	41.5
Ireland	0.4	2.5	14.0	40.1	43.0
Italy	3.0	10.1	22.2	51.6	13.1
Latvia	3.5	13.8	35.9	42.7	4.1
Lithuania	3.8	16.1	36.2	37.3	6.6
Luxembourg	1.5	6.4	19.6	46.5	26.0
Malta	0.7	3.4	25.1	47.1	23.7
Netherlands	**0.8**	**4.9**	**17.9**	**55.2**	**21.2**
Norway	1.2	7.3	18.3	48.8	24.4
Poland	2.5	12.4	27.5	39.7	17.9
Portugal	4.9	13.2	32.2	40.3	9.4
Romania	1.8	7.8	21.0	42.2	27.2
Slovakia	3.0	10.5	23.3	44.1	19.1
Slovenia	2.9	10.4	26.3	41.8	18.6
Spain	2.0	5.5	17.4	53.6	21.5
Sweden	1.0	3.7	15.4	41.4	38.5
Switzerland	0.7	2.9	15.2	49.8	31.4
United Kingdom	1.0	4.7	16.8	42.0	35.5

a starting point when drawing each concentration curve.[18] Starting from it, we create a table of cumulative distributions for all responses in all countries. Table 2 presents cumulative distributions for Greece and Netherlands. Each value may be interpreted as a probability that a person randomly selected from a given country perceives his or her health status as very bad, at least bad, at least fair, etc.

Table 2. Cumulative distributions for Greece and Netherlands

Country	Very bad	Bad	Fair	Good	Very good
Greece	0.027	0.090	0.236	0.494	1.000
Netherlands	0.008	0.057	0.236	0.788	1.000

We draw points: (0, 0), (0,027, 0,008), (0,090, 0,057), (0,236, 0,236), (0,494, 0,788), (1, 1) and connect them with segments. We receive a concentration curve of Dutch opinions relative to Greek opinions C (*Netherlands* : *Greece*) – see Fig. 1.

In a similar way we could plot the concentration curve for the inverse relationship, i.e. C (*Greece* : *Netherlands*). This curve would be a reflection of the first one over the diagonal of the coordinate system.

A concentration curve is composed of as many segments as there are characteristics of the objects being analysed. Each segment illustrates a comparison of a "concentration" of one characteristic in both objects. If a segment has a slope lower than 45°, the "concentration" of a characteristic is greater in the "horisontal" object than in the "vertical" one; a slope greater than 45° illustrates an inverse relation; a slope equal to 45° – identical "concentration" in both objects. In Fig. 1 we can notice that people perceiving their health status as very bad or bad are more numerous in Greece (slope lower than 45°), those who perceive their health status as fair or good are more numerous in Netherlands (slope greater than 45°), and responders perceiving their health as very good are more numerous in Greece. We will say, in GDA terminology, that responses "Very bad", "Bad", and "Very good" are **overrepresented** in Greece, and responses "Fair" and "Good" are **overrepresented** in Netherlands.[19]

[18] Note: It happens very often that we have a table with absolute values and not probabilities (e.g. here we could have a table with numbers of respondents). In this case we obtain a probability table by dividing each value by the sum of its row.

[19] Note: It has no impact on the interpretation of the overrepresentation value whether a certain segment of the concentration curve is under or above the diagonal. It is its slope what represents the overrepresentation.

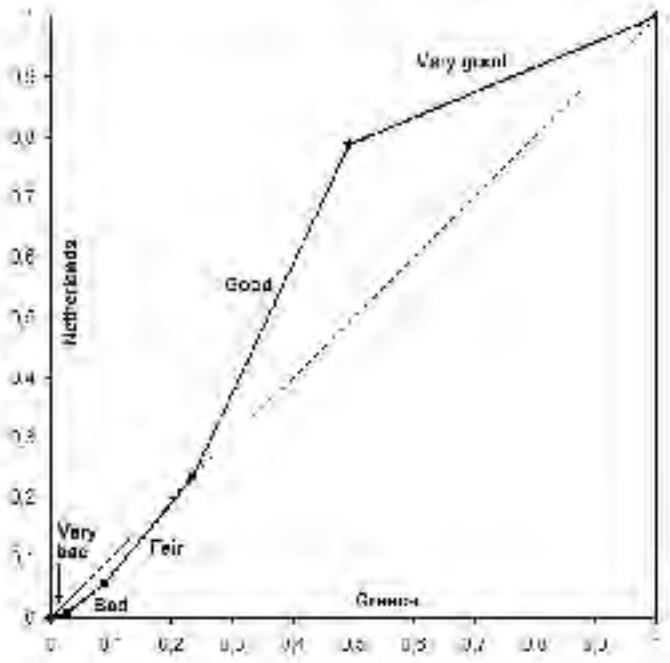

Fig. 1. Concentration curve of Dutch opinions relative to Greek opinions

There are as many concentration curves as permutations of segments which build them.[20] All concentration curves carry the same information about the overrepresentation of every particular attribute in compared objects. There is, however, one curve which also carries additional information: it shows total dissimilarity of two objects, taking into account all their attributes. It is a maximal concentration curve.

3.3.2 Maximal Concentration Curve

Some segments of the concentration curve are under the diagonal, some others – above it. Let us reorder segments based on their slopes, from the smallest slope to largest one. We obtain the curve presented in Fig. 2. It is the **maximal concentration curve**. It is convex and lies totally under the diagonal.

The area bounded from the bottom by the maximal concentration curve and from the top – by the diagonal has a special meaning. It is the largest of all the areas bounded by any concentration curve and the diagonal (proof

[20] This is also a number of permutations of objects' attributes.

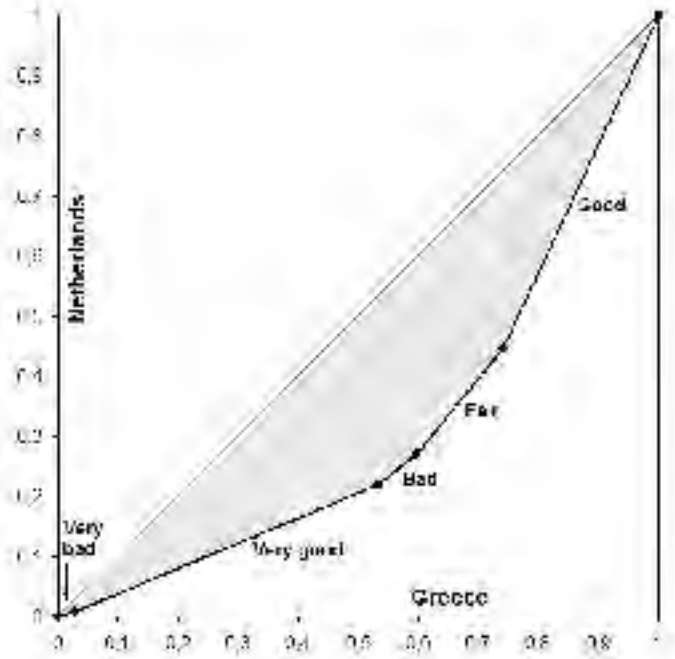

Fig. 2. Maximal concentration curve of Dutch opinions relative to Greek opinions

may be found in [19]), so it is a measure of dissimilarity of compared objects. It is an equivalent of a distance of objects used in more "traditional" cluster analysis methods. Here: this area represents dissimilarity of perceptions of their health by Greeks and by the Dutch.

We can compare the results of the survey for each pair of countries. Comparison of the areas between each maximal concentration curve and the diagonal will tell us the inhabitants of which two countries perceive their health status the most dissimilarly.

3.3.3 Concentration Curve For Columns

We compared two rows of the table, i.e. two countries. But we can also compare two columns, i.e. two perceptions of health status. For instance let us compare the answer "Very good" and the answer "Very bad".[21] Fig. 3

[21] In Section 3.3.1 we compared results of the survey for two countries. In this comparison countries were for us objects and answers of responders were attributes of these objects. Here we are going to compare two answers, so responders' answers are objects and countries are attributes.

presents the concentration curve for this comparison,[22] and Fig. 4 – the maximal concentration curve.

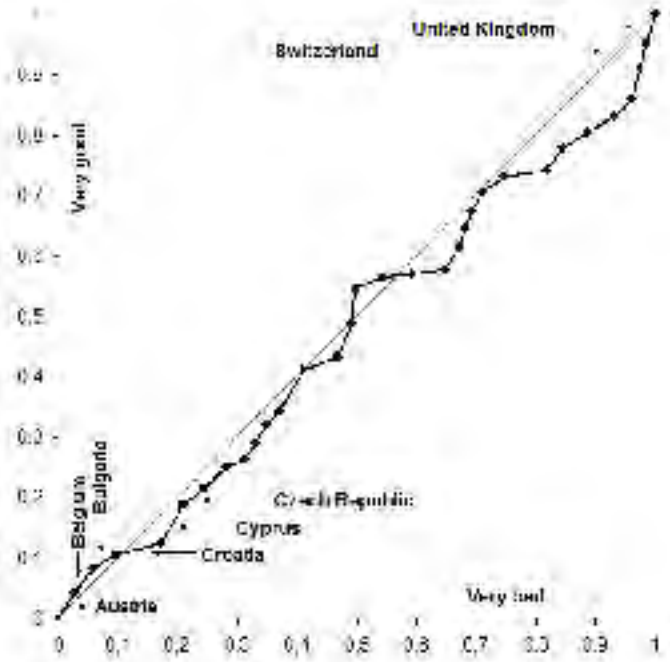

Fig. 3. Concentration curve of the answer "Very good" relative to the answer "Very bad"

Because we can compare pairs of rows of the table as well as pairs of columns, we can find both: dissimilar objects and dissimilar attributes of these objects.

Measuring dissimilarity of objects will be covered in more detail in Section 3.3.4.

[22]Note: To draw this curve, first we need to have cumulative distributions for both answers. To obtain them, we divide each value by the sum of its column, thus obtaining probabilities for both answers: "Very bad" and "Very good".
It is important to understand the difference between probabilities for countries (rows of the table, no data transformation necessary) and probabilities for health status perceptions (columns of the table after the operation described above). Values in rows are probabilities that a randomly chosen responder from a given country has given a certain answer (e.g. that a responder from Greece has answered *I perceive my health status as very good*). Values in columns are probabilities that a randomly chosen answer has been given by a responder from a given country (e.g. that a response *I perceive my health status as very good* has been given by a responder from Greece).

