ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 196, 2006

*Hanna Dudek**, *Arkadiusz Orlowski***

CLUSTERING OF EUROPEAN COUNTRIES WITH RESPECT TO FOOD CONSUMPTION

Abstract. Problem of clustering of European countries with respect to food consumption is considered. Data related to average yearly *per capita* consumption of 14 main categories of food products in 39 countries are collected and analysed. Food consumption data for two years: 2000 and 1993 are elaborated. The year 2000 was because there are no more recent data sets available. The year 1993 was chosen as a good reference point: data for that year are the oldest complete. To perform a reasonable grouping of countries the cluster analysis is performed. As a proper number of cluster is not known in advance, hierarchical methods offered by statistical packages Statgraphics are used. The desirable number of clusters is estimated by distance matrices analysis, dendrograms, and graphical representations of distance between clusters with respect to different clustering stages. Squared Euclidean distance is used as a measure of similarity. It is remarkable that all hierarchical methods applied in this paper, apart from nearest neighborhood approach, lead to very similar classification results. Therefore we believe that obtained results provide a valuable and objective insight into the problem of diversification of food consumption in Europe. It has been verified that in spite of visible changes in food consumption in investigated countries, sets of countries belonging to particular clusters obtained for 2000 and for 1993 are almost indistinguishable.

Key words: food consumption, cluster analysis.

1. INTRODUCTION AND DATA

In this paper the problem of clustering of European countries with respect to food consumption is considered. Data related to average yearly *per capita* consumption of 14 main categories of food products in 39 countries are collected and analyzed. Food consumption data for two years: 2000 and 1993 are elaborated. First of all we use the most recent data

* Ph.D., Department of Econometrics and Informatics, Warsaw Agricultural University.
** Professor, Department of Econometrics and Informatics, Warsaw Agricultural University and Institute of Physics, Polish Academy of Science.

available from the FAO data base published in Food Balance Sheet (2002), namely those regarding the year 2000. Moreover, to check if the structure of clusters does not change in time, we perform similar analysis for the data from the year 1993. The latter year is the earliest one for which the complete data coming from all considered countries are fully available – it should be noted that within the last 20 years quite a lot of new countries appeared in Europe after a decay of the former Soviet Union, a splitting of Czechoslovakia, and a break up of the former Yugoslavia. Due to the fact that all the data published by FAO are collected and prepared using the same methodology for all investigated countries they are well suited to perform meaningful comparisons.

The following 14 products (food categories) were used in our investigations:

- cereals,
- potatoes,
- sweeteners,
- pulses,
- vegetable oils,
- vegetables,
- fruits,
- stimulants,
- meat,
- offal's,
- animal fats,
- milk,
- eggs,
- fish and seafood.

All items describe the annual consumption in kilograms per person. All data, including milk and eggs consumption, are given in kilograms. We take into account 39 (thus almost all) European countries apart from the really small ones (as, e.g., Monaco and San Marino) for which no data are recorded.

2. METHODS OF ANALYSIS

In order to perform a reasonable grouping of European countries with respect to food consumption we use a well-known statistical method called the cluster analysis. The idea is to make such a grouping that leads to the clusters consisting of the maximally similar objects and, at the same time, which creates clusters that are maximally different from each other. In the

literature two main categories of the classification methods are distinguished: hierarchical and non-hierarchical (e.g. Ostasiewicz 1999, Dobosz 2001). In the former (hierarchical) approach each object forms at the beginning a separate cluster by itself. At the subsequent stages of the clustering procedure the investigated objects are incorporated into proper clusters using a chosen similarity measure. Typically the following similarity measures are used:

- Euclidean distance,
- square of Euclidean distance,
- city-block metrics,
- Mahalanobis distance,
- Tshebyshev distance.

To provide a more objective analysis the data should be normalized, (e.g. Rószkiewicz 2002a). There are many normalization procedures described and applied in the literature (e.g. Kukuła 2000). In this paper we use a standardization method.

Let us recall the main steps of any clustering method, (e.g. Rószkiewicz 2002b):

1. Defining a distance matrix.
2. Choosing the smallest value in the distance matrix (without taking into account the main diagonal) and creating a cluster of objects corresponding to that distance. These objects are then removed from the data set.
3. Re-computing the distance matrix again for the reduced set of objects. Distances between clusters (or objects) not affected by the step 2 do not change. Distances between newly created cluster and the existing ones are computed anew.

