ALICJA KOLASA-WIĘCEK[1]

THE USE OF CLUSTER ANALYSIS IN THE CLASSIFICATION OF SIMILARITIES IN VARIABLES ASSOCIATED WITH AGRICULTURAL GREENHOUSE GASES EMISSIONS IN OECD COUNTRIES

Abstract. The aim of the research was to group members of the Organization for Economic Co-operation and Development (OECD) into homogeneous subsets for similarities of agricultural variables affecting greenhouse gas emissions. Cluster analysis, which is a tool for exploratory data analysis, was used. This method is based on grouping of elements in a relatively homogeneous class. The most popular non-hierarchical clustering method is *k*-means. The method is based on an initial a priori assumption of input data set to a predetermined number of classes. In order to verify if the number of clusters was assumed properly, results were compared with another method of cluster analysis – a hierarchical method. Ward's method of classifying on the basis of minimizing the interclass variance was used. Countries qualified for each cluster derived using *k*-means were identical to those obtained using Ward's method. Analysis of the results lead to the conclusion that the geographical location of the countries was key to its inclusion in a cluster this was shown clearly in cluster 1 (Finland, Iceland, Norway, Sweden, Canada), cluster 2 (Austria, Czech Republic, Poland, Slovakia, Switzerland) and cluster 4 (Australia, New Zealand). Group 3 is a 15-element set of countries in predominantly highly industrialized regions.

Keywords: cluster analysis, *k*-means method, Ward's method, greenhouse gases, agriculture emissions

[1] The Author is researcher at Department of Economics and Regional Research, Faculty of Economy and Management, Opole University of Technology (e-mail: a.kolasa-wiecek@po.opole.pl).

INTRODUCTION

Key pollutants emitted as a result of agricultural production are methane and nitrous oxide. These pollutants are important because of their potential high global impact. The supposition is that greenhouse gas emissions may increase in the future, the main reason for which is related to a growth in the human population. Intensification of agricultural production will be a consequence of the growth in food demand. A growing demand for food forces the growth of the livestock population, greater use of nitrogen fertilisers and increasing areas under cultivation. The diet of developed countries is especially rich in animal protein. For example, meat consumption in European countries and the United States of America is estimated at 100 kg/person/year. In comparison, in central and southern Africa, this is approximately 13 kg/person/year [Pietrzak 2009: 143–158]. 6 million hectares of forested areas and 7 million hectares of other areas are converted to agricultural use in developing countries every year [Metz et al. 2007]. It is estimated that a similar trend will continue in the coming years [Green et al. 2005: 550–555]. The implementation of new practices in animal husbandry, land cultivation and fertiliser application has a significant role in reducing these emissions. The technical possibilities of reducing pollution from agriculture were estimated to be 4500 million tonnes carbon dioxide (Mt CO_2) equivalent [Caldeira et al. 2004: 103–129] or even 5,500–6,000 Mt CO_2 or equivalent by 2,030 [Smith et al. 2007].

Scientists, around the world, however take issue over the estimation of the volume and direction of the development of future emissions caused by agricultural livestock and crop production [Jarvis and Pain 1994: 27–38, Klimont and Brik 2004, Pathak and Wassmann 2005: 113–123, Li et al. 2005, Shih et al. 2008].

Accumulated studies and data can constitute a valuable source of information that has not been previously used. The increasing possibilities for data collection require methods which allow valuable information and complementary knowledge in the discipline to be found. This includes using algorithmic tools for data mining [Hand et al. 2005].

EXPERIMENTAL PROCEDURES

The aim of study was to group members of the Organization for Economic Co-operation and Development (OECD) into homogeneous subsets with similarities in the areas of agricultural variables affecting the main greenhouse gas emissions (GHG).

Cluster analysis was used in this research. It is a tool for exploratory data analysis. The method is based on the grouping of elements in relatively homogeneous classes. The basis of clustering in most algorithms was a similarity between elements, as expressed by a similarity function [Kaufman and Rousseeuw 2005].

Based on the analysis, the object is assigned to the class whose severity center lies closest in the sense of Euclidean distance. Cluster analysis can be used to detect structures in data without outputting interpretation and explaining reasons for their occurrence.