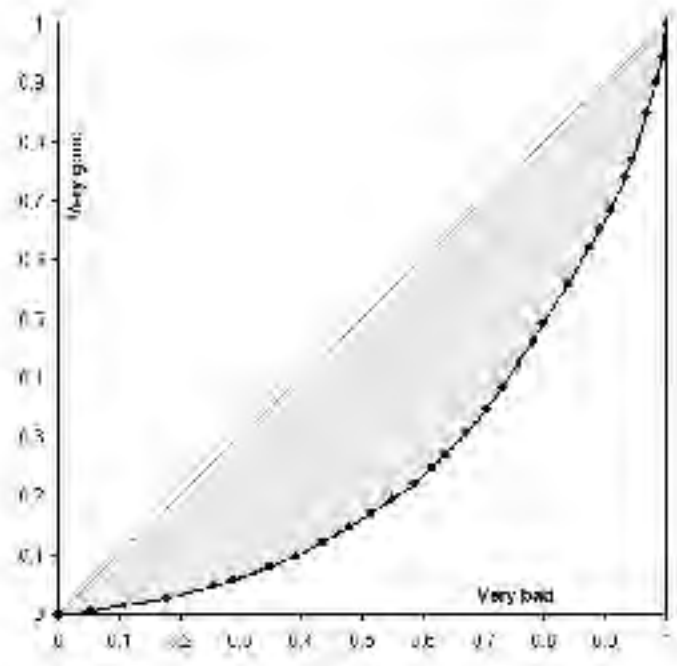

Fig. 4. Maximal concentration curve of the answer "Very good" relative to the answer "Very bad"

3.3.4 Concentration Index

Each concentration curve can be seen as a graph of the cumulative distribution function $F(X)$ of some continuous random variable X. This variable is defined on the interval [0; 1] and has the expectation (mean value)

$$E(X) = \int_0^1 xf(x)dx, \tag{1}$$

where $f(x)$ is its density function. The transformation of formula (1)

$$E(X) = \int_0^1 x\frac{dF(x)}{dx}dx = \int_0^1 xdF(x) = 1 - \int_0^1 F(x)dx \tag{2}$$

shows that this expectation is equal to the area above the concentration curve.

The expectation may be used as a measure of dissimilarity of objects being compared. When those are identical, the concentration curve coincides with the diagonal and the expectation is equal to $\frac{1}{2}$; when they are totally dissimilar, the expectation is either 0 or 1. However, it would be

fine to have a measure of dissimilarity which is equal to 0 for identical objects and equal to 1 or -1 when they are totally dissimilar. We can easily construct it:

$$ar = 2\left(E(X) - \frac{1}{2}\right). \tag{3}$$

We call this measure a **concentration index**.

A denotation ar comes from the word *area*. We will now see why. Let us look at Fig. 5. It presents the concentration curve of Dutch opinions relative to Greek opinions that we have already seen in Fig. 1. The areas bounded by the concentration curve and the diagonal have been shadowed.

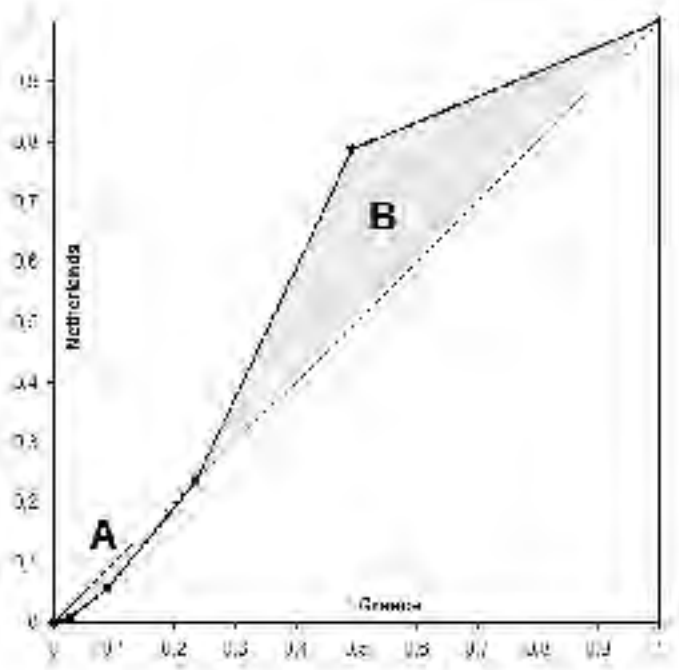

Fig. 5. Concentration curve of Dutch opinions relative to Greek opinions

As the expectation $E(X)$ is equal to the area above the concentration curve and the area above (or below) the diagonal is equal to $\frac{1}{2}$,

$$ar = 2\left(E(X) - \frac{1}{2}\right) = 2\left(A + \frac{1}{2} - B - \frac{1}{2}\right) = 2\,(A - B)\,. \tag{4}$$

Thus the concentration index is equal to the difference of two areas bounded by the concentration curve and the diagonal: the one under the diagonal and the one above it.

As we noticed in section 3.3.2, the maximal concentration curve is convex and lies totally under the diagonal. Thus a concentration index for the maximal concentration curve is

$$ar_{max} = 2\,(A - B) = 2A \geqslant 0. \tag{5}$$

We call it a **maximal concentration index**. It is a measure of dissimilarity of compared objects.

3.3.5 Grade Correspondence Analysis[23]

For each pair of objects (a pair of rows or a pair of columns of a table) we can find such a permutation of attributes that a concentration index reaches its maximum ($ar = ar_{max}$). However, usually it is not possible to do it for the whole table; if we maximize ar for one pair of objects (by changing the order of rows or the order of columns), at the same time we reduce the value of ar for other pairs. Nevertheless, it is possible to achieve a compromise: to find such an order of columns and rows of a table that all concentration indexes are close (though not necessarily equal) to their maximal values. This is done by Grade Correspondence Analysis (GCA) – an algorithm which constitutes the core of the GDA.

GCA's input is a probability table P_{mk} (m rows $\times$ k columns). Its parameter is a number of iterations n. Its output is a list of permutations of table P_{mk} ($List_1, List_2, \ldots, List_n$) being the local maxima of optimisation criteria: either Spearman's ρ^*, or Kendall's τ.

GCA based on Spearman's ρ^* aims to maximize the value of

$$\rho^* = 3 \sum_{i=1}^{m} \sum_{s=1}^{k} \left(p_{is}\,(2S_{row}(i) - 1)\,(2S_{col}(s) - 1)\right), \tag{6}$$

where

$$2S_{row}(i) = \left(\sum_{j=1}^{i-1} p_{j+}\right) + \frac{1}{2} p_{i+}, \tag{7}$$

$$2S_{col}(s) = \left(\sum_{t=1}^{s-1} p_{+t}\right) + \frac{1}{2} p_{+s}, \tag{8}$$

[23] See [17] and [19].

$$p_{j+} = \sum_{s=1}^{k} p_{js} \quad \text{is sum of } j\text{-th row}, \tag{9}$$

$$p_{+t} = \sum_{t=1}^{m} p_{ts} \quad \text{is sum of } t\text{-th column}, \tag{10}$$

p_{ij} is a value in i-th row and j-th column of a table, m is a number of table's rows, and k is a number of table's columns.

GCA based on Spearman's ρ^* permutes rows and columns of a table according to so called **grade regression** which may be described by formulas:

$$r_{col}(s) = \frac{\sum_{i=1}^{m} (p_{is} S_{row}(i))}{p_{+s}} \quad \text{for columns}, \tag{11}$$

$$r_{row}(i) = \frac{\sum_{s=1}^{k} (p_{is} S_{col}(s))}{p_{i+}} \quad \text{for rows}. \tag{12}$$

In each iteration of GCA algorithm the value of ρ^* increases.[24] As a number of permutations of rows and columns is finite (equal to $m! \times k!$), the algorithm must stop.

GCA based of Kendall's τ is not as straightforward as the previous method. It does not use the grade regression, but it sorts rows and columns using any algorithm comparing only adjacent rows and columns, e.g. bubble sorting. In each iteration it calculates concentration indexes for adjacent rows ($ar((i+1) : i; row)$) and columns ($ar((s+1) : s; col)$). If the result is negative, the rows or columns are swapped.

In the following text we limit ourselves to GCA based on Spearman's ρ^*.

3.3.6 Overrepresentation Index[25]

Table 3 presents so called **proportional distribution**. Its main characteristics is that for each p_{ij}

$$p_{ij} = p_{i+} \times p_{+j}, \tag{13}$$

where p_{i+} and p_{+j} are respectively: the sum of i-th row and the sum of j-th column.

[24] Ciok et al. have proved it in [6].

[25] See [21].

Table 3. Proportional distribution

	j = 1	j = 2	j = 3	j = 4	Sum p_{i+}
i = 1	0.12	0.10	0.14	0.04	0.40
i = 2	0.06	0.05	0.07	0.02	0.20
i = 3	0.12	0.10	0.14	0.04	0.40
Sum p_{+j}	0.30	0.25	0.35	0.10	1.00

Overrepresentation index is a ratio

$$c_{ij} = \frac{p_{ij}}{p_{i+} \times p_{+j}}. \tag{14}$$

Becase for any proportional distribution there is a relationship (13), $c_{ij} = 1$ for every i and each j. Table 4 presents a non-proportional distribution.

Table 4. Non-proportional distribution

	j = 1	j = 2	j = 3	j = 4	Sum p_{i+}
i = 1	0.25	0.05	0.04	0.06	0.40
i = 2	0.10	0.20	0.03	0.02	0.35
i = 3	0.10	0.05	0.03	0.07	0.25
Sum p_{+j}	0.45	0.30	0.10	0.15	1.00

Table 5 presents overrepresentation indexes for the above data. Note: Values in the table have been rounded; overrepresentation index for $i = 3$ and $j = 2$ is precisely $\frac{2}{3}$.

