Searching for suitable method for clustering the EU regions according to their agricultural characteristics

Ondřej Šimpach[1], Marie Pechrová[2]

Abstract. Not only Common Agricultural Policy, but also other policies at the EU level require unified approach. However, agriculture in various member states as same as in particular regions differs significantly. Therefore, the aim of the paper is to find the appropriate method which will group the most similar regions according to their agricultural characteristics and create appropriate number of groups which would be suitable for the application of the agricultural policy. These results from different methods of HCA are compared and discussed. Particularly single, average, complete, weighted average, median, centroid, and Ward's methods are applied. Also various metrics (Euclidean, Square Euclidean, absolute and maximum-value, Minkowski and Canberra distances, and correlation coefficient and angular separation measures) are used. The data about NUTS II regions of the EU are obtained from Eurostat for the latest available year (mainly 2013). Main indicators for the description of agricultural in particular region were: agricultural income, utilized agricultural area, labour and others. The most suitable well-balanced groups were created by Ward's method with Canberra distance.

Keywords: Clustering, measures, NUTS 2 regions, agricultural policy

JEL Classification: C38, Q18
AMS Classification: 62H30

1 Introduction

Spatial econometrics is an important tool for support of the policy-making decisions. The analyses enable to see the results of taken measures and suggest needed correction. For example research of Smith et al. [12] used spatial econometric techniques to evaluate RDPs in the European Union, at the NUTS 2 level. They focused specifically on labour productivity in the agricultural sector. At first side their results seemed to show that spending within the regional development programs on the competitiveness program (axis 1) had a statistically significant positive relationship with the increase of agricultural labour productivity in southern Europe. However, when their controlled properly for spatial effects (rural versus urban areas), the effect disappeared. "This shows how not taking spatial econometrics into account can lead to erroneous (policy) conclusions" [12]. Similarly Becker et al. [1] examined the effects of EU's structural policy (particularly of the Objective 1 facilitating convergence and cohesion within the EU regions). They observed the development of average annual growth of GDP per capita at purchasing power parity (PPP) during a programming period and average annual employment growth at NUTS 2 and NUTS 3 levels. Becker et al. [1] concluded that "Objective 1 treatment status does not cause immediate effects but takes, in the average programming period and region, at least four years to display growth effects on GDP per capita". Also Palevičienė and Dumčiuvienė [6] used multivariate statistical methods to analyse the EU's NUTS 2 level socio-economic data and to identify the clusters of socio-economic similarity. "The results showed that despite long lasting purposeful structural funds allocations there are still big regional development gaps between European Union member states" [6]. Pechrová and Šimpach [7] searched for the development potential of the NUTS 2 regions using hierarchical cluster analysis (Ward's method and Squared Euclidean distances), the regions were clustered into groups with same characteristics.

"The cluster analysis objective is to find out which objects are similar or dissimilar to each other" [9]. There are various clustering algorithms. Basic division is on hierarchical and non-hierarchical methods. First mentioned group contains agglomerative polythetic approach, two-dimension agglomerative clustering, division monothetic and division polythetic approaches. Non-hierarchical methods can be based on *k*-means algorithm or use fuzzy approach to cluster analysis. "When compared to standard clustering, fuzzy clustering provides more flexible and powerful data representation" [10]. Schäffer et al. [11] proposed new Bayesian approach for quantifying spatial clustering that employs a mixture of gamma distributions to model the squared distance of points to their second

[1] University of Economics Prague, Faculty of Informatics and Statistics, Department of Statistics and Probability, W. Churchill sq. 4, 130 67 Prague 3, Czech Republic, ondrej.simpach@vse.cz.

[2] Institute of Agricultural Economics and Information, Modelling of the Impacts of Agricultural Policy, Mánesova 1453/75, 120 00 Prague 2, Czech Republic, pechrova.marie@uzei.cz.

nearest neighbours. Ritter et al. [8] proposed a new method for autonomously finding clusters in spatial data belonging to the nearest neighbour approaches. "It is a repetitive technique which produces changing averages and deviations of nearest neighbour distance parameters and results in a final set of clusters" [8].

