Economic Insights – Trends and Challenges	Vol.V(LXVIII) No. 1/2016	65 - 72

Bootstrap Stability Evaluation and Validation of Clusters Based on Agricultural Indicators of EU Countries

Cristian Marinoiu

Petroleum-Gas University of Ploieşti, Bd. Bucureşti 39, 100680, Ploieşti, Romania
e-mail: marinoiu_c@yahoo.com

Abstract

In this paper we propose a classification of the European Union countries, according to two indicators: number of agricultural holdings (NAH) and utilized agricultural area (UAA). The results obtained are important from the perspective of emphasizing the similarities and differences between European Union countries in relation to the specified indicators. Since the clustering methods usually used for discovering a structure in a data set have a tendency to "detect" clusters even in very homogeneous data set, a special attention was paid to the validation of the obtained results. For this reason, in order to validate the cluster structure obtained by applying hierarchical agglomerative clustering, three techniques were used to estimate the number of clusters, techniques which have led to the same result. In addition to this, we estimated the stability degree of the three clusters obtained. Our results show that one of the clusters has a reasonable stability and the other two have an exceptionally good stability.

Keywords: *hierarchical agglomerative clustering; cluster stability; bootstrap*

JEL Classification: *C38; Q10*

Introduction

Due to its agricultural exceptional resources and to a common agricultural policy (European Commission, 2014) launched since 1962 and increasingly based on sustainable management, the European Union supports today more than 500 million consumers and it is prepared to remain an important player in ensuring food security of the whole world. Characteristics of agricultural holdings in the European Union countries are strongly influenced by climatic factors, geological and geographical factors and also by social factors. This reality is reflected in Eurostat by several indicators, among which (Eurostat-Statistics explained, 2016a): the size of agricultural holdings, the farms labour force, livestock units and agricultural land use. In this paper we intend to achieve a classification of members countries of the European Union based on two agricultural indicators, highlighted by the (Eurostat-Statistics explained, 2016b): number of agricultural holdings (NAH) as a percentage of the total number of EU agricultural holdings and utilized agricultural area (UAA) as a percentage of total utilized agricultural area in UE. The utilized agricultural area (UAA) describes the area used for farming and refers to (Eurostat-Statistics explained, 2016c): arable land, permanent grassland, and crops, other agricultural land. The classification obtained in the following paragraphs creates an overview of the similarities and differences between EU member states in terms of these two indicators.

Classification of EU Countries Using Clustering Techniques

The values of the two indicators (NAH) and (UAA) for the year 2013 are presented in Table 1.

Table 1. Values of NAH and UAA indicators for year 2013

No. Crt.	Country	NAH	UAA
1	Austria	1.3	1.6
2	Belgium	0.3	0.7
3	Bulgaria	2.3	2.7
4	Croatia	1.5	0.9
5	Cyprus	0.3	0.1
6	Czech Republic	0.2	2.0
7	Denmark	0.4	1.5
8	Estonia	0.2	0.5
9	Finland	0.5	1.3
10	France	4.4	15.9
11	Germany	2.6	9.6
12	Greece	6.5	2.8
13	Hungary	4.5	2.7
14	Ireland	1.3	2.8
15	Italy	9.3	6.9
16	Latvia	0.8	1.1
17	Lithuania	1.6	1.6
18	Luxembourg	0.0	0.1
19	Malta	0.1	0.0
20	Netherlands	0.6	1.1
21	Poland	13.2	8.3
22	Portugal	2.4	2.1
23	Romania	33.5	7.5
24	Slovakia	0.2	1.1
25	Slovenia	0.7	0.3
26	Spain	8.9	13.3
27	Sweden	0.6	1.7
28	United Kingdom	1.7	9.9

Source: Eurostat-Statistics explained, 2016b, *Farm_structure_YB2016*

In order to classify EU countries according to values of NAH and UAA indicators we applied the hierarchical agglomerative clustering method from (Theodoridis and Koutroumbas, 2009). The distance between two objects (i.e. countries) is the Euclidean distance and the distance between two clusters is Ward.D2 (a variant of Ward's method). According to (Shalizy, 2009, p.3) "Ward's method says that the distance between two clusters A and B is how much the sum of squares will increase when we merge them"... Mathematically, in this approach, the distance between clusters A and B is defined by (Shalizy., 2009 , p.3):

$$\Delta(A,B) = \sum_{i\in A\cup B}\left\|x_i - m_{A\cup B}\right\|^2 - \sum_{i\in A}\left\|x_i - m_A\right\|^2 - \sum_{i\in B}\left\|x_i - m_{\mathbf{B}}\right\|^2 - \frac{n_{A\cap B}}{n_{A\cup B}}\left\|m_A - m_B\right\|^2 \quad (1)$$

where:

x_i are items that we want to place in clusters;

m_i is the centroid of cluster i;

n_i represents the number of objects from cluster i.