Table 5. Overrepresentation indexes for data from Table 4

	j = 1	j = 2	j = 3	j = 4
i = 1	1.39	0.42	1.00	1.00
i = 2	0.63	1.90	0.86	0.38
i = 3	0.89	0.67	1.20	1.87

3.3.7 Overrepresentation Map[26]

Overrepresentation map gives a quick insight into the inner structure of a

[26] See [21].

bivariate distribution. It is composed of rows which have heights proportional to marginal distributions p_{i+} and columns which have widths proportional to marginal distributions p_{+j}. Its cells are shadowed adequately to the values of overrepresentation indexes as shown in Table 6.

Table 6. Colour code for overrepresentation maps

	$c_{ij} \leqslant$ 2/3
	2/3 $< c_{ij} \leqslant$ 0.99
	0.99 $< c_{ij} \leqslant$ 1 / 0.99
	1 / 0.99 $< c_{ij} \leqslant$ 3/2
	3/2 $< c_{ij}$

Table 7. Overrepresentation map for values from Table 4

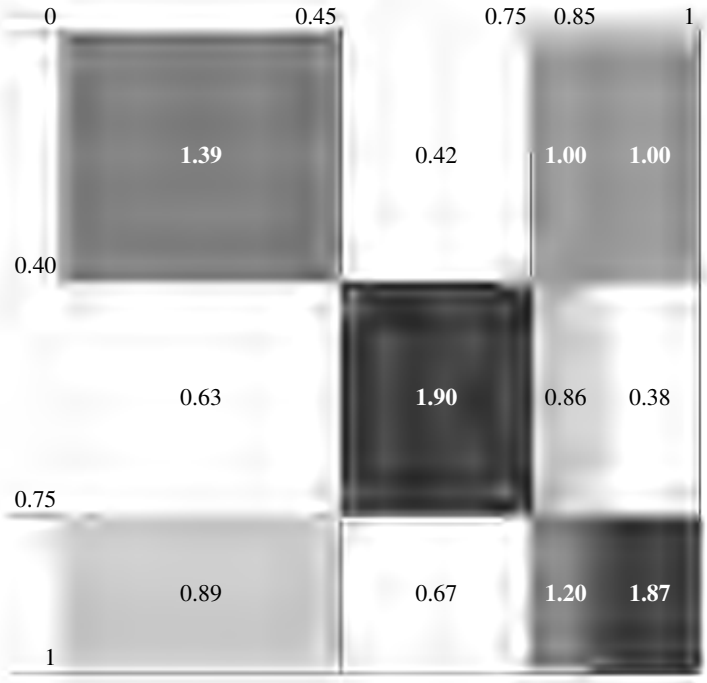

Table 7 presents an overrepresentation map for values from Table 4. Pay attention to the colour of the cell in row 3 and column 2 (overrepresentation

index is $\frac{2}{3}$). For any proportional distribution an overrepresentation map would be uniformly grey (because $c_{ij} = 1$ for every i and every j).

Overrepresentation maps are one of main GCA's advantages. They show relations between objects in a concise, easy to understand form.

Fig. 6 presents an overrepresentation map for the results of the survey on self-perceived health status (data from Table 1). This map has been prepared in GradeStat. Its columns have widths proportional to marginal distributions for particular answers ("Very bad", "Bad", "Fair", etc.). All rows have equal heights because a sum of values in a row is the same for each country: 100%.

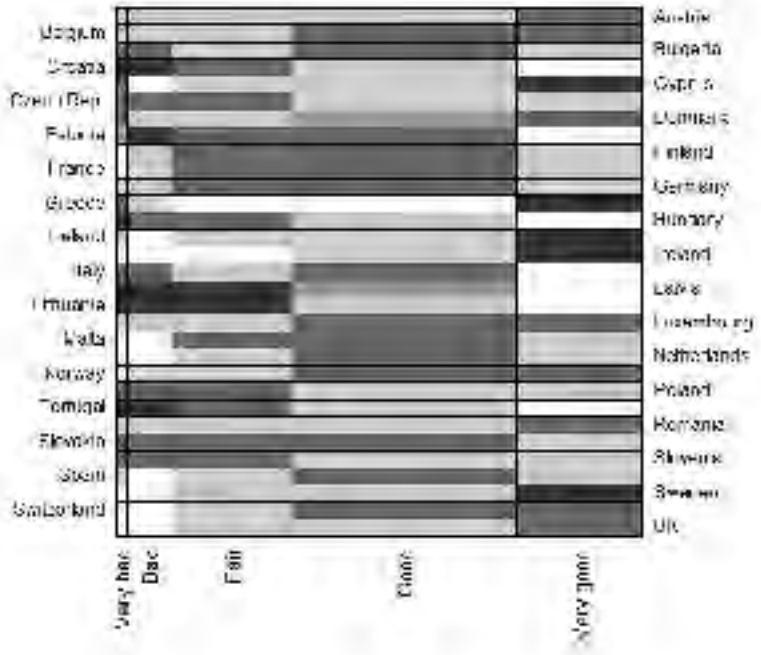

Fig. 6. Overrepresentation map for self-perceived health survey

3.3.8 Grade Correspondence Analysis

The GCA algorithm tends to such a permutation of rows and columns of a table of raw data that the value of Spearman's rank correlation coefficient ρ^* is as close to maximum as possible. After the GCA we receive an overrepresentation map which is more regular; its segments with the same degree of grey form tight areas, and the darkest ones are close to

the diagonal. Fig. 7 presents an overrepresentation map for the results of self-perceived health status survey after the GCA.

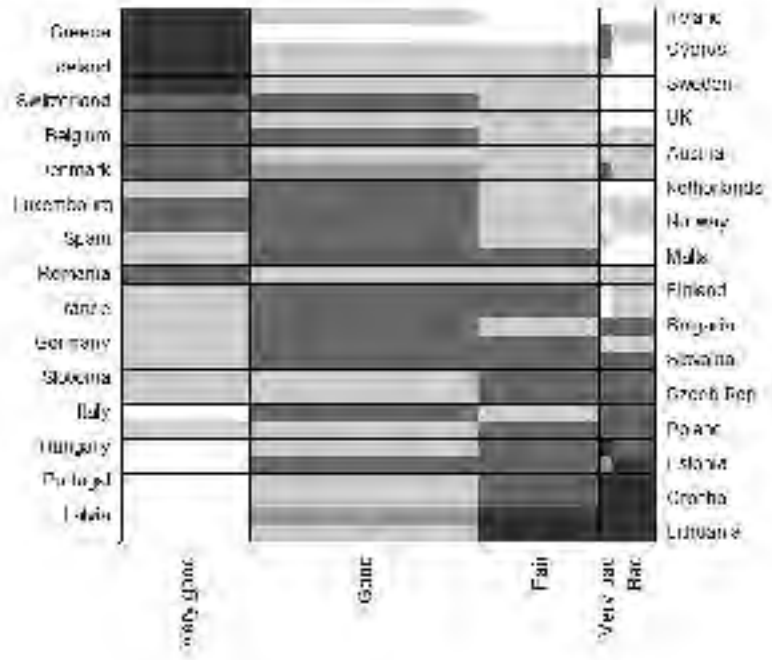

Fig. 7. Overrepresentation map for self-perceived health survey after the GCA

3.3.9 Cluster Analysis

As we noticed in previous section, an overrepresentation map after the GCA is more regular. Its rows and columns which are more similar become adjacent. Such a structure of the map permits us to perform a cluster analysis – both for rows and columns. So we can find groups of similar objects as well as groups of interrelated attributes.

As we said in Section 3.2, there is no clear rule on how many clusters to search for. Fig. 8 presents an overrepresentation map for the survey on self-perceived health status with 4 clusters for rows and 3 clusters for columns.

Clusters for columns seem to be natural: one is composed of the attribute "Very good", one contains the attribute "Good", and the last one contains three remaining attributes: "Fair", "Very bad" and "Bad". We could name these clusters "Excellent health", "Good health", and "Poor health".

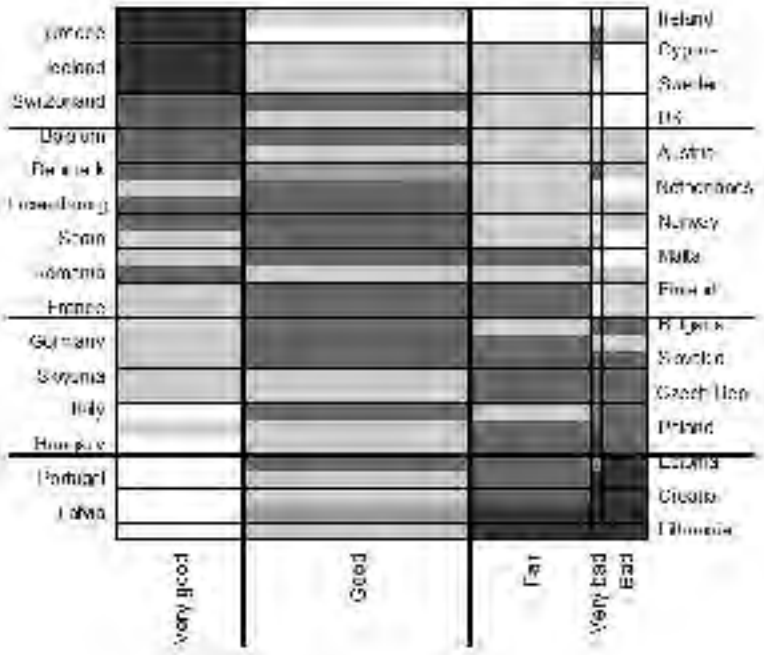

Fig. 8. Overrepresentation map for self-perceived health survey – cluster analysis

As for the clusters for rows, we could name them: "Very poor health" (Lithuania, Latvia, Croatia, Portugal, Estonia), "Rather poor health" (Hungary, Poland, Italy, Czech Republic, Slovenia, Slovakia, Germany, Bulgaria), "Rather good health" (France, Finland, Romania, Malta, Spain, Norway, Luxembourg, Netherlands, Denmark, Austria, Belgium), and "Very good health" (United Kingdom, Switzerland, Sweden, Iceland, Cyprus, Greece, Ireland). We should remember that we have just results of a survey on self-perception of health status. We do not know whether Lithuania, Latvia, Croatia, Portugal and Estonia got to the "Very poor health" cluster because of real poor condition of health of their inhabitants or because of their high expectations regarding health. Similarly, we do not know whether inhabitants of the United Kingdom, Switzerland, Sweden, Iceland, Cyprus, Greece and Ireland are really healthy or they just do not bother with their health.