2 Data and methods

The aim of the paper is to find the appropriate method which will group the most similar regions according to their agricultural characteristics and create balanced groups. Firstly, the data are introduced, than the hierarchical cluster analysis methods and metrics used in the article are described. The data matrix was obtained from Eurostat [5] for the latest available years (mainly 2013). The data were not available for all EU's regions (only for 2014). All NUTS 2 regions from Germany, Belgium and Slovenia were missing. Unfortunately, very important agricultural indicators such as the herds' sizes and types, structure and acreage of crops were not available for majority of regions. Hence, they were not included into the analysis. The data were checked whether there is the correlation between them. Based on pairwise Pearson correlation coefficients (the values higher than 0.9), some of indicators were excluded from the analysis. In Table 1 we present the descriptive characteristics only for those variables which were chosen to be used in the next step. The agricultural holdings created an output in average height of 1 533 thousand CZK in year 2013. They used around 706 thous. ha of agricultural area and 411 thous. ha of arable land on average. The average share of arable land was 56% on average, but it differed across the EU. Similarly the share of agricultural land was higher (81%), but varied across the EU regions. Indicators related to the number of workers were highly correlated. Therefore, only the number of sole holders working on the farm was used. There were 45 898 of them on average. High standard deviation points on higher variability in the data. Sometimes it is even twice as high (in case of the number of sole holders working on the farm). The presence of variability in the data can negatively influence the results of clustering methods. Zero values stand for region Inner London with no agricultural production and land.

Used variables	**Mean**	**Std. Dev.**	**Min**	**Max**
Agricultural output (million EUR)	1533	1593	0	9 390
Utilized agricultural area (ha)	706 038	752 940	0	5 295 680
Arable land (ha)	410 724	474 818	0	3 371 340
Share of arable land	56%	27%	0%	99%
Share of agricultural land	81%	19%	0%	100%
Sole holders working on the farm (pers.)	45 898	99 096	0	749 260

Table 1 Descriptive characteristics of data from Eurostat (2016); own elaboration

As the variables are in different units standardized. Consequently, hierarchical cluster analysis was applied. At first, the resemblance matrix was calculated. Choice of similarity or dissimilarity measure depends on the type of variables (nominal, ordinal, ratio, interval, and binary). We utilized Euclidean, squared Euclidean, absolute-value, and maximum-value distances, and Minkowski distance with *p* argument, Minkowski distance with *p* argument raised to power (2), Canberra distance, correlation coefficient similarity measure, and angular separation similarity measure. Euclidean and squared Euclidean distances are based on Pythagoras theorem. Euclidean distance between two data points (X_i and Y_i) is calculated as the square root of the sum of the squares of the differences between corresponding values (1).

$$d = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \tag{1}$$

(It is similar to Minkowski distance with argument $p = 2$.) The Euclidean Squared distance metric uses the same equation (1) without the square root. As a result, clustering with the Euclidean Squared distance metric is faster. Calculation of absolute value distance (i.e. Manhattan distance) is similar to Minkowski distance with argument $p = 1$. Maximum-value distance (Czebyshev) is calculated as Minkowski with $p \to \infty$. Minkowski distance with other arguments is calculated as (2).

$$d = \left(\sum_{i=0}^{n-1} |X_i - Y_i|^p \right)^{1/p} \tag{2}$$

It is often used when variables are measured on ratio scales with an absolute zero value. Its disadvantage is that even a few outliers with high values bias the result. Canberra metric is a dissimilarity coefficient defined in interval $a_{jk} = <0;1>$, where $a_{jk} = 0.0$ means maximum similarity when objects *j* and *k* are identical. "Each term in the sum is scaled between 0.0 and 1.0 equalizing the contribution of each attribute to overall similarity" [9]. Multiplier $1/n$ averages the *n* proportions (3):

$$a_{jk} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|X_{ij} - X_{ik}\right|}{\left(X_{ij} + X_{ik}\right)}. \tag{3}$$