The above procedure should be repeated until all the objects end up in a single cluster. Of course, to find a distance between clusters different agglomeration techniques can be applied. The most popular are the following:

- nearest neighbor method,
- farthest neighbor method,
- group average method,
- centroid method,
- median method,
- Ward's method.

They are thoroughly described in many textbooks (e.g. Marek 1989), Ostasiewicz 1999, Rószkiewicz 2002 a, b, Timm 2002). It is clear from the above that hierarchical methods of clustering are based on iterations: at each stage a newly created cluster consists of all earlier created ones.

In our study all the calculations were performed using statistical software Statgraphics. It offers three possible distance measures: square of Euclidean

distance, Euclidean distance, and the city-block metrics. Besides the above mentioned six agglomeration techniques there is also a possibility of choosing one non-hierarchical method, namely the *k*-averages method. However, as we do not know the proper number of clusters in advance, we stick in this paper to hierarchical methods. Using these methods we can observe distances between clusters at the intermediate clustering stages. Subsequent grouping steps can be visualized in so-called dendrograms, i.e., in plots with vertical axes describing agglomeration distances and horizontal axes displaying the objects themselves. Unfortunately there are no unique methods of defining the proper number of clusters (e.g. Timm 2002).

In this paper, to find out a proper number of clusters we analyzed the distance matrices, dendrograms, and plots of clusters distances (vertical axes) as functions of the clustering stages (horizontal axes), called agglomeration distance plots. A first pronounced jump in distance suggests a stage at which the clustering process should be terminated. The agglomeration distance obtained that way enables us to fix a proper number of clusters via dendrograms (e.g. Dobosz 2001). Some authors (e.g. Marek 1989) points out that the number of clusters should depend on the character of analyzed data.

3. RESULTS

As a similarity measure we decided to take the square of Euclidean distance. As three is hardly the universally best clustering method – all of them have odds and pluses (cf. Milligan, Cooper 1985) – we tried all the hierarchical methods described in the previous section and available in the Statgraphics package. For each method we separately established a proper number of clusters. Our procedure of performing this task is presented in Figures 1 and 2 (illustrated via Ward's method, e.g. Table 1).

From Figure 1 we see that the agglomeration distance should not exceed 60. Using this fact we infer from Figure 2 that it leads to 4 clusters. Another option is to fix the "critical" agglomeration distance at the level 25, what corresponds to 12 clusters, including many one-element clusters.

Using various methods we obtained the following numbers of clusters: with the nearest neighbour method – 6, with the farthest neighbor method – 13, with the centroid method – 10, with the group average method – 10, with Ward's method – 12, and with the median method – 8. The nearest neighbor method resulted in the so-called chain – a cluster connecting countries being quite far away each other, as e.g., Spain with Poland and with the Scandinavian countries.

