



Marek Sobolewski

Rzeszow University of Technology, Faculty of Management, Department of Quantitative Methods,
msobolew@prz.edu.pl

Małgorzata Markowska

Wroclaw University of Economics, Faculty of Economics, Management and Tourism, Chair
of Regional Economy, malgorzata.markowska@ue.wroc.pl

Hierarchical Clustering Methods With Territorial Integrity Criterion¹

Abstract: During the conference entitled *Spatial Econometrics and Regional Economic Analyses*, which took place in Lodz in 2014, there was a proposition to introduce the spatial coherence property into the Ward method, which is applied to group administrative units [1]. At each stage of agglomeration in the modified Ward method, there are included only those aggregates which are adjacent to each other. This work is an extension of this concept based upon other methods of hierarchical clustering, in particular the single and complete linkage method. The study highlighted the benefits of clustering methods with the coherence property, also emphasizing the limitations of these procedures. First of all, the introduction of the restricting condition during the procedure of the hierarchical clustering reduces the homogeneity of isolated clusters. Spatial constraints may also lead to a situation where the distance between the clusters linked at a later stage is smaller than at an earlier stage (graphically, we can talk about the dendrogram “backflow”). The method of complete linkage is free of these aberration where the distance between the clusters is defined as the maximum distance between their elements. The modified clustering algorithm was implemented as an extension of *STATISTICA* software. Examples of an application of the hierarchical clustering method with the coherence concern sector changes in the European regional space. The aim of the analysis was to isolate spatially coherent areas that demonstrate a similar direction and intensity of structural change in selected areas of the labour market.

Keywords: hierarchical clustering, territorial integrity

JEL: C38, O18, R23

¹ Supported by National Science Centre (Poland) grant 2015/17/B/HS4/01021.

1. Introduction

Taxonomic methods are related to the most primary cognitive need of human beings, i.e. arranging our surrounding reality into greater order. From this perspective almost every data analysis can be referred to as “taxonomic”, however, in terms of statistics, this concept is used with reference to clustering procedures and the ordering of multivariate data. Taxonomic methods are applied by researchers and practitioners representing many scientific disciplines. They are particularly common in the research in the area of social sciences – various classifications of administrative units are very popular with regards to residents’ quality of life or level. However, the application of classical procedures for administrative units’ clustering most frequently results in territorial divisions that lack integrity, which significantly impedes the interpretation of research outcomes. In practice, the areas featuring territorial integrity are far easier to manage. A certain conception for solving the problem of cluster incoherence for Ward’s method was suggested in 2014 at the conference entitled *Spatial Econometrics and Regional Economic Analyses* held in Łódź (Markowska, Sobolewski, 2014). In 2016 the analogical modification of *k*-means algorithm (Sobolewski, Sokołowski, 2016) was proposed at the SKAD [The Section of Classification and Data Analysis of Polish Statistical Association] conference in Bełchatów. The present study discusses the concept of coherent clustering using various hierarchical agglomerative clustering methods. Certain characteristics of adjusted algorithms were indicated and an assessment of the quality of clustering was conducted using modified methods against the results achieved by means of conventional algorithms. Importantly, the suggested methods were implemented as an extension of *STATISTICA* software, making them available to interested readers for their own practical application.²

2. Algorithms of hierarchical clustering

Hierarchical clustering methods (and more specifically hierarchical agglomerative clustering, as opposed to division methods) are described in detail in numerous publications (Grabiński, 1992), therefore the present study touches upon them only very briefly. Basically, the algorithm of hierarchical clustering can be presented in three points:

1. Each object creates a separate cluster in the initial partition.
2. The distances between clusters are determined [in the space of diagnostic variables] using an appropriate “clustering” convention (see Table 1).

² The developed application shall be provided to all interested persons without any limitations via e-mail.

3. The most adjacent clusters are linked – if their number is greater than 1 we return to point 2.

In the final stage of clustering, one set including all elements is obtained – obviously this result (similarly to the initial situation) is not of particular interest for a researcher. The ultimate division (or alternative divisions) is performed at a certain stage of hierarchical clustering by identifying either an arbitrarily set, or the most distinctive, number of clusters.

The algorithms of hierarchical clustering differ in their method of defining distance between clusters, which in some situations has a significant impact on the research outcomes (Table 1). Ward’s method is considered the most effective one since it is based on the idea of minimizing the intra-cluster variance and comes as the result of many simulation studies.