Correlation coefficient similarity measure is defined in interval r_{jk} = < -1;1 >, where r_{jk} = 1.0 represents maximum correlation between the variables. Its advantage is that it can be applied on non-normalized data and its accuracy of the score increases. The calculation is as follows (4):

$$r_{jk} = \frac{\sum_{i=1}^{n} X_{ij}X_{ik} - \frac{1}{n}\sum_{i=1}^{n} X_{ij}\sum_{i=1}^{n} X_{ik}}{\sqrt{\left(\sum_{i=1}^{n} X_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_{ij}\right)^2\right)\left(\sum_{i=1}^{n} X_{ik}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_{kj}\right)^2\right)}}. \tag{4}$$

Angular separation similarity measure represents cosine vectors between two angles (5). It is defined in interval s_{jk} = < -1;1 >, where higher value indicates the similarity. The cosine of 0° is 1, for any other angle is < 1.

$$s_{jk} = \frac{\sum_{i=1}^{n} X_{ij}X_{ik}}{\left(\sum_{i=1}^{n} X_{ij}^2 \sum_{r=1}^{n} X_{ir}^2\right)^{1/2}} \tag{5}$$

Both above stated distances place anti-correlated objects maximally far apart (see e.g. [14]). Extended discussion on similarity measures can be found e.g. in [3].

Joining clusters by dissimilarity coefficients two clusters identified with the smallest dissimilarity coefficient values are merged. When using similarity coefficient, two clusters identified with the largest similarity coefficient value are joined. The article uses different clustering methods: single linkage (SLINK), average linkage method (i.e. unweighted pair-group method using arithmetic averages, UPMGA), complete linkage (CLINK), weighted average linkage, median linkage, centroid linkage, and Ward's linkage. Single linkage method finds the two most similar spanning objects in different clusters. SLINK tents to produce compacted trees. It is useful only when clusters are obviously separated. When objects are close to each other, SLINK tends to create long chain-like clusters that can have a relatively large distance separating observations at either end of the chain. Average linkage method defines the similarity between two clusters as the arithmetic average of the similarities between the objects in one cluster and objects in the other cluster. It usually produces trees which are between extremes of those created by single or complete linkage. It also tends to give higher values of the cophenetic correlation coefficient. "On average UPGMA produces less distortion in transforming the similarities between objects into a tree" [9]. It can also have a weighted form. In complete linkage clustering method the similarity between any two clusters is determined by the similarity of their two most dissimilar spanning objects. SLINK tends to produce clusters with similar diameters and extended trees. It is useful when the objects form naturally separated clusters. However, the results can be sensitive to outliers. Median linkage belongs to the averaging techniques, but uses medians instead of arithmetic means. This enables to mitigate the effect of possible outliers in the data. With the median linkage method, the distance between two clusters is the median distance between an observation in one cluster and an observation in the other cluster. Centroid method determines the distance of clusters by the distance of their centres (calculated as averages of real values). Its weighted form is useful when different size of clusters is expected. It usually utilizes the squared Euclidean distance metrics. Ward's method [15] merges the clusters with minimal within-cluster sum of squared deviations from objects to centroids. The distances of objects are again usually measured by squared Euclidean distance. It tends to create relatively small clusters because of the squared differences, but with similar numbers of observations. However, it is sensitive to outliers. Another disadvantage is that the distance between clusters calculated at one step of clustering is dependent on the distance calculated in previous step.

The clustering was cut when the number of clusters was 5 to create reasonable number of groups for policy treatment. The calculations were done in Stata 11.2 where above stated metrics and methods are available.