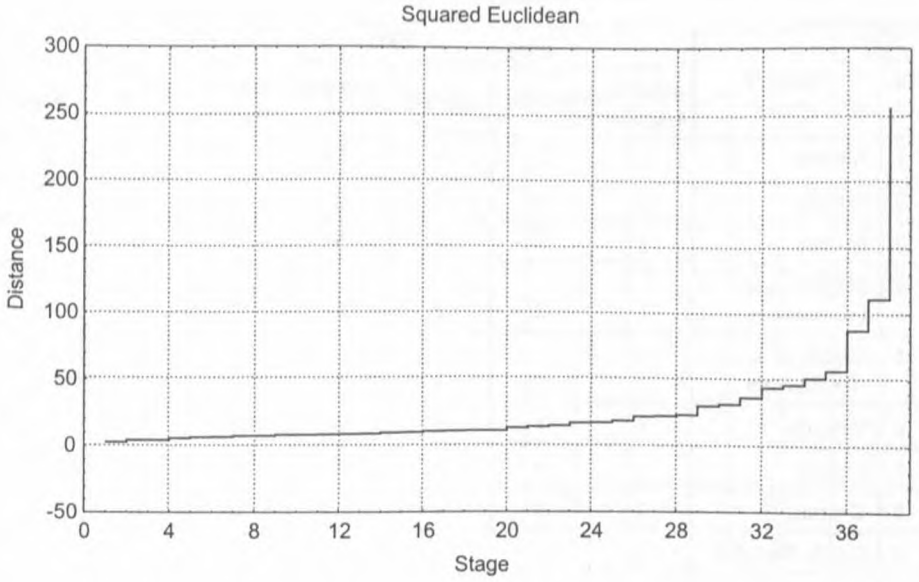

Fig. 1. Agglomeration distance plot Squared Euclidean

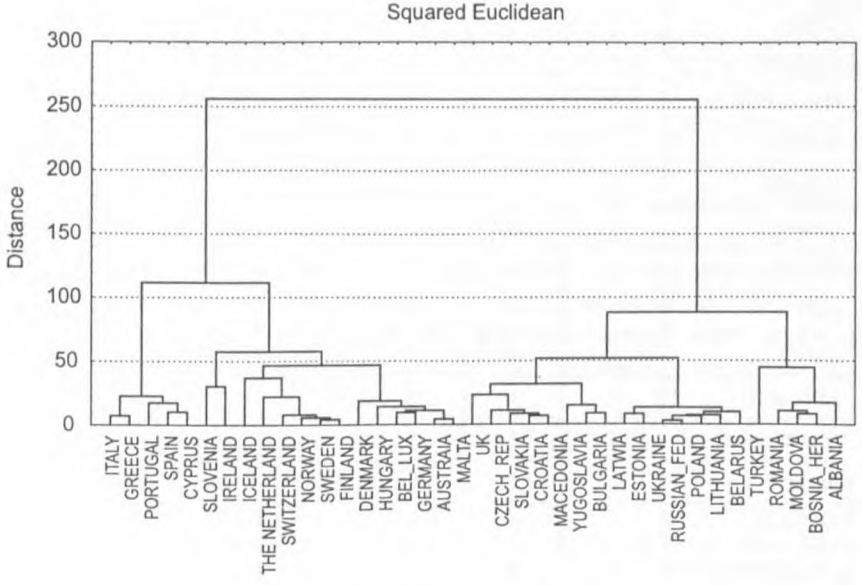

Fig. 2. Ward's method Squared Euclidean

Table 1. Countries and clusters (Ward's method, squared Euclidean distance)

No.	Country	2000						1993
		nearest neighbor	median	centroid	average link	farthest neighbor	Ward's	Ward's
1	Albania	1	1	1	1	1	1	1
2	Austria	1	2	2	2	2	2	2
3	Belarus	1	3	3	3	3	3	3
4	Belgium and Luxembourg	1	2	2	2	2	2	2
5	Bosnia and Herzegovina	1	1	3	1	1	1	1
6	Bulgaria	1	1	3	4	4	4	5
7	Croatia	1	1	3	3	5	5	5
8	Cyprus	1	4	4	5	6	6	6
9	Czech Republic	1	1	3	3	5	5	5
10	Denmark	1	2	2	2	7	2	2
11	Estonia	1	1	3	3	3	3	3
12	Finland	1	2	2	2	7	7	7
13	France	1	2	2	2	2	2	2
14	Germany	1	2	2	2	2	2	2
15	Great Britain	1	1	3	3	5	5	5
16	Greece	1	4	4	5	6	6	6
17	Hungary	1	2	2	2	2	2	2
18	Iceland	3	6	6	7	9	9	7
19	Ireland	2	5	5	6	8	8	8
20	Italy	1	4	4	5	6	6	6
21	Latvia	1	3	3	3	3	3	3
22	Lithuania	1	1	3	3	3	3	3
23	Macedonia	1	1	3	4	4	4	1
24	Malta	1	2	7	8	10	10	6
25	Moldova	1	1	3	1	1	1	1
26	The Netherlands	1	2	2	2	7	7	2
27	Norway	1	2	2	2	7	7	7
28	Poland	1	1	3	3	3	3	3

Table 1. (condt.)

No.	Country	2000						1993
		nearest neighbor	median	centroid	average link	farthest neighbor	Ward's	Ward's
29	Portugal	4	4	8	5	11	6	6
30	Russia Fed.	1	1	3	3	3	3	3
31	Romania	1	1	3	1	1	1	1
32	Slovakia	1	1	3	3	5	5	3
33	Slovenia	5	7	9	9	12	11	5
34	Spain	1	4	4	5	6	6	6
35	Switzerland	1	2	2	2	7	7	2
36	Sweden	1	2	2	2	7	7	7
37	Turkey	6	8	10	10	13	12	9
38	Ukraine	1	1	3	3	3	3	3
39	Yugoslavia	1	1	3	4	4	4	4

Source: own computations performed in Statgraphics.