It is worth noting that Greece and Netherlands (countries we compared in Section 3.3.1) have been separated in two different clusters.

3.3.10 Outliers Detection

As we said in Section 3.3.5, the aim of GCA is such a permutation of rows and columns of a table of raw data, that all concentration indexes ar (for rows and columns) are as near maximum (ar_{max}) as possible. In each iteration GCA tries to minimize an average of differences $ar - ar_{max}$ which we denote $AvgDistA_{Row}$ for rows and $AvgDistA_{Col}$ for columns. As proved in [19], they are given by formulas:

$$\begin{aligned} &AvgDistA_{Row}(i;P) = \\ &\sum_{s=1}^{i-1} \frac{ar_{max}(i:s;row(P)) - ar(i:s;row(P))}{(m-1)\sqrt{2}} + \\ &\sum_{s=i+1}^{m} \frac{ar_{max}(s:i;row(P)) - ar(s:i;row(P))}{(m-1)\sqrt{2}}, \\ &i = 1,...,m; \end{aligned} \tag{15}$$

$$\begin{aligned} &AvgDistA_{Col}(j;P) = \\ &\sum_{t=1}^{j} \frac{ar_{max}(j:t;col(P)) - ar(j:t;col(P))}{(k-1)\sqrt{2}} + \\ &\sum_{t=j+1}^{k} \frac{ar_{max}(t:j;col(P)) - ar(t:j;col(P))}{(k-1)\sqrt{2}}, \\ &j = 1,...,k. \end{aligned} \tag{16}$$

Elements of a table (rows or columns) which have an average $AvgDistA$ much greater than others may be considered as outliers, i.e. elements that deviate from the general trend. Fig. 9 presents values of $AvgDistA$ for rows, i.e. for countries.[27] We may observe that the most optimistic in assessing their health are Greeks, and – to a lesser extent – Cypriots and the the Dutch.

How important is the outliers detection can be understood when we look at the GCA-processed overrepresentation maps in (Fig. 7 and Fig. 8). There are two countries "at the edge of the map": Ireland and Lithuania. We could intuitively consider them as outliers – which, in fact, they are not.

[27] Because of limited space only 11 countries are shown.

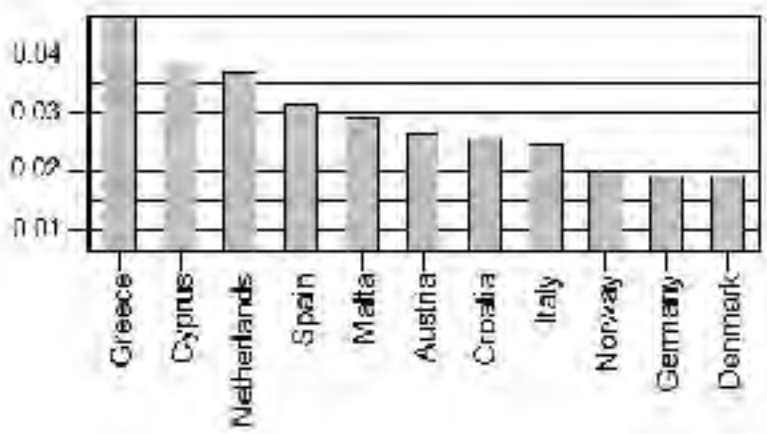

Fig. 9. $AvgDistA_{Row}$ – outliers detection for countries

Fig. 10 presents values of $AvgDistA$ for columns, i.e. for responders' answers. We can see that these are "extremities" (assessments "Very good" and "Very bad") which have the biggest influence on the results of the survey.

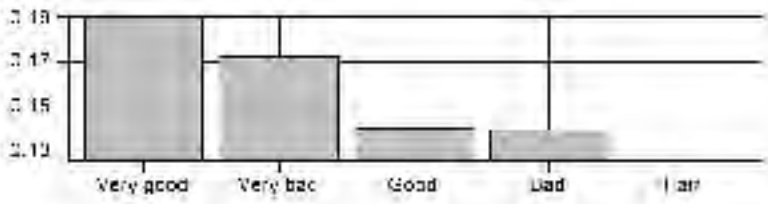

Fig. 10. $AvgDistA_{Col}$ – outliers detection for answers

4 EUROPEAN AGRICULTURE. GDA APPLIED

In Section 2 we said that every assessment is more or less subjective. Grade Data Analysis is not an exception. As J. Koronacki observed, GDA is *a tool, which allows, in an orderly manner – although dependent on the experience and subjective preferences* [of analysts], *not only on objective indicators – to extract from data different trends that have regular distributions, and which, at the same time, significantly differ among themselves.*[28]

[28] See [17], p. 5.

Thus we do not try to make an objective analysis of EU agriculture; we only wish to present it *in an orderly manner, although dependent on our experience and subjective preferences*. We must also emphasize that the purpose of this paper is not a comprehensive analysis of agriculture in the European Union, but only a presentation of a practical application of the GDA on an example of EU agriculture.

As mentioned in Section 2.3, our assessment will be composed of

- choosing characteristics (attributes) of the objects,
- data pre-processing,
- data processing,
- analysis of the results.

The following sections cover these steps in details.

4.1 Choosing Characteristics of the Objects

Objects of our assessment are EU-member countries. Because we want to assess agriculture in each of them, we should first decide what values may be considered as measures of "agricultural condition". Those values will be attributes of the objects. In Section 2.2 we said that many researchers, to assess the condition of agriculture, measure its **productivity**, i.e. ratio(s) of outputs to inputs. As in agriculture there are three basic inputs: **land, labour, and capital**, we will analyse productivity of these three elements.[29]

It should be noted that productivity of land, productivity of labour, and productivity of capital are interrelated. Yields depend on soil quality, labour effort, as well as on many capital-depending factors (e.g. machines usage). Thus it is not possible to present a list of "clear" measures of productivity of land; each measure will contain some labour- and capital-related elements. There are no "clear" measures of productivity of labour or productivity of capital, either. Moreover, each of three basic "inputs of agriculture": land, labour, and capital may be – to some extent – a substitute of the other two, e.g. poor soil quality may be compensated by more labour effort or by more money spent of fertilizers. This substitutability is at the core of Common Agricultural Policy (CAP) of the European Union;

[29] An interesting analysis of productivity of these three elements may be found in [4]. In [8] authors go farther: they analyse economic, social, environmental, demographic, and territorial challenges for agriculture. Land, labour and capital, along with **competitiveness** and **weak bargaining power in the food chain**, compose the first group of challenges.

agricultural subsidies (i.e. capital) are a tool to shape the condition of agriculture in member countries.[30]

We will use for our assessment the following measures, keeping in mind the imperfection of our choice.[31] For productivity of land (see Table 8):

- agricultural output at producer prices[32] per utilised agricultural area [1000 PPS[33] per ha] – code: Out-Area,
- yields of cereals for the production of grain (including rice and seed) [100 kg/ha] – code: Cereals,
- average production of milk and milk products (in milk equivalents) per dairy cow[34] [kg] – code: Milk-Cow,
- average fat contents in cow's milk [% of product weight] – code: Milk-Fat,
- average protein contents in cow's milk [% of product weight] – code: Milk-Prot.

For productivity of labour (see Table 9):

- crop output at producer prices per labour [1000 PPS per AWU[35]] – code: Crop,
- animal output at producer prices per labour [1000 PPS per AWU] – code: Animal,
- agricultural services output at producer prices per labour [1000 PPS per AWU] – code: Service,
- utilised agricultural area per labour [ha per AWU] – code: Area-Work.

For productivity of capital (see Table 9):

- share of inputs[36] in production [%] – code: Inputs.

To be able to relate these measures to the "general" characteristics of each country, we will also include in our study (see Table 10):

[30] See [18] for more details.

[31] More in-depth discussions about measuring productivity in agriculture may be found in [15], [16], [24], [31], [32], [34], and [35].

[32] Procuder price, as defined by the World Bank, is the amount receivable by the producer inclusive of taxes on products except deductible value added tax and exclusive of subsidies on products.

[33] Purchasing Power Standard. An artificial currency unit which eliminates differences in price levels between countries. Theoretically, one PPS can buy the same amount of goods and services in each country.

[34] Data for 2011. Newer data not available.

[35] Annual Work Unit. The work performed by one person who is occupied on a full-time basis, during one year.

[36] Seeds and reproductive material, energy and lubricants, fertilizers and soil improvers, crop protection products and pesticides, veterinary expenditure, animal feed, maintenance of machinery, maintenance of buildings, agricultural services, other products and services.