K-means is the most popular non-hierarchical clustering method. The method is based on an initial assumption of a priori of input data set for a predetermined number of classes. Then, in order to minimize the variability inside clusters and maximize variability between clusters, it is advisable to move objects between those clusters, until in this iteration, all the objects do not change their class [Grabiński 1992].

Because the stored output variables are expressed in various units and have different areas of variability, data was normalized. In practice, standardization is often used, in accordance with the equation [Ostasiewicz 1998]:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{1}$$

where:
x_{ij} – value for the *i*-th object and the *j*-th characteristic.
$\bar{x}_j$ – average of the *j*-th characteristic.
S_j – standard deviation of the *j*-th characteristic.

Grouping of the member countries of OECD was conducted in order to evaluate a similarity of factors associated with greenhouse gas carbon (GGC) emissions. Analysis was performed using the Statistica package 10.0.

20 variables were identified, directly or indirectly responsible for agricultural GGC emissions guided by the databases of OECD[2], the Food and Agriculture Organization of the United Nations (FAO)[3] and the International Fertilizer Industry Association (IFA)[4]. These were: arable and permanent crop area, ha – X1, the pasture area, ha – X2, other categories of agricultural land, ha – X3, the forests, ha – X4, wheat, ha – X5, barley, ha – X6, oats, ha – X7, sugar beet, ha – X8, maize, ha – X9, rape, ha – X10, peas, ha – X11, triticale, ha – X12, rye, ha – X13, number of cattle, pc – X14, number of pig, pc – X15, number of sheep, pc – X16, number of poultry, pc – X17, number of horses, pc – X18, number of goat, pc – X19, the consumption of nitrogenous fertilizers, t – X20. Because of incomplete data or a lack of it, some countries of the OECD organization were not included in the analysis.

RESULTS AND DISCUSSION OF RESULTS

Analysis was performed for clusters 2–6. For the final cluster centres six objects were selected. The optimal clustering was found after two iterations. Table 1 describes measures of the variation of inter- and intra-group differences following diagnostic variables with corresponding degrees of freedom (df). Snedecor F-statistic values obtained as the ratio of inter-group variation for intra-group variation enabled them to establish a hierarchy of variables because of their discriminatory power. Most of the variables were qualified for characteristics that significantly affect the division into groups.

[2] htttp://stats.oecd.org/index.aspx, available: 20.06.2012
[3] http://faostat.fao.org/site/573/DesktopDefault.aspx?PageID=573#ancor, available: 3.08.2012
[4] http://www.fertilizer.org/ifa/ifadata/search, available: 23.08.2012

TABLE 1. Analysis of variance
TABELA 1. Analiza wariacji

Variables	Inter-group variation	df	Intra-group variation	df	F	Significance p
X1	0.79723	5	1.158286	23	3.166	0.025549
X2	7.33867	5	4.791921	23	7.045	0.000400
X3	49.22104	5	2.899943	23	78.076	0.000000
X4	7.75523	5	6.074142	23	5.873	0.001226
X5	0.08069	5	0.089148	23	4.164	0.007700
X6	0.11551	5	0.041675	23	12.750	0.000005
X7	0.16809	5	0.058134	23	13.300	0.000004
X8	0.16976	5	0.062198	23	12.555	0.000006
X9	0.18465	5	0.059790	23	14.206	0.000002
X10	0.12639	5	0.060890	23	9.549	0.000049
X11	0.18462	5	0.066385	23	12.793	0.000005
X12	0.15949	5	0.066312	23	11.063	0.000016
X13	0.14754	5	0.072082	23	9.415	0.000054
X14	2.25719	5	0.716127	23	14.499	0.000002
X15	7.06155	5	1.501173	23	21.638	0.000000
X16	9.16959	5	6.213348	23	6.789	0.000506
X17	80.62230	5	0.828263	23	447.759	0.000000
X18	0.17344	5	0.063431	23	12.578	0.000006
X19	15.81002	5	0.052765	23	1378.308	0.000000
X20	0.15466	5	0.048411	23	14.696	0.000002

Detailed results of the grouping which indicated the distance of the country from the centre of its concentration are shown in Table 2.

It is noted that two countries – Denmark and Luxembourg are characterised by a homogeneity with regard to the rest of countries acting as outlying points. The classification of countries in which the variables were analysed shows similarities between Scandinavian countries and Canada.

Group 2 was assigned to Central European countries in which Poland was included. Cluster 3 comprised of 15 countries, which are mainly industrialized nations. Cluster 4 consists of two countries with similar geographical characteristics.