All the methods produced one-element clusters consisting of Ireland, Iceland, Slovenia, and Turkey, respectively. Moreover the methods of centroids, the farthest neighbor, group average, and Ward's resulted in Malta being a one-element cluster. Using the nearest neighbor method we found Portugal forming another one-element cluster. The group average method, the nearest neighbor method, and Ward's methods give very similar results. In fact the last two methods lead to practically identical classifications (modulo Portugal and Denmark). As they are in a sense complimentary to each other (according to Milligan and Cooper 1985), the nearest neighbor method is less influenced by outliers and Ward's method, influenced by outliers, performs better with noisy data, we decided to restrict our further analysis to the results obtained via the latter one.

In Table 2 average values and in Table 3 standard deviations of consumption of all 14 products in given clusters are presented.

For the year 2000 we obtained the following results. The first cluster consists of Albania, Bosnia and Herzegovina, Moldova, and Romania. All these countries are geographically close to each other. These countries are characterized by high consumption of cereals and vegetables as well as by low consumption of potatoes, animal fats, meat, fish and seafood, stimulants, and sweeteners.

Table 2. Average values of products for 2000 (clustering via Ward's method)

Products	In total	Clusters											
		1	2	3	4	5	6	7	8	9	10	11	12
Cereals	131.6	178.4	111.9	148.1	111.9	109.1	131.8	122.6	128	82.3	178.2	137.7	213.7
Potatoes	83.6	56.9	79.4	135.5	38.6	93.9	71.1	83.7	125.4	50.3	99.5	83.6	64.5
Sweeteners	38.4	27.4	48.0	35.8	29.2	38.8	35.1	37.4	43.9	58.6	51.4	16.4	29.1
Pulses	2.7	2.1	1.8	1.1	4.6	3.2	4.9	3.1	2.9	0.8	3.3	0.9	11.4
Vegetable oils	14.3	9.8	17.2	10.3	12.0	15.2	22.8	15.5	16	7.7	8.7	9.5	17.9
Vegetables	117.9	146.6	110.9	92.0	142.4	85.6	195.2	125.2	73.6	51.8	146.9	61.1	238.7
Fruit	88.2	58.3	107.0	49.1	73.3	75.1	139.7	88.8	84.7	100	62.4	135.6	110.3
Stimulants	6.5	1.6	9.0	4.2	3.5	5.6	6.8	5.8	5.8	14.2	8.4	12.1	2.7
Animal Fats	11.1	2.1	23.0	9.4	8.4	9.0	7.3	11.4	15.6	13.8	10	17.9	1.9
Meat	69.0	27.5	97.9	48.7	64.4	63.5	97.9	74.5	93.7	81.8	72.3	92.6	20.9
Offal's	4.1	3.0	3.6	4.6	3.8	3.1	4.6	3.9	17.4	7.7	2.6	7.9	1.1
Milk	212.4	187.8	233.9	176.6	145.4	172.8	212.9	188.3	301.2	240.5	211.5	222.5	119.9
Eggs	11.0	6.9	13.6	10.9	9.0	12.1	11.1	11.3	7.2	7.2	17.3	11.6	9
Fish. seafood	20.7	3.1	18.8	16.2	4.1	12.4	39.2	18.1	16	90.7	37.8	6.7	7.3

Source: own computations based on FAO data.

Table 3. Standard deviations for products in 2000 (clustering via Ward's method)