Table 8. Measures of productivity of land

Country	Out-Area	Cereals	Milk-Cow	Milk-Fat	Milk-Prot
Austria	2.16	60.10	6 437.52	4.20	3.39
Belgium	5.59	88.10	7 153.39	4.10	3.40
Bulgaria	2.00	36.70	3 177.37	3.68	3.28
Cyprus	6.61	18.60	7 332.96	3.70	3.38
Czech Republic	1.93	45.40	6 992.56	3.83	3.38
Denmark	3.25	63.30	8 408.16	4.28	3.42
Estonia	1.26	34.10	7 392.82	4.00	3.38
Finland	1.52	35.50	8 707.13	4.27	3.48
France	2.38	73.00	7 095.68	3.98	3.41
Germany	3.11	69.70	7 645.91	4.13	3.41
Greece	2.11	43.40	6 672.65	3.95	3.31
Hungary	2.68	37.60	7 026.81	3.63	3.21
Ireland	1.31	67.40	5 456.14	3.94	3.36
Italy	3.63	52.80	6 151.37	3.78	3.38
Latvia	0.72	37.70	5 503.58	4.16	3.33
Lithuania	1.59	40.20	5 407.90	4.15	3.27
Luxembourg	2.44	55.10	7 558.31	4.16	3.39
Malta	13.87	46.94	6 565.96	3.38	3.23
Netherlands	12.52	85.70	8 023.14	4.40	3.53
Poland	2.66	37.10	4 935.94	4.00	3.22
Portugal	2.08	40.90	7 218.20	3.78	3.25
Romania	2.04	23.40	3 400.64	3.81	3.26
Slovakia	1.77	38.30	5 806.44	3.78	3.37
Slovenia	2.97	57.80	5 268.15	4.15	3.37
Spain	1.87	28.40	7 308.10	3.62	3.26
Sweden	1.49	51.00	8 519.41	4.22	3.42
United Kingdom	1.47	62.10	7 393.18	4.07	3.26

Table 9. Measures of productivity of labour and of productivity of capital

Country	Crop	Animal	Service	Area-Work	Inputs
Austria	15.57	15.93	1.41	15.24	60.9
Belgium	57.38	70.18	0.69	22.94	70.3
Bulgaria	8.97	3.91	0.92	6.89	61.9
Cyprus	29.21	28.53	0.02	8.73	53.7
Czech Republic	24.20	14.97	1.04	20.79	71.9
Denmark	45.73	77.54	6.11	39.84	73.1
Estonia	21.22	18.79	1.87	33.24	61.0
Finland	13.11	16.18	0.71	19.77	67.4
France	52.90	30.39	4.60	36.87	59.3
Germany	38.92	36.46	2.58	25.09	68.6
Greece	14.67	6.42	0.91	10.41	53.1
Hungary	24.91	14.63	2.01	15.48	65.7
Ireland	20.21	51.38	3.84	57.76	75.0
Italy	26.78	17.77	5.31	13.72	47.4
Latvia	11.47	6.91	0.56	26.13	73.6
Lithuania	24.45	13.56	0.89	24.45	63.4
Luxembourg	40.22	35.53	0.87	31.46	71.4
Malta	11.66	17.20	0.00	2.08	54.9
Netherlands	51.91	44.12	11.93	8.63	67.3
Poland	10.01	9.30	0.42	7.40	61.6
Portugal	8.11	6.29	0.48	7.16	66.5
Romania	6.62	2.93	0.09	4.73	57.0
Slovakia	25.77	20.64	2.53	27.68	78.8
Slovenia	9.48	8.56	0.32	6.19	63.9
Spain	35.99	22.90	0.61	31.81	49.2
Sweden	23.28	21.05	2.48	31.52	73.4
United Kingdom	27.07	39.58	3.20	47.65	64.4

– GDP[37] at current prices per inhabitant [PPS] – code: GDP,
– share of agricultural GVA[38] in the GDP [%] – code: Agri-GDP,
– share of employment in agriculture in the total civilian working population [%] – code: Agri-Work,
– share of utilised agricultural area in the total area of the country [%] – code: Agri-Area,
– utilised agricultural area per holding [ha] – code: Area-Hold.

Table 10. General characteristics of countries

Country	GDP	Agri-GDP	Agri-Work	Agri-Area	Area-Hold
Austria	32 300	1.0	4.5	34.32	19.17
Belgium	29 500	0.6	1.3	44.48	31.69
Bulgaria	11 600	4.2	18.9	40.32	12.08
Cyprus	23 300	1.9	3.6	12.80	3.05
Czech Republic	20 100	0.9	3.3	44.17	152.38
Denmark	31 400	1.5	2.4	61.41	62.87
Estonia	17 300	2.1	4.7	20.80	47.98
Finland	28 600	0.9	4.6	6.77	35.87
France	27 200	1.6	2.8	50.70	53.94
Germany	30 500	0.6	1.6	46.77	55.84
Greece	20 200	2.8	12.2	39.23	7.16
Hungary	16 400	2.7	7.4	50.37	8.12
Ireland	32 600	1.1	4.7	71.02	35.68
Italy	25 100	1.6	3.8	42.67	7.93
Latvia	14 800	1.4	7.9	27.82	21.54
Lithuania	16 800	3.5	8.8	42.00	13.72
Luxembourg	66 000	0.3	1.1	50.70	59.60
Malta	21 800	0.8	3.2	36.23	0.91
Netherlands	32 600	1.4	2.5	50.12	25.89
Poland	16 200	2.4	12.6	46.20	9.59
Portugal	19 700	1.3	11.0	39.91	12.02
Romania	12 600	4.7	30.6	55.82	3.45
Slovakia	18 500	0.8	3.1	38.66	77.49
Slovenia	21 000	1.1	8.3	23.81	6.47
Spain	23 900	2.1	4.2	47.00	24.00
Sweden	31 600	0.5	2.1	6.85	43.13
United Kingdom	27 400	0.5	1.2	69.16	90.37

[37] Gross Domestic Product.
[38] Gross Value Added = net output + subsidies - taxes.

All data have been taken from Eurostat portal[39], FADN portal[40], and the European Commission portal[41] and they date from 2012. The analysis does not include Croatia which joined the EU in 2013.

4.2 Data Pre-processing

As we said in Section 2.3, data pre-processing usually means their weighting and/or normalizing.

4.2.1 Data Normalizing – First Step

In GDA, normalizing data is performed in two steps. In the first step each value in the table is divided by the sum of all values in the group to which this value belongs.

Each group may be composed of one or more columns. It is up to a researcher to decide whether to put several columns in the same group.[42] We may place in the same group the attributes that describe similar characteristics, which are measured in the same units, and which take values from the same interval. For instance we may (but we do not have to) put in a common group three measures of productivity of labour: Crop (crop output per labour), Animal (animal output per labour), and Service (agricultural services output per labour). They all describe agricultural output relative to labour effort, they are all measured in PPS per AWU, and they take values from the same interval. Moreover, a sum of these three attributes for a particular country has a concrete sense: it is the total agricultural output relative to labour effort for this country. But it would be a great mistake to put in this group the fourth measure of productivity of labour: Area-Work (utilised agricultural area per labour).

After the normalization, all values in the table belong to the interval [0; 1]. This prevents the analysis from being dominated by one (or some) of the attributes of objects. In our case these would be: Milk-Cow and GDP which have values much higher than all the other ones (see Table 8 and Table 10).

We will put Crop, Animal, and Service in one common group. All other attributes will belong to their individual groups.

[39] http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/themes, access: July 27th, 2014.

[40] Farm Accountancy Data Network: http://ec.europa.eu/agriculture/rica/, access: July 27th, 2014.

[41] http://ec.europa.eu/agriculture/statistics/agricultural/index_en.htm, access: July 27th, 2014.

[42] After the selection of attributes, it is the next source of subjectivity of the analysis.

4.2.2 Data Weighting

After the first step of the normalization, each value in the table is multiplied by the weight of its group. A weight of a group is a reflection of its "importance" in the whole analysis. By default, all groups have the same weights equal to 1, which means: they are equally important. Those weights may be changed if there is a reason for it.

Once again, it is a researcher's decision whether to differentiate the weights of attributes. Though each choice of weights is more or less subjective, it should always have a clear justification. It should be preceded by a thorough analysis of the problem being studied and of relationships between attributes. In our analysis, we will stay with all weights equal to 1.

NB. It should be noted what are the consequences of putting the three measures of productivity of labour (Crop, Animal, Service) in a common group. They have now a **joint weight** equal to 1. If each of them belonged to an individual group, each of them would have an **individual weight** equal to 1.

4.2.3 Data Normalizing – Second Step

Finally, each value in the table is divided by the sum of the whole table. Such a pre-processed table is the starting point for the data processing: the GCA and the cluster analysis.

4.3 Data Processing

Let us have a look at the overrepresentation map of our data (see Fig. 11). We may notice that widths of nearly all columns are almost equal – which means marginal distributions of attributes are similar. Columns corresponding to the measures of productivity of labour are exceptions because we have put them in the same group assigning to it a weight equal to 1.

Rows have heights proportional to the "participation in European agriculture" of particular countries. We can see that they differ, though differences are not big.

Colours of segments are very diverse (we would like to say that they resemble bird's-eye view of crop fields). This diversity of colours reflects the diversity of individual measures in particular countries.

When the GCA is done, we will be able to see the similarities and differences between countries as well as between their attributes.

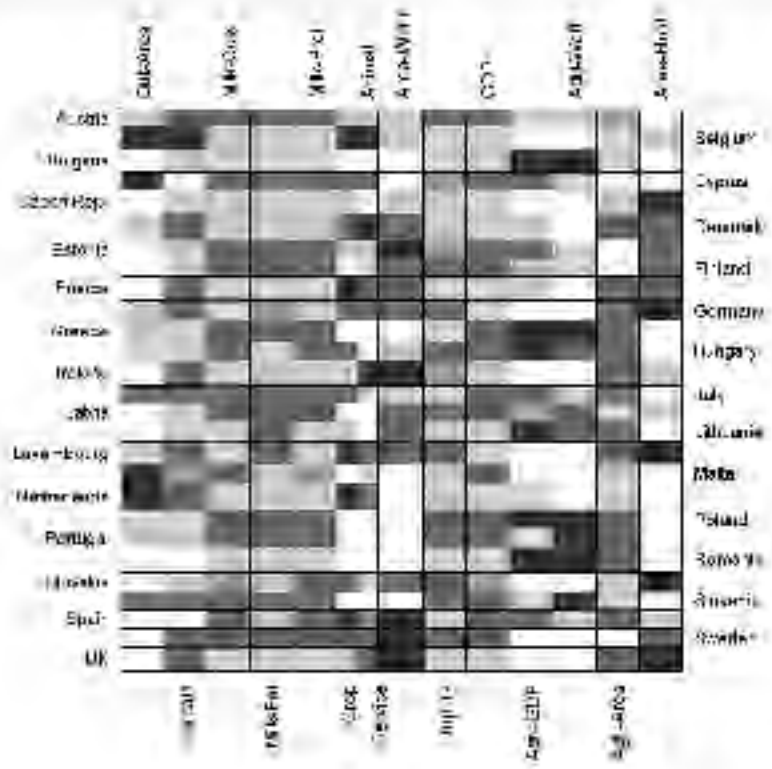

Fig. 11. EU agriculture – overrepresentation map

4.4 Analysis of the Results

4.4.1 GCA

Fig. 12 presents the overrepresentation map produced by GradeStat after the GCA procedure. Both: the sequence of countries and the sequence of attributes have changed. Segments with similar degree of grey form bigger areas, the darkest of them are arranged close to the diagonal.