3 Results and discussion

For the policy making purposes, it is important that the clusters contain similar number of regions. We created 5 clusters by each method. Usual disadvantage of single linkage method – the tendency to chaining – appeared also in our application. Regardless the distance metrics created four clusters with 1 or 2 regions and 1 cluster (number 5) with others. Therefore, it will not be further taken into account. Average linkage method produced also groups where majority of regions was included in one cluster (mostly 1 or 2) and the remaining clusters included only few regions. Canberra distance rearranged the regions differently (majority of them was in 5th cluster, than in 1st). Correlation coefficient and angular separation similarity measures crated more balanced clusters in case of using average linkage clustering method. Complete linkage with majority of used metrics was also not optimal. Almost all regions were clustered to the 1st group. Using Canberra distance, majority of regions was included in 3rd group. It may be due to the presence of outliers in the data. Correlation coefficient and angular separation similarity measures created the most balanced clusters for complete linkage method. Weighted average linkage grouped to 3rd cluster more regions than its unweighted variant. Canberra distance in this case put more regions in 3rd cluster while in unweighted case it was the 5th regions. The number of regions in each cluster was balanced in case of using correlation coefficient and angular separation similarity measures. Median linkage put almost all regions into the 1st cluster. Only using angular separation similarity measure the 3rd cluster emerged as the biggest. Despite that it should mitigate the influence of the outliers to some extent, this was not the case. Probably London region with minimal values in almost all variables represented a problem to this method. Similar problem is with centroid linkage. All regions are in the 1st cluster, only with correlation coefficient and angular separation similarity measures the clusters are more balanced.

Despite that Ward's linkage is normally used with squared Euclidean distance, more feasible seems absolute and maximum values metrics (see Table 2). While the first mention creates one cluster with only three regions (cluster 5), the number of regions grouped by the other stated methods is more equal; but maximum-value method put only 10 regions in the fourth cluster. Therefore, Canberra distance seems to be optimal in terms of the number of regions in each cluster. Minkowski distance with p argument provided similar results to Euclidean distance and Minkowski distance with p argument raised to power to squared Euclidean distance in some cases as the value of parameter p was 2. Correlation coefficient and angular separation similarity created the most suitable groups. As the first one is only a special case of the latter one (angular separation standardized by centring the coordinates on its mean value), we may suggest using the angular separation similarity measure. Canberra distance enabled to create relatively well balanced groups in the last category. Despite that it has certain disadvantage with higher number of variables, this is not our case and we can recommend its usage.

Distance	Cluster's number				
	1	2	3	4	5
Euclidean distance	62	3	58	73	18
squared Euclidean distance	98	22	59	32	3
absolute-value distance	24	22	63	68	37
maximum-value distance	79	57	29	10	39
Minkowski distance with p argument	62	3	58	73	18
Minkowski distance with p argument raised to power	98	22	59	32	3
Canberra distance	71	42	23	45	33
correlation coefficient similarity measure	49	31	64	46	24
angular separation similarity measure	46	27	57	61	23

Table 2 Number of regions in clusters using Ward's linkage method and various metrics; own elaboration

As Ward's (and average) method is according to [4] the most suitable in majority of cases, we suggest combining it with Canberra distance method. Analogical results for UPGMA and SLINK methods were achieved by Bouguettaya et al. [2]. They compared distance measure functions and found out that Canberra distance seems better than the Euclidean counterpart. „With no noticeable difference in computational cost, correlation achieved by the Canberra method is consistently higher than the correlation obtained by the Euclidean method on a same data set with either UPGMA or SLINK" [2]. In our case, this approach grouped regions with minimal values in almost all variables to the fifth cluster and with maximal values to the first cluster (see Table 3).

Cluster	Agric. output (million EUR)	Utilized agric. area (ha)	Arable land (ha)	Share of arable land	Share of agric. land	Sole holders working on the farm (pers.)
1	2 827	1 474 755	871 046	62%	89%	96 704
2	1 189	510 578	224 192	45%	86%	43 615
3	1 198	560 681	473 554	83%	61%	21 287
4	827	160 985	79 111	51%	94%	6 579
5	382	145 464	66 148	44%	54%	10 266

Table 3 Characteristics of clusters created by Ward's method using Canberra distance; own elaboration

Regions grouped to the fifth cluster produced the less agricultural output with the less acreage of UAA and arable land. It contained mainly regions from France, Italy, and Poland. Regions in first cluster produce the highest agricultural output, utilize the most agricultural area and arable land in absolute values and had the most sole holders working on the farm. It included e.g. regions from Austria. The results of Ward's clustering method using different metrics are presented at Figure 1 in Appendix.