The distance $\Delta(A, B)$ between the clusters A and B is interpreted as "the merging cost of combining the clusters A and B" (Shalizy., 2009 , p.3). The principle of Ward's method is to merge, at every step, the clusters A and B for which this cost is minimal. By applying the hierarchical agglomerative clustering method we obtained the dendogram shown in Figure 1.

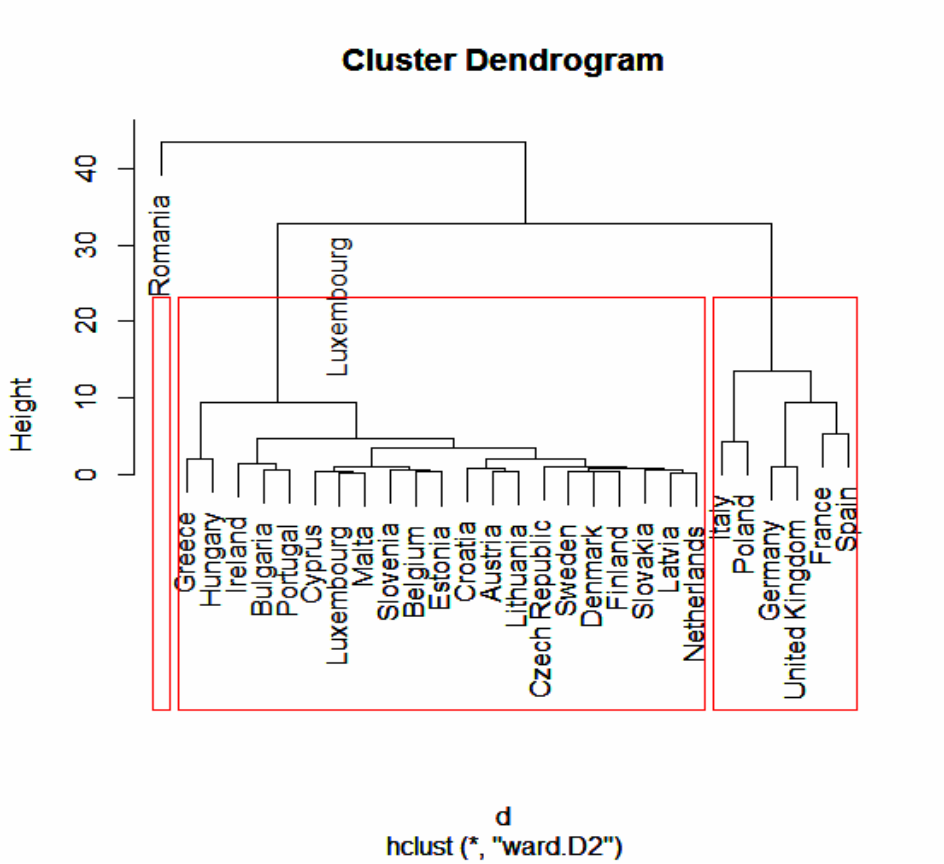

Fig. 1. Dendogram from agglomerative hierarchical clustering

Source: made by the author in R, using data from Table 1

The Validation of the Clusters Structure

The dendogram in Figure 1 suggests the existence of a structure consisting of two or three clusters. In order to better determine the number of clusters, we additionally used the suggestions offered by the plots in Figures 2, 3 and 4.

The graph in Figure 2 (the so-called scree plot (Peeples , M., 2011)) is a plot of the within group sum of squares distances against the number of clusters. For each cluster this distance is computed between the object and the cluster centroid. For the obtained structure the total sum of

squares of each cluster is interpreted as a measure of overall error. In general, by representing graphically the overall error value against to the number of clusters of the structure, we obtain a plot which is similar to that shown in Figure 2, i.e. a plot which rapidly decreases to a point from which the decrease slows down. In this point the plot makes an elbow. The 'right' number of clusters is indicated by the number of clusters which correspond to this elbow of the plot: a greater number of clusters than this number does not produce a significant decrease in the value of global error. As the plot in Figure 2 suggests this number is 3.