While analysing the averages in each group (Figure 1), particular attention was paid to focusing cluster 1 on countries with a high share of the other categories of agricultural land (case 3) and forests (case 4). The selection of cluster 4 is also characteristic – with high levels of variables 2 and 3, i.e. the areas of pasture and other agricultural land and 16 – the number of sheep. Highly significant variables were attributed to livestock farming. For countries which qualified for clusters 2, 3 and 5,

TABLE 2. Members of each cluster and the distance from the centre of the cluster
TABELA 2. Elementy skupienia oraz odległości od środka właściwego skupienia

Members of cluster	Distance from the centre of the cluster
Cluster 1	
Canada	0.097664
Finland	0.289702
Iceland	0.458118
Sweden	0.126286
Norway	0.247375
Cluster 2	
Austria	0.131997
Czech Republic	0.113748
Poland	0.150845
Slovakia	0.118199
Switzerland	0.177267
Cluster 3	
Belgium	0.150818
France	0.056840
Greece	0.158914
Germany	0.168327
Hungary	0.108595
Italy	0.052355
Japan	0.097748
Korea	0.088240
Mexico	0.198784
Netherlands	0.136378
Portugal	0.066783
Spain	0.165688
Turkey	0.138322
Great Britain	0.205686
USA	0.190120
Cluster 4	
Australia	0.483828
New Zealand	0.483828
Cluster 5	
Denmark	0.00
Cluster 6	
Luxembourg	0.00

there was a very high share of poultry production (case 17). Untypical, isolated cases among the analysed countries are Denmark and Luxembourg. Denmark has a high level of pig and poultry production and Luxembourg a high share of goat production.

An excellent verification as to whether the assumed members of the clusters was suitably selected was to compare the results with other methods of cluster analysis – hierarchical (agglomeration). Moreover, in the case of data mining, it is preferred that more than one algorithm is used to solving the problem [Tadeusiewicz 2006].

The result of this grouping is a hierarchical tree – dendrogram (Figure 2).

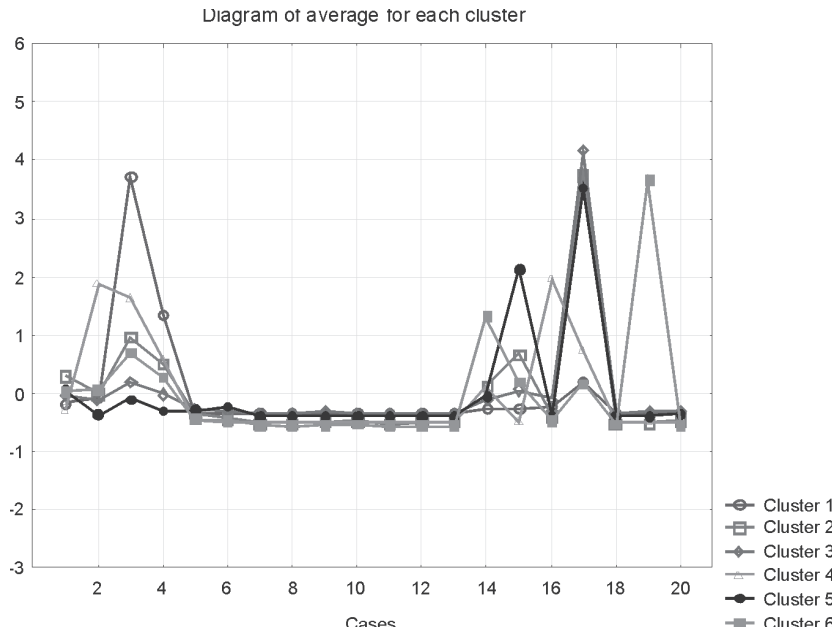

FIGURE 1. Diagram of averages for clusters
RYSUNEK 1. Wykres średnich dla skupień