We can see a clear rule according to which the GCA has changed the ordering of rows. When we analyse the map from top to bottom, we see declining utilised agricultural area per holding (Area-Hold) and growing share of employment in agriculture in the total civilian working population (Agri-Work). However, Malta and Cyprus are exceptions; though they have low values of Agri-Work, they have been put at the bottom of the map. These exceptions will become clear when we make a cluster analysis for countries.

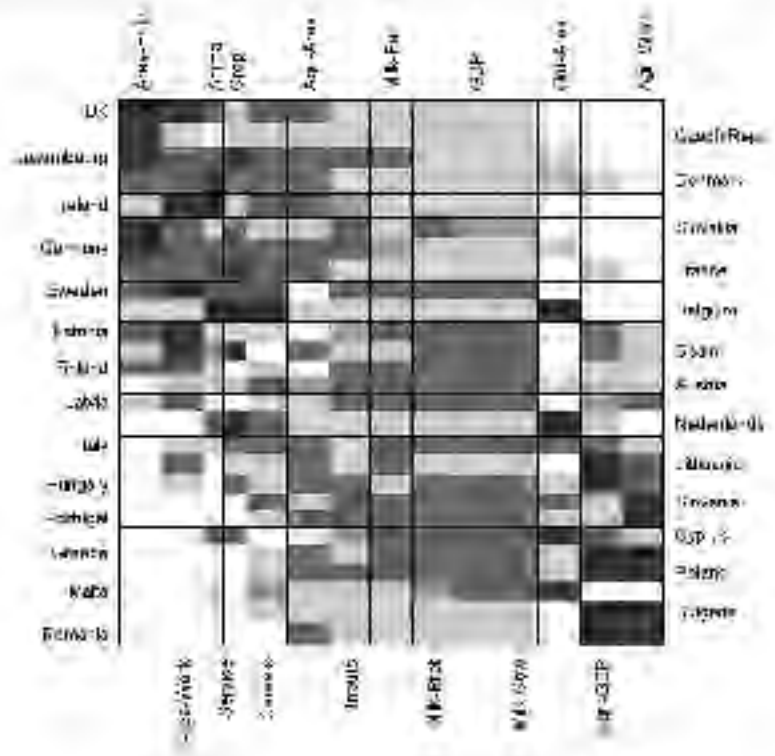

Fig. 12. EU agriculture – overrepresentation map after the GCA

There is also a rule according to which the GCA has reordered columns. Two leftmost ones represent the cultivated area: per holding (Area-Hold), and per labour (Area-Work). Next to them is a group of "hard" productivity measures: agricultural outputs per labour (Crop, Animal, and Service), and yields of cereals for the production of grain per hectare (Cereals).

On the right side of the map we may find columns representing the following attributes:

- share of employment in agriculture in the total civilian working population (Agri-Work),
- share of agricultural GVA in the GDP (Agri-GDP),
- agricultural output per utilised agricultural area (Out-Area).

The first two are not surprising: they measure how much "agricultural" is a particular country. As for the third one, we would rather expect it to be

put by the GCA on the left side of the map, together with three columns representing agricultural output per labour.

In the middle of the overrepresentation map the GCA has put:

- average fat contents in cow's milk (Milk-Fat),
- average protein contents in cow's milk (Milk-Prot),
- average production of milk and milk products per dairy cow (Milk-Cow),

i.e. attributes representing productivity of agricultural animals,

- share of utilised agricultural area in the total area of the country (Agri-Area),
- GDP per inhabitant (GDP),

i.e. attributes representing macroeconomic parameters of each country, and

- share of inputs in production (Inputs)

representing productivity of capital.

To understand why the GCA has reordered columns in such a way, we should look at the colours of their cells. These "at the edges" of the over-representation map contain all shades of grey, while those in the middle of the map are almost uniformly grey. Thus, the basis for the ordering of countries consists of the following attributes:

- farm size (cultivated area per holding or per labour),
- agricultural output per labour,
- yields of cereals per hectare,
- presence of agriculture in country's economy (percentage of working population employed in agriculture, share of agriculture the GDP),
- agricultural output per utilised area.

4.4.2 Cluster Analysis

Let us now turn to the cluster analysis. As we already said, there is no clear rule on how many clusters to search for. Fig. 13 presents five clusters for countries and three clusters for attributes.

We have the following clusters of countries:

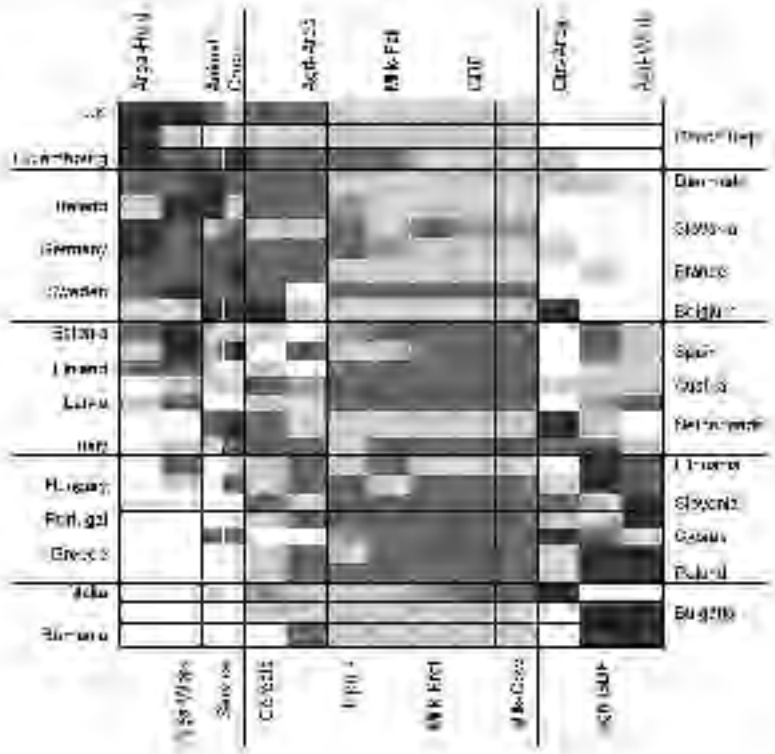

Fig. 13. EU agriculture – cluster analysis

– United Kingdom, Czech Republic, Luxembourg. Countries with large farms and a small share of agriculture in national economies. United Kingdom and Luxembourg have also a high per-worker agricultural output.
  Czech Republic, in spite of its more modest per-worker output, has been put in this cluster because of its extremely high overreprsentation in the "area per holding" attribute: 152.38 ha, the highest value in the whole European Union.
– Denmark, Ireland, Slovakia, Germany, France, Sweden, Belgium. Countries with large or big farms, a high per-worker agricultural output, and a small share of agriculture in national economies.
– Estonia, Spain, Finland, Austria, Latvia, Netherlands, Italy. Countries with big or middle-size farms, a good or average per-worker agricultural output, and an average share of agriculture in national economies.
– Lithuania, Hungary, Slovenia, Portugal, Cyprus, Greece, Poland. Countries with middle-size or small farms, poor per-worker agricultural out-

put (with the exception of Cyprus and Hungary), and a high share of agriculture in national economies.
Cyprus and Hungary, in spite of their good per-worker output, have been put in this cluster because of their high overrepresentation in the Agri-GDP attribute.

– Malta, Bulgaria, Romania. Countries with small farms, poor per-worker agricultural output, and a high share of agriculture in national economies (with the exception of Malta).
Malta, in spite of its low overrepresentation in Agri-GDP and Agri-Work attributes, has been put in this cluster because of its high over-representation in the Out-Area attribute, which is much higher than any other overrepresentation for this country.

Clusters for columns correspond with the division of attributes we described earlier:

– Left cluster: farm size and agricultural output per labour. It is surprising that this cluster does not contain yields of cereals per hectare which belong to the middle one. Yields of cereals are clustered with with crop output (but not with animal output) per labour if there are 4 clusters, but when there are 5 clusters these two attributes are separated again (see Fig. 14). That means yields of cereals are more correlated with attributes grouped in the middle cluster than with those in the left one.

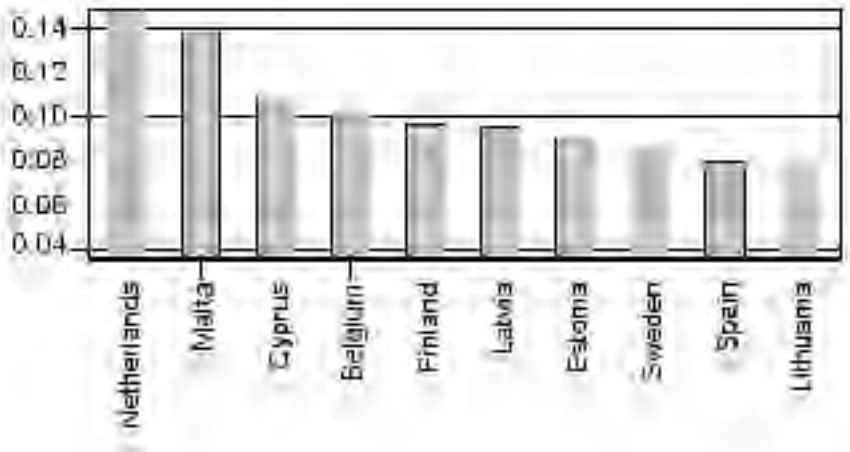

Fig. 14. EU agriculture – 4 and 5 clusters for attributes

– Right cluster: presence of agriculture in country's economy, and agricultural output per utilised area.
– Middle cluster: productivity of land (including yields of cereals per hectare) and of animals, and macroeconomic parameters of a country.