Table 1. Determining the distance between clusters for the selected hierarchical methods

Method	Distance between clusters <i>A</i> and <i>B</i>
Single-linkage clustering	$\min_{x \in A, y \in B} d(x, y)$
Complete-linkage clustering	$\max_{x \in A, y \in B} d(x, y)$
Average linkage clustering	$\frac{\sum_{x \in A, y \in B} d(x, y)}{N_A \cdot N_B}$
Centroid linkage clustering	$d(\bar{x}_A, \bar{y}_B)$
Ward’s minimum variance method	$\frac{d(\bar{x}_A, \bar{y}_B) \cdot N_A \cdot N_B}{N_A + N_B}$

A, B – clusters, *x, y* – elements of clusters, *d(x, y)* – distance between two points.

Source: own elaboration

It should be emphasized that the clustering results clearly depend on the method for determining the distance between objects in the space of diagnostic variables (Euclidean distance is most frequently applied in this case). Due to size limitations, the presented study discusses the detailed results of conventional clustering and clustering with coherence criteria for two methods: Ward’s and complete-linkage clustering methods. It seems that by introducing the coherence criterion to the algorithm of clustering (discussed in more detail in the next point) the impact of the clustering method selection on the obtained results is even larger. Moreover, after introducing the limiting criteria to the algorithm of clustering, the distances between subsequent clusters do not have to follow a non-decreasing sequence.

3. The introduction of territorial integrity criterion in the clustering procedure

The algorithm of hierarchical clustering with coherence criterion is developed by imposing certain limitations on an original algorithm. In order to present the simplicity of the discussed concept, three stages of clustering with coherence criterion are described below. They differ from the conventional method by the introduction of one condition:

1. In the initial division, each object (administrative unit) creates a separate cluster.
2. The distances between clusters are determined using an appropriate “clustering” convention.
3. The most proximate geographically adjacent clusters are linked – if their number is greater than 1 we return to point 1.

The introduction of coherence criterion to the algorithm of hierarchical clustering (in general terms – the limiting criterion) can result in the phenomenon of dendrogram “fall back”. This situation is illustrated in Figure 1, based on the example of the classification of provinces against the labour market situation in the period 1999–2012.

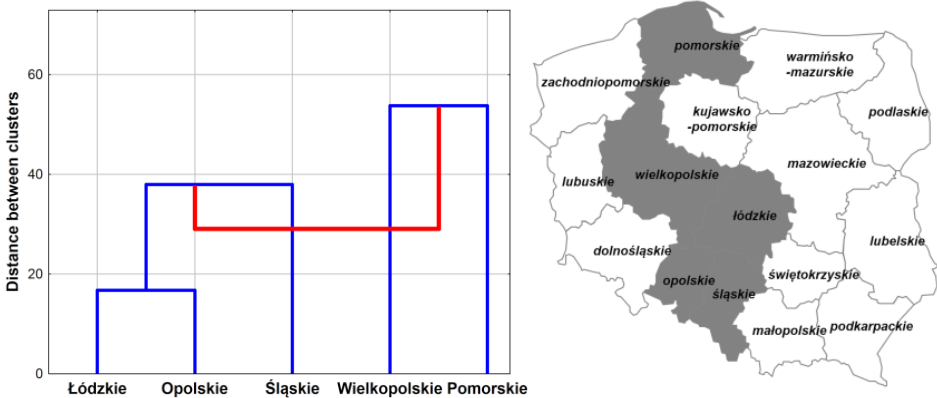


Figure 1. The phenomenon of dendrogram “regression” in Ward’s method with coherence criterion

Source: own elaboration

In the course of subsequent clustering stages, Ward’s method with coherence criterion was used to link the following provinces: Łódzkie and Opolskie, to be followed by Śląskie. A separate cluster was created by the neighbouring Pomorskie and Śląskie provinces. In the fourth agglomeration stage the dendrogram “regression” was observed – the distance between clusters identified in stage 3

was smaller than in the two previous linkage cases. The reason of such aberration is obvious – Pomorskie province remains very similar to Silesian province with regards to diagnostic variables, whereas their linkage, along with maintaining territorial integrity, was possible only after the complete linkage with Łódzkie and Wielkopolskie provinces.

The complete linkage method is characterized by its resilience to dendrogram “fall back” occurrence, which results from the method of defining distance between clusters as the distance between the two most distant objects.