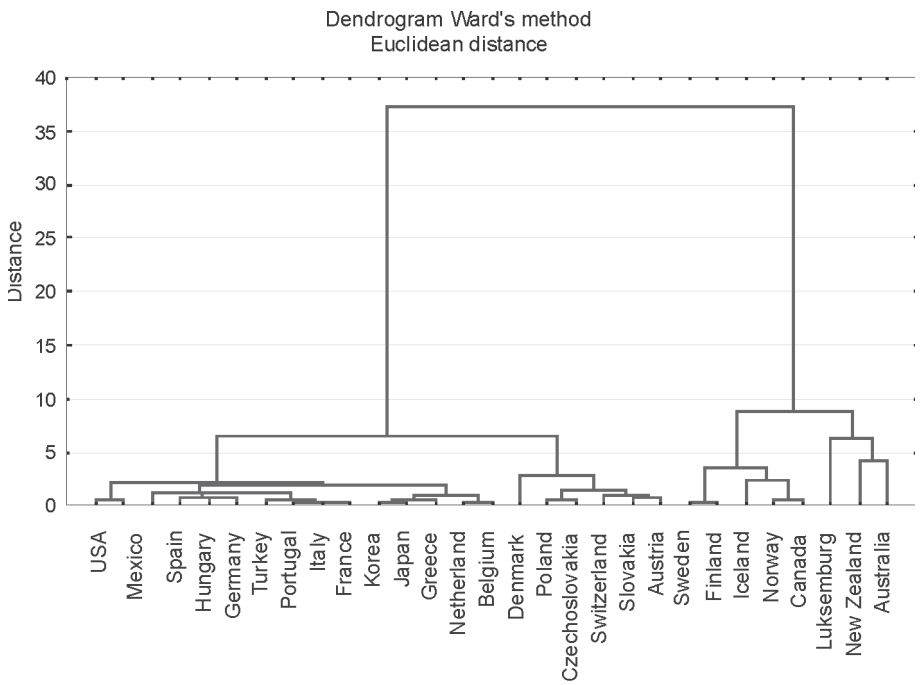

FIGURE 2. Dendrogram using Ward's method
RYSUNEK 2. Dendrogram używany w metodzie Warda

An important parameter affecting the results of the clustering was to choose the method of agglomeration. This is a method of calculating the distance between the clusters [Migut 2009]. In this case Ward's method was used for classifying the basis of minimizing interclass variance. This is a reliable method because of the criterion of the effectiveness of regeneration the real structure of data [Sokołowski 1992]. Allocation of Countries to clusters by *k*-means were identical to those obtained using Ward's method.

CONCLUSION

The use of two different analytical cluster analysis methods can validate the results of research. The analysis shows the existence of clearly varied structures in the grouped countries for similarities in the variables under the study. There are six characteristic groups.

Analysis of the results leads to the conclusion that the geographical location of countries in combination with the selection of variables played a key role in the forming of clusters. Countries included in clusters are often neighbours. Cluster 1 (Finland, Iceland, Norway, Sweden, Canada), cluster 2 (Austria, Czech Republic, Poland, Slovakia, Switzerland) and cluster 4 (Australia, New Zealand). Group 3 comprises of a 15-element set of states and is dominated by highly industrialized regions.

REFERENCES

Caldeira K., Morgan M.G., Baldocchi D., Brewer P.G., Chen C.T.A., Nabuurs G.J., Nakicenovic G.J., Robertson G.P., 2004: *A portfolio of carbon management options.* In: C.B. Field, M.R. Raupach (ed.). *The Global Carbon Cycle. Integrating Humans, Climate, and the Natural World.* SCOPE 62, Island Press, Washington DC: 103–129.

Grabiński T., 1992: *Methods of axonometry*. Akademia Ekonomiczna w Krakowie, Kraków.

Green R.E., Cornell S.J., Scharlemann J.P.W., Balmford A., 2005: *Farming and the fate of wild nature.* "Science" 307: 550–555.

Hand D., Mannila H., Smyth P., 2005: *Data Mining*. WNT, Warszawa.

Jarvis S.C., Pain B.F., 1994: *Greenhouse Gas Emissions from Intensive livestock Systems: Their Estimation and Technologies for Reduction.* "Climatic Change" 17 (1): 27–38.

Kaufman L., Rousseeuw P. J., 2005: *Finding groups in data: an introduction to cluster analysis.* Wiley, New York.

Klimont Z., Brink C., 2004: *Modelling of Emissions of Air Pollutants and Greenhouse Gases from Agricultural Sources in Europe.* Interim Report IR-04-048, International Institute for Applied Systems Analysis, Luxemburg.