Products	In total	Clusters											
		1	2	3	4	5	6	7	8	9	10	11	12
Cereals	48.6	17.0	7.8	20.9	13.5	15.2	25.5	4.8	0	0	0	0	0
Potatoes	38.5	23.1	18.6	19.0	8.6	22.2	35.3	5.8	0	0	0	0	0
Sweeteners	14.3	3.1	5.8	7.3	6.2	4.9	4.5	1.6	0	0	0	0	0
Pulses	2.1	2.1	0.7	1.0	1.5	1.9	0.8	0.6	0	0	0	0	0
Vegetable oils	6.2	2.7	5.6	2.4	2.7	2.7	6.1	1.3	0	0	0	0	0
Vegetables	56.7	53.4	23.2	20.2	46.0	12.5	55.8	17.7	0	0	0	0	0
Fruit	39.3	24.2	18.6	19.6	24.5	10.5	16.9	5.7	0	0	0	0	0
Stimulants	3.8	0.8	3.3	2.6	1.0	1.4	0.8	1.1	0	0	0	0	0
Animal Fats	6.9	0.9	4.1	3.8	5.1	2.6	4.0	1.6	0	0	0	0	0
Meat	31.1	14.0	13.7	14.1	28.4	15.3	11.9	6.4	0	0	0	0	0
Offal's	2.7	1.3	3.0	1.9	2.6	1.1	1.0	0.8	0	0	0	0	0
Milk	82.6	73.0	39.4	20.6	34.1	49.5	44.2	19.6	0	0	0	0	0
Eggs	4.1	2.8	2.1	1.5	2.5	3.2	1.2	0.6	0	0	0	0	0
Fish. seafood	17.6	0.7	9.7	5.5	1.6	6.6	22.4	3.7	0	0	0	0	0

Source: as Table 2.

The second cluster consists of Austria, Belgium and Luxembourg, Denmark, France, Germany, and Hungary. What characterizes these countries is high consumption of fruits, animal fats, meat, fish and seafood, stimulants, and milk as well as by a rather low consumption of cereals.

The third cluster consists of Belarus, Estonia, Lithuania, Latvia, Poland, Russia, and Ukraine. In these countries we observe the highest consumption of potatoes in Europe as well as by the lowest consumption of fruits in Europe.

The fourth cluster consists of Bulgaria, Yugoslavia, and Macedonia. These countries are characterized by higher than average consumption of pulses and vegetables as well as by quite low consumption of potatoes, milk, and fish and seafood.

The fifth cluster consists of Croatia, Czech Republic, Slovakia, and Great Britain. For these countries we have lower than average consumption of cereals and vegetables. Consumption of other products seems to be at the average European level.

The sixth cluster consists of Cyprus, Greece, Italy, Spain, and Portugal. This cluster of countries is characterized by high consumption of pulses, vegetable oils, vegetables, fruits, meat, and fish and seafood.

The seventh cluster consists of Finland, Norway, and Sweden as well as Switzerland and the Netherlands. These countries are characterized by very high consumption of sweeteners, stimulants, milk, and fish and seafood. Somewhat surprising seems to be the presence of Switzerland and the Netherlands in otherwise Scandinavian environment. Also these two countries differs the geographical location, climate, and tradition. The consumption of fish and seafood in these two countries is lower than in Scandinavian countries. Probably the reason for them to belong to this cluster is the very high consumption of milk, sweeteners, and stimulants.

The remaining 5 clusters numbered from 8 to 12 are just one-element clusters and consists of Ireland, Iceland, Malta, Slovenia, and Turkey, respectively.

4. DISCUSSION AND SUMMARY

Results presented in Table 1 show that, apart from just a few exceptions (marked in boldface), most countries belong to the same clusters in 2000 as they did in 1993. This strongly suggests the existence of very stable consumption patterns. In the following we discuss the main differences observed as comparing the two years: 2000 and 1993.

Macedonia left the Balkan cluster no. 1 and moved to cluster no. 4. More detailed analysis reveals quite a substantial change in the consumption pattern: consumption of cereals decreased from 155.2 kg to 125.9 kg; consumption increased for potatoes (from 30.2 kg to 48.4 kg), sweeteners (from 23.8 kg to 35.7 kg), and vegetable oils (from 5.2 kg to 13.7 kg).

Bulgaria moved from cluster no. 5 to cluster no. 4. This could be due to decreasing sweeteners consumption (just opposite happened for other countries from cluster no. 5) and to keeping the level of stimulants consumption (3 kg) – close to the level characteristic for cluster no. 4.

Malta left cluster no. 6 and formed its own one-element cluster. Consumption of potatoes increased from 65.3 kg to 99.5 kg and consumption of fruits decreased from 101.3 kg to 62.4 kg. Trends in the remaining countries of cluster no. 6 were just the opposite.

Slovakia moved from cluster no. 3 (middle-east Europe) to cluster no. 5. Main reason is decreasing consumption of potatoes, milk, and eggs as well as increasing consumption of vegetable oils to the level characteristic for cluster no. 5.