4 Conclusion

Agriculture in European Union varies in particular regions. As Common Agricultural Policy requires unified approach, the aim of the paper was to find the appropriate method which will group the most similar regions according to their agricultural characteristics and create appropriate number of groups which would be suitable for the application of the agricultural policy. The results from different methods of hierarchical cluster analysis showed that finding a suitable method for clustering regions is not an easy task. Each clustering method or distance metric provided different results (see also comparison in [13]). Besides, majority of them proved to be sensitive to outliers. Mostly the best results were provided by correlation coefficient and angular separation similarity measures. Surprisingly, often used Ward's linkage method with squared Euclidean distances did not provide well balanced groups. Normally this combination of methods tends to create relatively small clusters because of the squared differences, but with similar numbers of observations which is important for the application in the area of agricultural policy-making. Therefore, we recommend using Ward's method when clustering EU regions for policy purposes using rather Canberra distance. This measure provided well balanced groups with clearly different characteristics which enable to formulate appropriate political measures.

Acknowledgements

The research was supported by the Czech Science Foundation project no. P402/12/G097 DYME – "Dynamic Models in Economics" and also by Thematic task no. 4107/2016 of the Institute of Agricultural Economics and Information.

References

[1] Becker, S. O., Egger, P. H., von Ehrlich, M.: Going NUTS: The effect of EU Structural Funds on regional performance. *Journal of Public Economics* **94** (2010), 578–590.

[2] Bouguettaya, A., Yub, Q., Liub, X., Zhouc, X., Songa, A.: Efficient agglomerative hierarchical clustering. *Expert Systems with Applications* **42** (2015), 2785–2797.

[3] Cha, S. H.: Comprehensive Survey on Distance / Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences* **4** (2007), 300–307.

[4] Dunn, G., Everitt, B. S.: *An introduction to mathematical taxonomy*. Cambridge University Press, 1982.

[5] Eurostat: Agriculture, forestry and fisheries. [on-line] Available at: http://ec.europa.eu/eurostat/data/database [cit. 2016-02-20].

[6] Palevičienė, A., Dumčiuvienė, D.: Socio-Economic Diversity of European Regions: Finding the Impact for Regional Performance. *Procedia Economics and Finance* **23** (2015), 1096–1101.

[7] Pechrová, M., Šimpach, O.: The Development Potential of the Regions of the EU. In: *Region in the Development of Society*. Mendel University in Brno, Brno, (2013), 322–335.

[8] Ritter, G. X., Nieves-Vázquez, J.-A., Urcid, G.: A simple statistics-based nearest neighbor cluster detection algorithm. *Pattern Recognition* **48** (2015), 918–932.

[9] Romesburg, H. C.: *Cluster Analysis For Researchers*. Lulu Press, North Carolina, 2004.

[10] Rovetta, S., Masulli, F.: Visual stability analysis for model selection in graded possibilistic clustering. *Information Sciences* **279** (2014), 37–51.

[11] Schäffer et al.: A Bayesian mixture model to quantify parameters of spatial clustering. *Computational Statistics and Data Analysis* **92** (2015), 163–176.

[12] Smith, M. J., van Leeuwen, E. S., Florax, R. J. G. M., de Groot H. L. F.: Rural development funding and agricultural labour productivity: A spatial analysis of the European Union at the NUTS2 level. *Ecological Indicators* **59** (2015), 6–18.

[13] Šimpach, O.: Application of Cluster Analysis on the Demographic Development of Municipalities in the Districts of Liberecky Region. In: *7th International Days of Statistics and Economics*. Slaný: Melandrium, (2013), 1390–1399.

[14] van Dongen, S. Enright, A.J.: Metric distances derived from cosine similarity and Pearson and Spearman correlations. [on-line] Available at: http://arxiv.org/abs/1208.3145 [cit. 2016-04-14].

[15] Ward, J. H.: Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58** (1963), 236–244.