Li C., Frolking S., Xiao X., Moore B., Boles S., Qiu J., Huang Y., Salas W., Sass R., 2005: *Modelling impacts of farming management alternatives on CO_2, CH_4, and N_2O emissions: A case study for water management of rice agriculture of China.* "Global Biogeochemical Cycles" 19 (3), doi:10.1029/2004GB002341.

Metz B., Davidson O.R., Bosch P.R., Dave R., Meyer L.A. (ed.) 2007: *Climate Change 2007: Mitigation of Climate Change.* Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridde University Press, Cambridge, United Kingdom and New York, NY, USA.

Migut G., 2009: *Application of techniques cluster analysis and decision trees for market segmentation.* Statsoft Polska.

Ostasiewicz W., 1998: *Statistical methods for data analysis*. Akademia Ekonomiczna we Wrocławiu, Wrocław.
Pathak H., Li C., Wassmann R., 2005: *Greenhouse gas emissions from Indian rice fields: calibration and upscaling using the DNDC model.* "Biogeosciences" 2, 113–123.
Pietrzak S., 2009: *Formation of the nitrogen cycle in macro-and microsystems farming*. "Water – Environment – Rural Areas" 9, 3 (27): 143–158.
Shih J.S., Burtraw D., Palmer K., Siikamaki J., 2008: *Air Emissions of Ammonia and Methane from Livestock Operations: Valuation and Policy Options*. Air & Waste Management Association, Washington.
Smith P., Martino D., Cai Z., Gwary D., Janzen h.H., Kumar P., McCarl B., Ogle S., O'Mara F., Rice C., Scholes R.J., Sirotenko O., Howden M., McAllister T., Pan G., Romanenkov V., Schneider U., Towprayoon S., Wattenbach M., Smith J.U., 2008: *Greenhouse gas mitigation in agriculture.* Philosophical Transactions of the Royal Society of London. B 363, 1492, 789–813.
Sokołowski A., 1992: *Empirical significance tests in the taxonomy.* Akademia Ekonomiczna w Krakowie. Zeszyty Naukowe. Monografie 108.
Tadeusiewicz R., 2006: *Data mining as an opportunity for relatively cheap carrying out scientific discoveries through digging seemingly fully exploited empirical data.* In: *Statistics and Data Mining in research.* Ed. J. Wątroba. StatSoft, Kraków.

WYKORZYSTANIE ANALIZY SKUPIEŃ W KLASYFIKACJI PODOBIEŃSTW W OBSZARZE ZMIENNYCH POWIĄZANYCH Z ROLNICZYMI EMISJAMI GAZÓW CIEPLARNIANYCH W KRAJACH WSPÓLNOTY OECD

Streszczenie. Celem badań było pogrupowanie państw członkowskich *Organization for Economic Co-operation and Development* (OECD) w jednorodne podzbiory pod kątem podobieństwa w obszarach zmiennych oddziałujących na rolnicze emisje głównych gazów cieplarnianych (GGC). W tym celu wykorzystano analizę skupień, która jest narzędziem służącym do eksploracyjnej analizy danych. Metoda polega na grupowaniu elementów we względnie jednorodne klasy. Najbardziej popularną niehierarchiczną metodą skupień jest metoda *k*-średnich. Polega ona na wstępnym założeniu *a priori* wejściowego zbioru danych na z góry określoną liczbę klas. W celu weryfikacji, czy liczba założonych skupień jest odpowiednio dobrana, porównano otrzymane wyniki z hierarchiczną metodą analizy skupień. Wykorzystano metodę Warda klasyfikującą na zasadzie minimalizacji wariancji wewnątrzklasowej. Analiza wyników skłania do wniosku, że na podstawie badanych zmiennych otrzymano skupiska, w których kluczową rolę odgrywa położenia geograficzne państw, czego przykładem jest skupienie 1 (Finlandia, Islandia, Norwegia, Szwecja i Kanada), skupienie 2 (Austria, Czechy, Polska, Słowacja i Szwajcaria) oraz skupienie 4 (Australia i Nowa Zelandia). Skupienie 3 jest 15-elementowym zbiorem państw, w których dominują wysoko uprzemysłowione regiony.

Słowa kluczowe: analiza skupień, metoda *k*-średnich, metoda Warda, metoda hierarchiczna, metoda niehierarchiczna, gazy cieplarniane, rolnictwo