4. The description of computer software for clustering with territorial integrity criterion

The methods of coherent hierarchical clustering were implemented as an extension of *STATISTICA* software. The algorithm of clustering with coherence criterion is relatively complex since at each stage only the adjacent clusters are linked, which requires the spatial relationships of objects to be analysed each time based on the adjacency matrix. After the two clusters are linked, both matrices have to be transformed, i.e. economic distance and adjacency matrices.

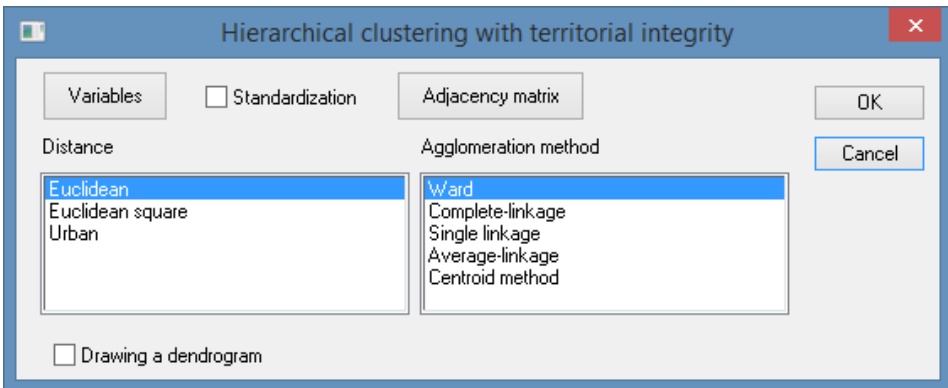


Figure 2. The software interface

Source: own elaboration

Figure 2 presents the input window of the software for coherent clustering. The application that was developed allows to:

- 1) identify the file with adjacency matrix used in clustering;
- 2) determine the linkage method;
- 3) specify the type of distance;
- 4) select the adjacency matrix;

- 5) define the number of clusters identified based on the conducted analysis (it is possible to provide several values and the divisions corresponding to the provided figures will be saved in the input sheet data in separate columns).

If the *Dendrogram drawing* option is selected, then, prior to determining the number of clusters, the diagram of clustering is displayed, which allows for “interactive” determining of the clusters’ number based on its interface.

5. Proposed procedure as applied to European regional space

The algorithm of hierarchical clustering with coherence criterion was applied in the European Union regions’ clustering at NUTS–2 level with regards to unemployment levels in the period 1999–2012. The practical objective of the conducted analysis was to identify the spatially coherent areas characterised by a similar level and direction of changes in the labour market situation to help fight unemployment in the spatially homogeneous EU areas.

The number of analysed units was 269. The clustering of EU regions was conducted based on unemployment rate values in the period 1999–2012, therefore 14 diagnostic variables were analysed. Due to the fact that the range of values of diagnostic variables was comparable, and their measurement was performed using the same units, they were not subject to standardization.

The conducted analysis applied two selected hierarchical clustering methods – Ward’s and complete-linkage clustering methods. The distance between objects was determined based on the squared Euclidean distance.³

Eight clusters were identified and their number was determined based on the dendrogram for the conventional Ward’s method. To make the comparison of results easy the same number of clusters was used in case of the other applied clustering methods.

The method for determining adjacency matrix used in clustering methods with coherence criterion requires a separate discussion. Based on the digital coordinates of the EU regions (available on the Eurostat website), the matrix of common land borders length between regions was determined. On this basis the adjacency matrix was defined and adjusted to take into account close relationships between regions which do not have common land borders. For the purposes of the analysis the following regions were considered “adjacent”: e.g. British and French regions located on both sides of the English Channel, the islands of Sardinia and

³ It should be pointed out that in case of complete-linkage clustering method it is irrelevant whether the Euclidean distance or its square is selected since the monotonic transformation of distance matrix does not influence the clustering results using this algorithm.

Corsica with the regions of southern France, Sicily with the regions of southern Italy, and Malta with Sicily. The same was applied to the regions located on both sides of the Danish straits and the Greek islands, forming separate regions. It is generally known that the subject literature on clustering quality measures and the consistency of alternative divisions is abundant. It was decided, however, to apply the popular coefficient of determination in the present study. It is useful in determining the percentage of unemployment rate variability in the EU regional space, which was explained by the division made. The value of the R^2 coefficient presented below is averaged over all diagnostic variables (which, as a reminder, represented unemployment rates in the years 1999–2012).