Slovenia left cluster no. 5 and formed its own one-element cluster. Consumption of stimulants almost doubled (from 6.8 kg to 12.1 kg). Also consumption of fruits essentially increased (from 76.5 kg to 135.6 kg). Such trends in the remaining countries of cluster no. 5 were not observed.

The Netherlands and Switzerland moved from cluster no. 2 to the Scandinavian cluster no. 7. In Switzerland we observe increased consumption of sweeteners and milk and decreased consumption of fruits (from 119.1 kg to 91.8 kg). In the Netherlands we observe increasing consumption of fish and seafood (from 13.6 kg to 20.6 kg) and milk (from 306.1 kg to 335.1 kg) as well as quite essentially decreasing consumption of pulses to the level characteristic for Scandinavia.

Iceland left cluster no. 7 and formed its own one-element cluster. Consumption of cereals decreased (in other Scandinavian countries increased). Tends in consumption of animal fats were also quite opposite – it increased in Iceland and decreased in the rest of Scandinavia.

Of course we can see clear changes in consumption patterns in all investigated countries. It should be noted, however, that these changes were usually similar in countries belonging to the same cluster so they did not result in any essential rearrangements of the clusters content. The presented results are just the preliminary ones and we plan to continue our investigations along similar lines in forthcoming papers.

REFERENCES

Dobosz M. (2001), *Wspomagana komputerowo statystyczna analiza wyników badań*, Akademicka Oficyna Wydawnicza EXIT, Warszawa.
Food Balance Sheet (2002), FAO, Roma.
Gulbicka B. (2000), *Wyżywienie polskiego społeczeństwa w ostatniej dekadzie XX wieku*, Instytut Ekonomiki Rolnictwa i Gospodarki Żywnościowej, Warszawa.
Kukuła K. (2000), *Metoda unitaryzacji zerowanej*, Wydawnictwo Naukowe PWN, Warszawa.
Milligan G. W., Cooper M. C. (1985), "An Examination of Procedures for Determining the Number of Clusters in a Data Set", *Psychometrika*, **50**, 159–179.
Marek T. (1989), *Analiza skupień w badaniach empirycznych. Metody SAHN*, Państwowe Wydawnictwo Naukowe, Warszawa.
Ostasiewicz W. (red.), (1999), *Statystyczne metody analizy danych*, Wydawnictwo Akademii Ekonomicznej, Wrocław.
Rószkiewicz M. (2002a), *Metody ilościowe w badaniach marketingowych*, Wydawnictwo Naukowe PWN, Warszawa.
Rószkiewicz M. (2002b), *Narzędzia statystyczne w analizach marketingowych*, Wydawnictwo C. H. Beck, Warszawa.
Timm N. H. (2002), *Applied Multivariate Analysis*, Springer-Verlag, New York.

Hanna Dudek, Arkadiusz Orłowski

GRUPOWANIE PAŃSTW EUROPEJSKICH ZE WZGLĘDU NA SPOŻYCIE ŻYWNOŚCI

(Streszczenie)

W artykule rozważono zagadnienie pogrupowania państw europejskich ze względu na konsumpcję żywności. Zgromadzono dane o rocznym spożyciu na osobę 14 głównych grup produktów żywnościowych w 39 państwach. Dane dotyczą konsumpcji żywności w latach 2000 oraz 1993. W celu pogrupowania państw wykorzystano analizę skupień. Z uwagi na brak przesłanek dotyczących liczby skupień zastosowano hierarchiczne metody aglomeracyjne, oprogramowane w pakietach statystycznych Statgraphics. Liczbę skupień ustalono na podstawie analizy macierzy odległości, dendrogramów oraz wykresów odległości skupień względem etapów grupowania. Za miarę podobieństwa przyjęto kwadrat odległości euklidesowej. Ustalono, że poza metodą najbliższego sąsiedztwa, wszystkie hierarchiczne metody aglomeracyjne prowadzą do skupień o zbliżonym zestawie państw. Na podstawie wykonanej analizy skupień stwierdzono, że mimo zmian w spożyciu produktów żywnościowych w poszczególnych krajach, zestawy państw w otrzymanych skupieniach w roku 2000 i 1993 były niemal identyczne.