4.4.3 Outliers Detection

Outliers detection for countries may be surprising (see Fig. 15). There are four outliers: Netherlands, Malta, Cyprus, and Belgium (though the latter two to a lesser degree). None of them is a country "on the edge" of the overrepresentation map. They are outliers because of their high overrepresentation in agricultural output per utilised agricultural area (Out-Area). A high value of this attribute is an advantage; it demonstrates good use of agricultural land. So our intuitive pejorative perception of the term "outlier" in this case is misleading.

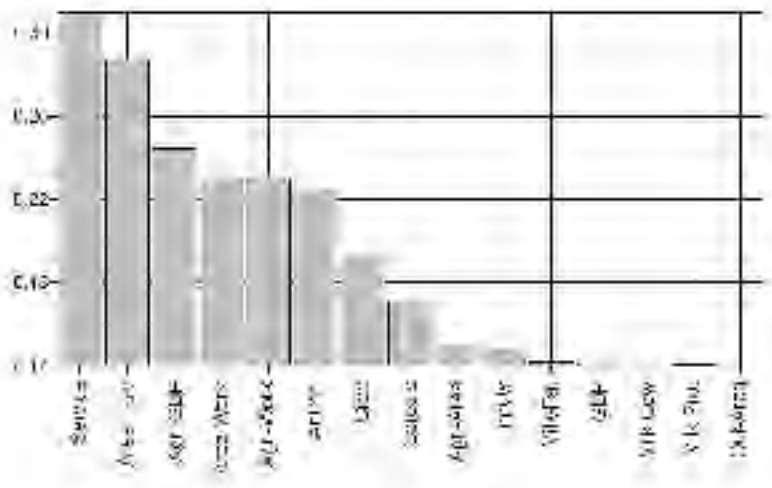

Fig. 15. EU agriculture – outliers detection for countries

Outliers detection for attributes also brings surprises (see Fig. 16). The first outlier is "per-worker" output of agricultural services (Service) – an attribute barely visible on the overrepresentation map. We may explain it when we look at the Table 9. No other attribute is as "internally diversified" as this one.

The second outlier is utilised agricultural area per holding (Area-Hold) – an attribute "on the edge" of the overrepresentation map, being the basis to distinguish clusters of countries.

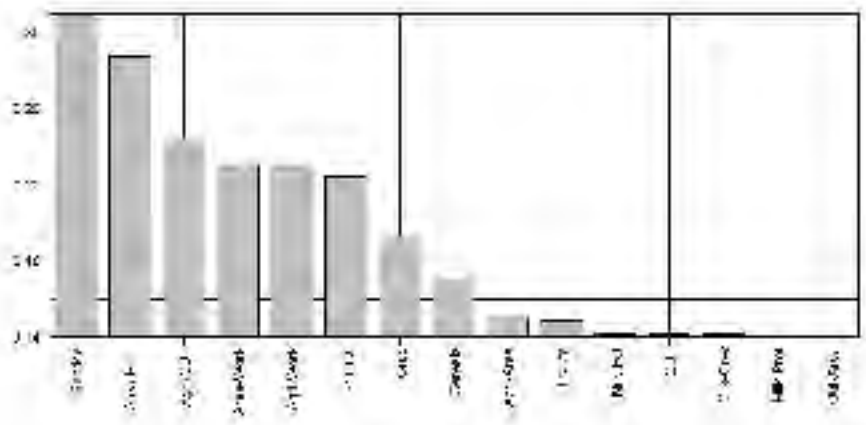

Fig. 16. EU agriculture – outliers detection for attributes

4.4.4 Outliers Rejection. New GCA

Outliers are elements that deviate from the general trend observed in the population. Thus, they are elements which introduce some chaos to the analysis and deform its results. To better understand the interrelationships between objects (as well as between their attributes), we should reject outliers and then perform a re-analysis of the data.

First we will do this for attributes. The two most outlying ones are: agricultural services output per labour (Service) and utilised agricultural area per holding (Area-Hold). Rejecting only the first one (Service) has no impact on the results; after the GCA there is still the same order of rows and columns. The conclusion is simple: there is no sense to include this attribute in the analysis. When we reject the second outlying attribute (Area-Hold), and then we perform a new GCA, we obtain an overrepresentation map which differs much from that analysed before – see Fig. 17.

This new map is much smoother than the previous one, its organization is more logical, its areas with the same intensity of grey are more compact and are arranged nearer the diagonal. Clusters have been reorganised on the basis of a more comprehensible criteria. Since the volume of this paper is limited, we will not go deeply into details in the analysis of this new map; however, let us remark three details:

– Romania, Bulgaria, and Greece have the highest overrepresentation in two attributes: the share of employment in agriculture in total civilian working population (Agri-Work) and the share of agricultural GVA in the GDP (Agri-GDP). Previously, Romania and Bulgaria have been

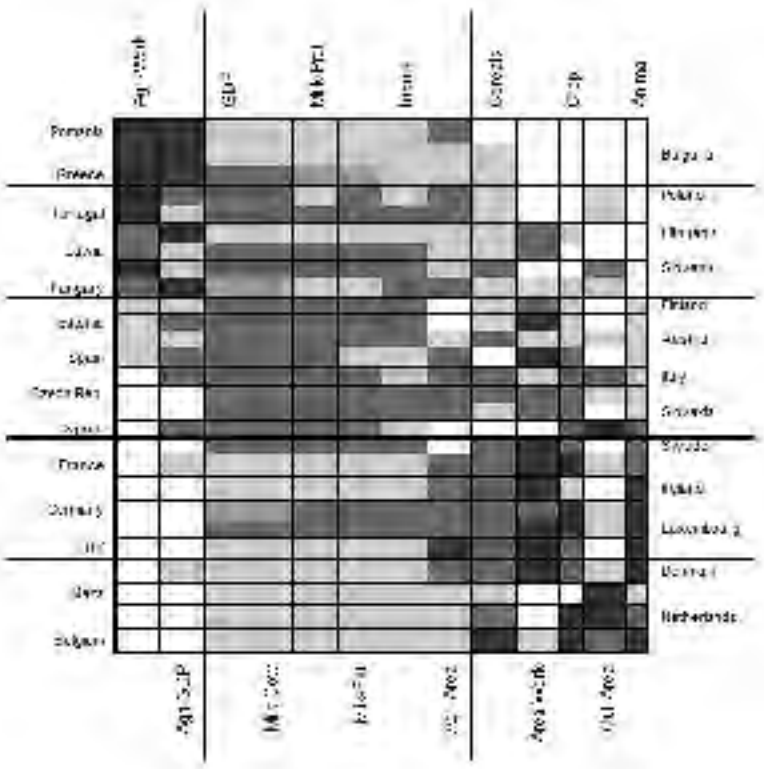

Fig. 17. EU agriculture – overrepresentation map (with clusters) after the GCA without two outlying attributes

combined into one common cluster with Malta, while Greece has found itself in a different cluster. Now Romania, Bulgaria, and Greece – the most "agricultural" countries in the whole European Union – are grouped in one (almost homogeneous) cluster, and Malta has been moved to the opposite edge of the table.

– Belgium and Netherlands are adjacent in the table, which is natural as they are very "similar". The GCA carried out previously has separated them in two different clusters.
– Clusters of attributes are very natural. Yields of cereals (Cereals) are clustered with crop and animal outputs per labour (Crop and Animal), and with agricultural output per utilised area (Out-Area). Previously, these attributes have been put in three clusters.

Now let us remove from the analysis two outlying countries: Netherlands and Malta. Fig. 18 presents the new overrepresentation map.

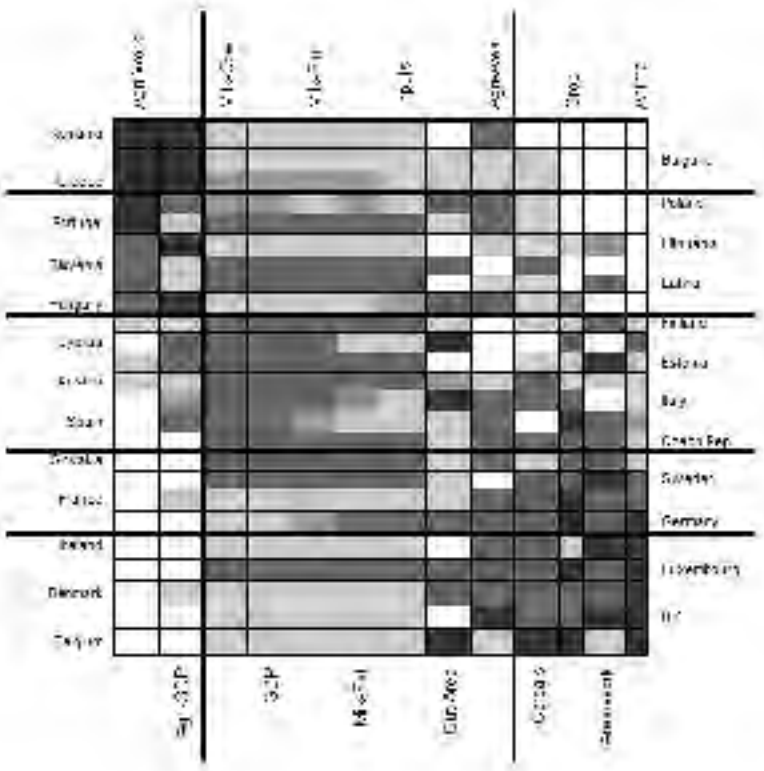

Fig. 18. EU agriculture – overrepresentation map (with clusters) after the GCA without two outlying attributes and two outlying countries

Compared to the previous map, there are only some minor changes:

- Ireland, Luxembourg, Slovakia, and UK have been moved to the neighbouring clusters.
- agricultural output per utilised area (Out-Area) has been moved to the middle cluster of attributes,
- GDP per inhabitant (GDP) and average production of milk per diary cow (Milk-Cow) have been swapped.

As we can see, after having rejected two outlying countries, we receive nearly the same information, but for a smaller group of countries.

5 CONCLUSION

Agriculture plays a vital role in each country's economy. Therefore, it is natural that governments, international communities as well as many re-

searchers try to assess its condition. Although there are a large number of publications on this subject, very few of them use data mining and cluster analysis. This paper attempts to fill this gap.