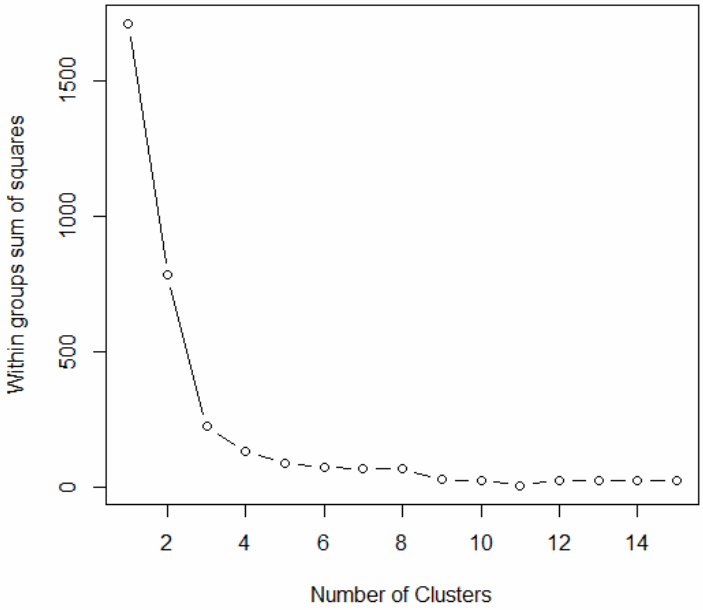

Fig. 2. The plot of the sum of squared error against number of clusters (scree plot)
Source: made by the author in R, using data from Table 1

The second method used is based on *gap statistic* (Tibshirani, Walther and Hastie, 2001). The Gap *statistic* plot from Figure 3 is the plot of the function values defined by the relation (3), from (Tibshirani, Walther and Hastie, 2001, p. 412), against the number of clusters k. According to Tibshirani, Walther and Hastie (2001, p. 415) the estimated number of clusters is the lowest k for which

$$Gap[k] > Gap[k+1] - s_{k+1}, \quad (2)$$

where:
$Gap[k]$ is the value of the function in point k
s_{k+1} is the standard error in point $k+1$

In a less accurate interpretation, but more intuitive, the estimated number of clusters is the lowest k from which the increase of the function Gap slows near the point $(k, Gap(k))$. Because the vertical bars in the Figure 3 represent the values of the standard errors s_k in the points k, we observe immediately that the estimated number of clusters is 3.

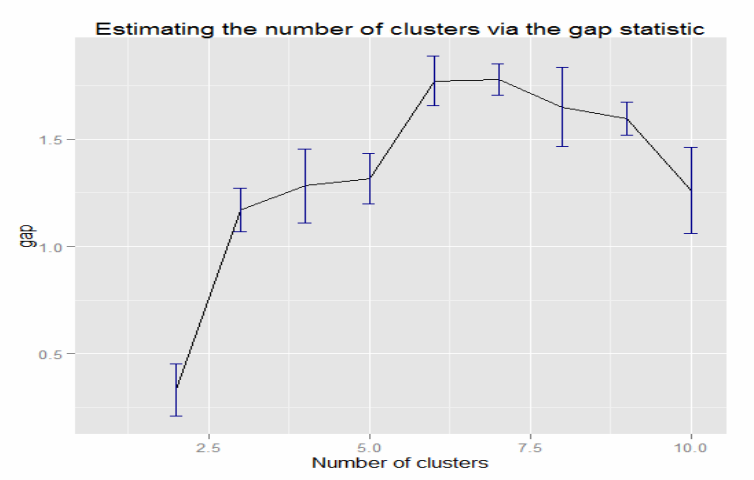

Fig. 3. The gap statistic plot

Source: made by the author in R using functions defined by Chen (2010)

In order to evaluate the quality of the structure based on three clusters it is also useful to plot the average silhouette width against number of clusters. The results are shown in Table 2.

Table 2. Average silhouette widths based on the number of clusters

Number of clusters	2	3	4	5	6	7	8	9	10	11	12	13
Average silhouette width	0.66	069	0.67	0.67	0.54	0.42	0.40	0.37	0.39	0.35	0.38	0.38

Source: made by the author with results obtained in R

We find that the maximum of these values, named Silhouette Coefficient (SC) is reached for k =3, which reinforces the idea that the number of clusters is estimated correctly (Stuyf, Hubert and Rousseeuw, p. 10). In Figure 4 we show the silhouette plot of the structure found for the number of clusters k = 3.