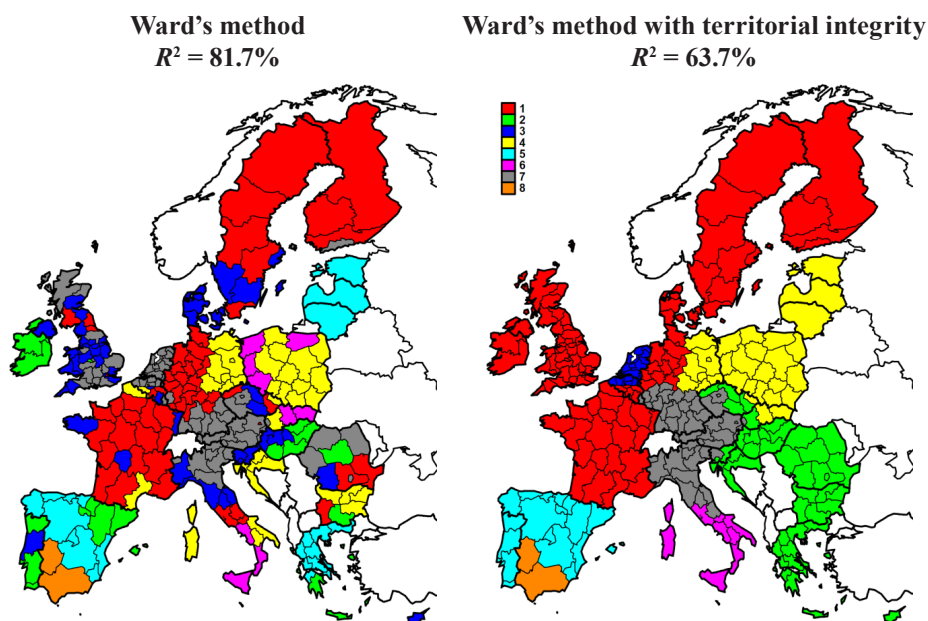


Figure 3. The results of clustering using the conventional Ward's method and Ward's method with territorial integrity

Source: own elaboration

Figure 3 presents the clustering results of the European Union regions in terms of unemployment rate, in the years 1999–2012, obtained using the conventional Ward's method and Ward's method with territorial integrity. From the cognitive viewpoint, the spatially consistent results are much easier to describe and to derive practical actions from (obviously, on condition that the division based on territorial integrity remains homogeneous in terms of the identified clusters). The coefficient of determination for the division maintaining territorial integrity is lower

from the results of the conventional division (81.7 vs. 63.7%), but the difference is not big. Therefore it seems that the proposal of the division based on territorial integrity has significant practical advantages which cannot be offered by the results achieved based on the conventional Ward's method.

Table 2. The characteristics of clusters obtained using Ward's method with territorial integrity

Cluster	Average unemployment rate in the identified groups													
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1	8.3	7.2	6.3	6.5	6.7	7.0	7.1	7.0	6.5	6.3	8.0	8.2	8.0	8.1
2	10.0	10.0	10.1	10.1	9.5	9.7	9.3	8.7	7.7	7.1	8.7	10.6	12.1	14.2
3	5.3	3.9	3.2	4.0	4.6	5.2	5.7	5.0	4.4	3.9	4.6	5.3	5.0	5.7
4	14.5	16.8	18.2	18.8	18.6	18.9	17.5	14.4	11.1	9.2	10.8	11.6	10.8	10.7
5	9.9	8.4	6.6	7.4	8.0	7.9	7.4	7.1	6.9	8.4	12.9	14.5	16.6	19.6
6	19.6	18.3	16.9	16.2	16.0	12.8	12.1	10.9	9.9	10.4	10.7	11.2	11.4	14.0
7	5.6	5.0	4.4	4.8	5.2	5.8	5.9	5.4	4.6	4.3	5.2	5.1	4.7	5.1
8	22.7	21.9	17.1	16.7	14.7	15.7	15.0	14.4	15.0	17.7	23.0	25.5	27.8	33.5

Source: own calculations

The description of clusters identified using the division based on territorial integrity is presented in Table 2. For example, clusters no. 3 and 7, covering respectively Benelux and the Alpine regions of Germany, Italy and Austria represent areas featuring low unemployment rate in the entire discussed period. Cluster no. 5 covers Spanish and Portuguese regions characterized by a relatively low unemployment rate till the beginning of the financial crisis and the very difficult labour market situation that followed 2008. Extremely high unemployment rates can be observed in two Spanish regions included in cluster no. 8. A different situation occurs in the central and southern parts of Italy (cluster 6), where the unemployment rate was very high at the beginning of the analysed period, it dropped later to go up again to a relatively high level, however, not so drastically as in Spain, as a consequence of the financial crisis. Poland belongs to cluster no. 4 characterised by territorial integrity along with Lithuania, Latvia and Estonia, the regions of eastern Germany and northern Slovakia.