Grade Data Analysis is a universal method of acquiring useful information from large data sets. It can be used in various fields. It allows for a uniform analysis of different types of data measured on different scales and in different units. It is a pity, therefore, that this tool, being developed for many years, is still little known.

While the GDA is a great tool for data mining, we should be aware of its limitations. Although the GDA can discover data dependencies which researchers are not able to see, it will never replace them in drawing conclusions. It is used to organize data, to group objects being investigated, to "find order in chaos". The result of its application is the creation of a series of questions: about the causes of such an organisation of the population, and about the nature of interconnectedness among objects and among their attributes. Here the role of the GDA ends; replies on such questions must be given by researchers.

We hope that this paper has encouraged readers to take an interest in the GDA and in its practical application: GradeStat software.

References

1. Abdullah A., Brobst S., Pervaiz I., Umer M., Nisar A. (2006) Learning Dynamics of Pesticide Abuse through Data Mining, *Journal of Research and Practice in Information Technology*, **3**, Vol. 38, 229-249.
2. Aldenderfer M.S., Blashfield R.K. (1984) Cluster Analysis, *Quantitative Applications in the Social Sciences*, **44**, Sage University Papers, Newbury Park.
3. Baer-Nawrocka A., Markiewicz N. (2013) Relacje między czynnikami produkcji a efektywność wytwarzania w rolnictwie Unii Europejskiej, *Journal of Agrobusiness and Rural Development*, **3 (29)**, 5-16.
4. Blaas G. (2004) Productivity of factors in the enlarged EU, *Agricultural Economics – Czech*, **11 (2004)**, Vol. 50, 509-513.
5. Ciok A. (2004) *On the number of clusters – a grade approach*, Instytut Podstaw Informatyki PAN, Warszawa.
6. Ciok A., Kowalczyk T., Pleszczynska E., Szczesny W. (1995) Algorithms of grade correspondence – cluster analysis, *The Collected Papers of Theoretical and Applied Computer Science*, **1-4**, Vol. 6, 3-20.
7. Darku A.B., Malla S., Tran K.C. (2012) Sources and Measurement of Agricultural Productivity and Efficiency in Canadian Provinces: Crops and Livestock, http://ag-innovation.usask.ca/cairn_briefs/publications for download/S Malla paper1.pdf (access: June 27th, 2014), 1-37.
8. Davidova S., Thomson K. (2014) *Family Farming in Europe: Challenges and Prospects. In-depth Analysis*. Document requested by the European Parliament's Committee on Agriculture and Rural Development.

9. Dharmasiri L.M. (2011) Measuring Agricultural Productivity Using the Average Productivity Index (API), *Sri Lanka Journal of Advanced Social Studies*, **2**, Vol. 1, 25-44.
10. Fayyad U., Piatetsky-Shapiro G., Smyth P. (1997) From Data Mining to Knowledge Discovery in Databases, in *Advances in knowledge discovery and data mining*, American Association for Artificial Intelligence, Menlo Park, 1-34.
11. Ghodke B.D. (2009) Determination of Agricultural Productivity in Daund Tahasil of Pune District, *International Research Journal*, **6**, Vol. II, 857-858.
12. Gollin D., Lagakos D., Waugh M.E. (2014) The Agricultural Productivity Gap, *The Quarterly Journal of Economics*, **2**, Vol. 129, 939-993.
13. Gollin D., Lagakos D., Waugh M.E. (2012) The Agricultural Productivity Gap in Developing Countries, http://www.aeaweb.org/aea/2013conference/program/retrieve.php?pdfid=238 (access: June 27th, 2014), 1-51.
14. Grabowska G., Wiech M. (2009) Grade analysis of data from the European Economic Survey 2005 on Economic Climate in Polish Servicing Sector, *Control and Cybernetics*, **3**, Vol. 28, 783-810.
15. Grotkiewicz K., Michałek R. (2009) Postęp naukowo-techniczny a wydajność ziemi i pracy w rolnictwie, *Inżynieria Rolnicza*, **6 (115)**, 109-116.
16. Jarka S. (2013) *Ekonomiczna i społeczna wydajność pracy w przedsiębiorstwach rolniczych*, IX Kongres Ekonomistów Polskich, Warszawa.
17. Jarochowska E., Grzegorek M., Hirny J., Maryja O., Wiech M. (2005) *Analiza danych medycznych i demograficznych przy użyciu programu GradeStat*, Instytut Podstaw Informatyki PAN oraz Instytut Pomnik – Centrum Zdrowia Dziecka, Warszawa.
18. Knieć W. (2012) *Wspólna Polityka Rolna Unii Europejskiej a zrównoważony rozwój obszarów wiejskich Polski. Analiza socjologiczna*, Wydawnictwo Naukowe UMK, Toruń.
19. Kowalczyk T., Pleszczyńska E., Rulad F. (2004) Grade Models and Methods for Data Analysis with Applications for the Analysis of Data Populations, *Studies in Fuzziness and Soft Computing*, **151**, Springer Verlag, Berlin – Heidelberg – New York.
20. Korzeniewski J. (2012) *Metody selekcji zmiennych w analizie skupień. Nowe procedury*, Wydawnicwto Naukowe Uniwersytetu Łódzkiego, Łódź.
21. Lenkiewicz St. (2012) Gradacyjna analiza danych – idea i przykład zastosowania, *Współczesne Problemy Zarządzania*, 1/2012, 63-98.
22. Matuszczak A. (2010) Alokacja czynników wytwórczych a wyniki działalności rolniczej w regionach rolnych UE-25. Ocena taksonomiczna, *Zeszyty Naukowe SGGW w Warszawie – Problemy Rolnictwa Światowego*, **10** (25), 71-79.
23. Melfou K., Theocharopoulos A., Papanagiotou E. (2007) Total factor productivity and sustainable agricultural development, *Economics and Rural Development*, **1**, Vol. 3, 32-38.
24. Michałek R., Grotkiewicz K., Peszek A. (2009) Wydajność ziemi i pracy w wybranych krajach Unii Europejskiej, *Inżynieria Rolnicza*, **1** (110), 199-205.
25. Mroczek R. (2007) Konkurencyjność produktów polskiego rolnictwa po wejściu do UE, *Zeszyty Naukowe SGGW w Warszawie – Problemy Rolnictwa Światowego*, **2** (17), 267-276.
26. Mucherino A., Papajorgji P.J., Pardalos P. (2009) *Data Mining in Agriculture*, Springer Dordrecht, Heidelberg – London – New York.
27. Nowak A. (2013) Produktywność rolnictwa polskiego w kontekście jego konkurencyjności, *Folia Pomeranae Universitatis Technologiae Stetinensis – Oeconomica*, **299** (70), 159-168.
28. Nowak A. (2011) Zmiany wydajności rolnictwa Polski i innych krajów Unii Europejskiej, *Zeszyty Naukowe Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie – Problemy Rolnictwa Światowego*, **1**, Vol. 11 (26), 130-139.
29. Romesburg H.Ch. (2004) *Cluster Analysis for Researchers*, Lulu Press, North Carolina.
30. Rungsuriyawiboon S., Lissitsa A. (2006) Total Factor Productivity Growth in European Agriculture, http://www.nesdb.go.th/econsocial/macro/tnce/Download/1/supawat.pdf (access: June 27th, 2014), 1-25.

31. Skarżyńska A. (2012) Cropping intensity vs. profitability of selected plant production activities in Poland, *Studies in Agricultural Economics*, **114** (2012), 31-38.
32. Szeląg-Sikora A. (2008) Mierniki oceny uwarunkowań ekonomiczno-technicznych gospodarstw rolnych, *Inżynieria Rolnicza*, **10** (108), 237-244.
33. Yeboah O., Gunden C., Shaik S., Allen A., Li T. (2011) Measurements of Agricultural Productivity and Efficiency Gains from NAFTA, http://ageconsearch.umn.edu/bitstream/98726/2/Yeboah_Gunden_Shaik_Allen_Li_SAEA.pdf (access: June 27th, 2014), 1-12.
34. Wieliczko B. (2011) Rozpoznanie wpływu kryteriów podziału środków finansowych pomiędzy państwa członkowskie na poszczególne przewagi konkurencyjne polskiego rolnictwa, *Przewagi konkurencyjne polskiego rolnictwa w różnych modelach płatności bezpośrednich*, expert report prepared for the Ministry of Agriculture and Rural Development under the leadership of Prof. W. Jóźwiak, Warszawa. Courtesy of Ms. B. Wieliczko, PhD.
35. Wieliczko B. (2013) *Państwo a rynek w rolnictwie – rolnictwo Polski i UE w pierwszych dekadach XXI wieku*, IX Kongres Ekonomistów Polskich, Warszawa.
36. Ząbkowski T., Szczesny W. (2012) Badanie atrakcyjności oferty dostępu do Internetu za pomocą analizy gradacyjnej, in *Metody ilościowe w badaniach ekonomicznych*, Vol. XIII/3, 276-287.

ANALIZA ROLNICTWA UNII EUROPEJSKIEJ PRZY UŻYCIU GRADACYJNEJ ANALIZY DANYCH

Streszczenie. Artykuł prezentuje wyniki analizy rolnictwa w Unii Europejskiej. Na podstawie 15 kluczowych cech dokonano podziału krajów członkowskich na grupy złożone z krajów o podobnej kondycji rolnictwa. W badaniach zastosowano gradacyjną analizę danych, a do przetwarzania danych wykorzystano program GradeStat. Tekst składa się z czterech części. Pierwsza część to krótka analiza rolnictwa jako działu gospodarki, druga zaś jest poświęcona pomiarowi kondycji rolnictwa. Część trzecia prezentuje narzędzia badawcze: analizę skupień i gradacyjną analizę danych, a czwarta przedstawia wyniki zastosowania owych narzędzi do oceny rolnictwa Unii Europejskiej.
Słowa kluczowe: rolnictwo, produktywność, analiza skupień, gradacyjna analiza danych, nadreprezentacja, element odstający

ISBN 83-894-7555-3