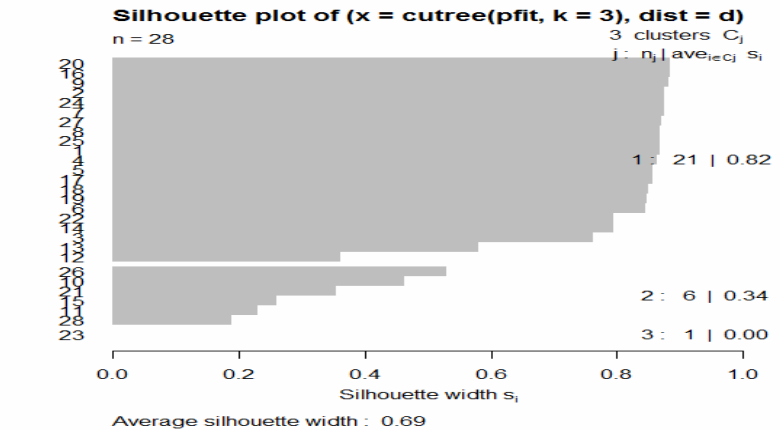

Fig. 4. Silhouette plot of the obtained structure. Source: made by the author in R

The obtained average silhouette width (0.69) shows that a "reasonable structure has been found" (Stuyf, Hubert and Rousseeuw, p. 10).

The Stability of the Clusters Structure

In Hennig's (2006, p.2) opinion "stability means that a meaningful valid cluster shouldn't disappear easily if the data set is changed in a non–essential way". Also in (2006, p.2), Hennig indicates more practical modalities in order to evaluate the stability of the clusters. In this paper we used his algorithm, implemented by the function *clusboot* from R. In this algorithm, the evaluation of the clusters stability is done by using Jaccard's coefficient. The change of initial data "in a non –essential way" is realized by using the bootstrap method (Efron and Tibshirani, 1993).

In order to easily understand the algorithm, let us denote by x_i $i = 1,2,...28$ the line vector whose components are given by the indicators values NAH and UAA from line i of Table 1. A randomized version of the initial data set, which we denote by $x^*_1, x^*_2,...x^*_{28}$ is a random sample of size 28 drawn with replacement from the initial population $x_1, x_2,...x_{28}$.

The Jaccard's coefficient of two sets A and B is defined in (Jaccard, 1912) by:

$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$, if sets A and B are nonempty and by $J(A,B) = 1$, if sets A and B are empty,

where $A \cap B$ is the intersection of sets A and B,

$A \cup B$ is the reunion of sets A and B,

$|A|$ represents the cardinal of the set A.

In principle, the method proposed by Hennig can be described by the following algorithm:

1. For the initial data set $x_1, x_2,...x_{28}$ estimate the number of clusters and choose a clustering method;
2. Determine the component of the obtained clusters C_i through the chosen method
3. Determine B (for example $B = 100$) bootstrap versions $x^*_1, x^*_2,...x^*_{28}$ of the initial data set $x_1, x_2,...x_{28}$
4. For each bootstrap version $b \in \{1,2,...B\}$:
   - 4.1 Determine the component of the clusters by the chosen method at point 1
   - 4.2 For each initial cluster C_i calculate $J\max_i^b$, as being the maximum of Jaccard's coefficients calculated for C_i and the obtained clusters in current bootstrap version b
5. Calculate the stability index of each cluster C_i as being $Stab_i = \frac{1}{B}\sum_{b=1}^{B} J\max^b{}_i$
6. Stop

From the definition of the Jaccard's coefficient we notice that it takes values between 0 and 1 and so $0 <= Stab_i <= 1$. The closer to 1 is the stability value $Stab_i$, the more stable is the cluster C_i. Based on this observation we find that the values of stability indexes presented in Table 3 reveal very good values for clusters C1 and C2 and an reasonable value for the cluster C3.

Table 3. Stability indexes for the found clusters

Cluster	C1	C2	C3
Stability index	0.98	0.86	0.66

Source: made by the author with results obtained in R

High values of stability indexes for clusters C1 and C2 indicate they are real clusters, i.e. not an artificial creation of the clustering algorithm. Cluster C3, with a reasonable stability index (0.66) does not seem to be a cluster of confidence. He appeared most likely due to the exceptional values that Romania's indicators NAH and UAA have: Romania is the country that has the largest number of land properties of all EU countries and ranks sixth in terms of utilized agricultural area.

Results and Discussions

The structure of the three clusters estimated based on the dendogram from the Figure 1 has been validated by three methods: scree plot, gap statistic and the method based on the use of silhouettes. In addition, the stability of the obtained clusters was analysed using the method proposed in (Efron and Tibshirani, 1993), which is based on the use of bootstrap technique. The three obtained clusters (highlighted in dendogram from Figure 1 by enclosing each one in a rectangle) have the following composition:

- Cluster 1: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Portugal, Slovakia, Slovenia, Sweden.
- Cluster 2: France, Germany, Italy, Poland, Spain, United Kingdom.
- Cluster 3: Romania.