Appendix

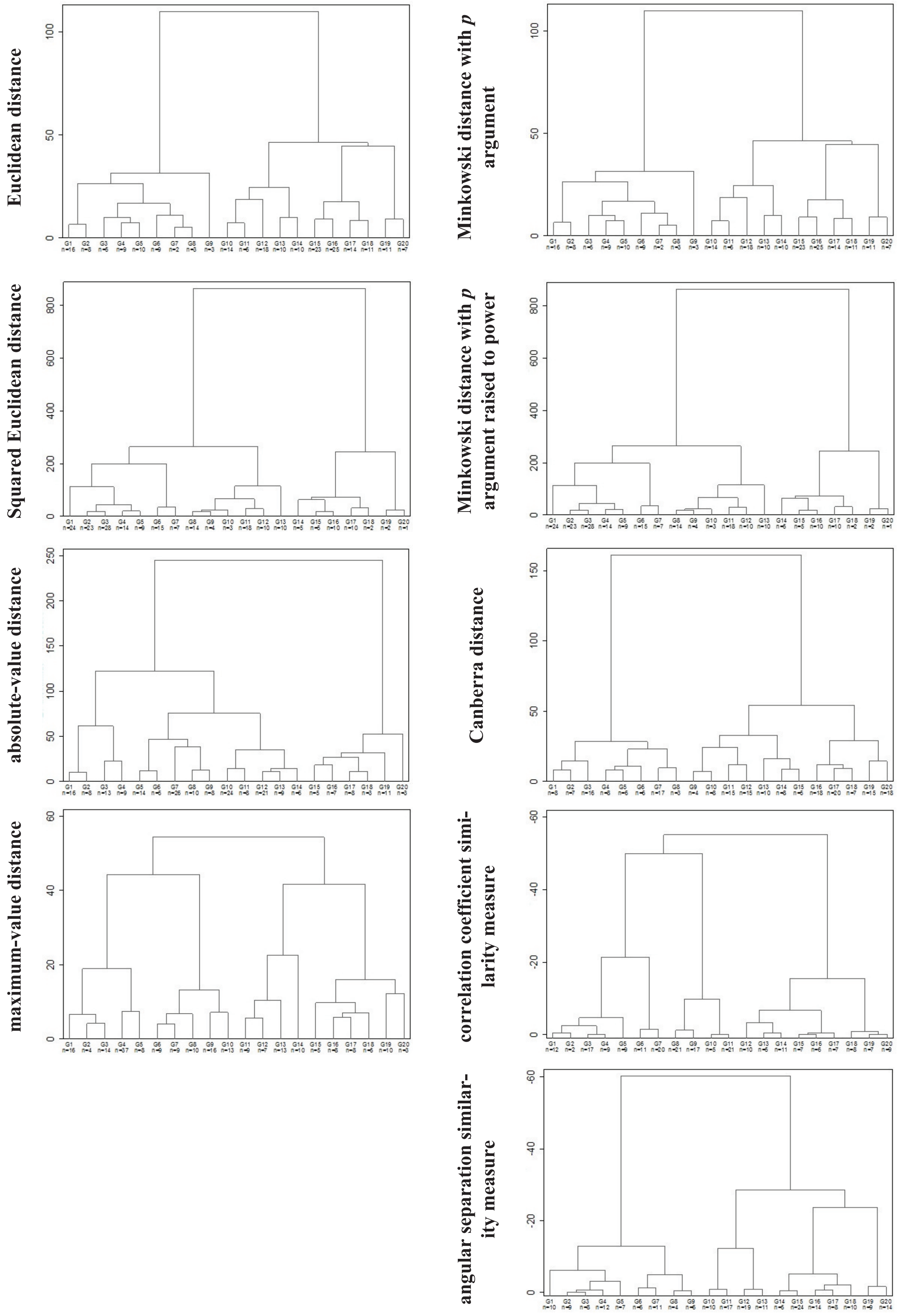

Figure 1 Dendrograms for Ward’s clustering method and various distance metrics; own elaboration