Countries that belong to cluster 2 are among the top seven countries in EU in terms of utilized agricultural area (UAA). Together they own 63.9% of the total utilized agricultural area in EU and 40.1% of the number of agricultural holdings.

Countries that belong to cluster 1 own 28.6% of total utilized agricultural area in EU and 26.4% of the number of agricultural holdings.

Romania (cluster 3) has a particular situation: it owns 7.5% in terms of utilized agricultural area (UAA), ranking 6th in the EU and it has a third of EU agricultural holdings (33.5%), the largest of the EU countries.

It may be observed that the number of agricultural holdings in Romania is 7.1% higher than the number of agricultural properties for the countries in cluster 2 although the utilized agricultural area in Romania is almost four times smaller.

Conclusions

In this paper we classify EU countries in relation to two indicators: number of agricultural holdings as a percentage of the total number of agricultural properties (NAH) in EU and utilized agricultural area (UAA) as a percentage of the total utilized agricultural area in EU. The method used was hierarchical agglomerative clustering based on Ward's distance and led to finding a three cluster structure. The number of clusters was validated by using three graphical methods (scree plot, gap statistic and silhouette plot), and we concluded that a reasonable structure has been found. We also studied the stability of the obtained clusters using an algorithm based on the bootstrap method. While clusters 1 and 2 have exceptionally good stability indexes (0.98 and respectively 0.86) cluster 3 has a reasonable stability index (0.66). The 0.66 value of stability index of cluster 3 reflects the special situation of its only member, Romania: a country with a large agricultural area (sixth country in the EU), but with the highest number of agricultural holdings.

References

1. Chen, E., 2010, *Gap statistic* available at https://github.com/echen/gap-statistic/commit/1d6dc43a331d227c86a1d8349c5773efff0020e9 [accessed on 22.03.2016].
2. Efron, B. and Tibshirani, R., 1993. *An introduction to the bootstrap*, Chapman& Hall, Inc, New York, U.S.A.
3. European Commission , 2014 , *The European Union explained :Agriculture* , [online] available at http://europa.eu/pol/pdf/flipbook/en/agriculture_en.pdf [accessed on 16.03.2016].
4. Eurostat-Statistics explained, 2016a, *Farm structure statistics*.[online] available at http://ec.europa.eu/eurostat/statistics-explained/index.php/Farm_structure_statistics [accessed on 16.03.2016]
5. Eurostat-Statistics explained, 2016b, *Farm_structure_YB2016*, [online] available at https://www.google.ro/search?q=Farm_structure_YB2016+ef_kvaareg&ie=utf-8&oe=utf-8&client=firefox-b&gws_rd=cr&ei=uSJVV7OuL8bwUqTsvdAH [accessed on 16.03.2016]
6. Eurostat-Statistics explained, 2016c, *Glossary:Agricultural area (AA)* [online] available at http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Agricultural_area_(AA) [accessed on 16.03.2016]
7. Hennig, C., *Clusterr-wise assesment of cluster stability,* Research Report no. 271, Department of Statistical Science, University College London, December 2006, available at http://www.cnmd.ac.uk/statistics/research/pdfs/rr271.pdf [accessed on 12.04.2016].
8. Jaccard, P., 1912. The distribution of the flora in the alpine zone, *The new phytologist*, vol XI, no. 2 February 1912, available at http://onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.1912.tb05611.x/pdf [accessed on 14.04.2016].
9. Peeples ,M., 2011 *R Script for K-Means Cluster Analysis* [online] available at http://www.mattpeeples.net/kmeans.html [accessed on 21.03.2016].
10. Shalizy, C., 2009 *Distances between Clustering, Hierarchical Clustering* Statistics 36-350: Data Mining course available at http://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf [accessed on 17.03.2016].
11. Stuyf, A., Hubert, M. and Rousseeuw, J., Clustering in an Object-Oriented Environment. *Journal of Statistical Software*, vol.1 , number 1, (1997) available at http://www.jstatsoft.org/v01/i04/paper [accessed on 13.04.2016].
12. Theodoridis S. and Koutroumbas K., 2009. *Pattern Recognition, Academic* Press Elsevier
13. Tibshirani, R., Walther, G. and Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic, *Journal of the Royal Statistical Society B*, 63, Part 2, pp. 411-423 available at http://www.web.stanford.edu/~hastie/Papers/gap.pdf [accessed on 7.04.2016].