In case of the complete linkage method the division quality defined by the coefficient of determination is 78.5% in the case of a conventional division and 63.3% for the division with territorial integrity. While analyzing the obtained divisions in terms of their substantive usefulness it should be observed that the complete linkage method with territorial integrity resulted in a more diversified division in terms of cluster size (picture 5). On the one hand, the areas particularly affected by the crisis are well separated (Spain, Portugal, southern Italy and Greece), however on the other, two one-element clusters were created (Brussels and Malta). Furthermore, a large part of the European Union was not diversified at all in terms

of unemployment rate in the period 1999–2012 (the cluster marked in red covers as many as 185 regions, i.e. more than two thirds of all regions!).

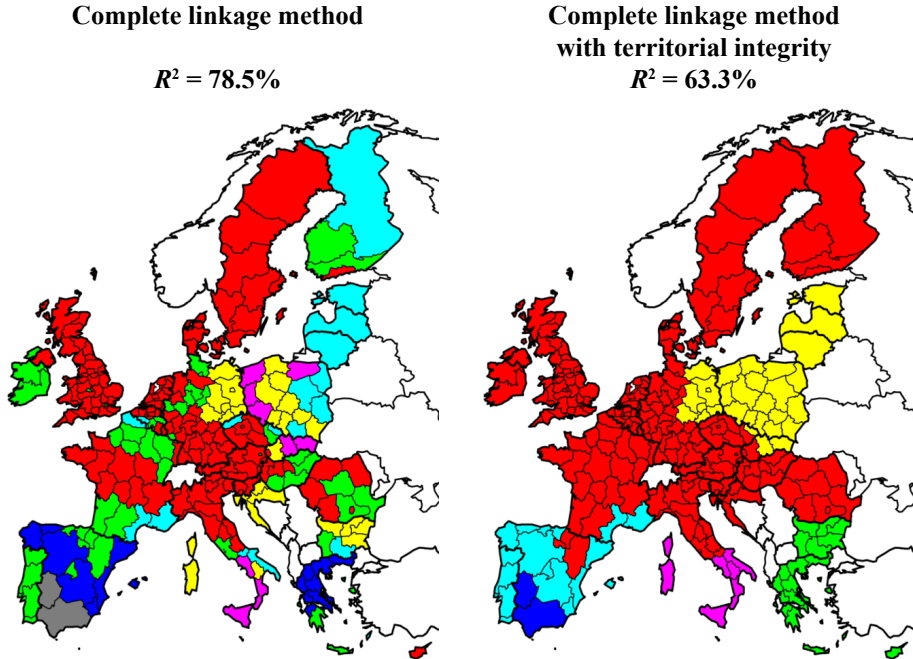


Figure 4. The results of clustering using the conventional complete linkage method and complete linkage method with territorial integrity

Source: own elaboration

6. Conclusions

The modification of the algorithm of hierarchical clustering presented in the study allows obtaining the spatially coherent divisions of administrative units according to the values of diagnostic variables. The suggested algorithm presents an interesting alternative for the conventional clustering methods, however, it should be borne in mind that by introducing the coherence criterion the spatial (territorial) picture of clustering is “improved”, but the division distinctiveness in the space of diagnostic variables deteriorates. If, however, the decline in division quality is insignificant, it seems that the division based on territorial integrity is worth suggesting, since such areas are easier to manage.

The study highlights the fact that the distances between clusters linked following coherent clustering are arranged non-monotonically – which was defined as the phenomenon of dendrogram “regression”. It was observed, that the complete linkage method remains an exception here.

The discussed tool for coherent clustering can become the subject of interesting generalizations. First of all, the application of other matrices can be noticed, alleviating the criterion of territorial integrity – for example the second degree adjacency, maximum distance criteria. The coherence criterion “fuzziness”, by taking the common borders’ length into account, also seems a reasonable concept. The links between regions with a longer common border would be preferred if such an approach were to be taken.

References


- Grabiński T. (1992), *Metody taksonometrii*, Cracow University of Economics Press, Cracow.
- Markowska M., Sobolewski M. (2014), *Wrażliwość regionalnych rynków pracy Unii Europejskiej na kryzys ekonomiczny. Klasyfikacja metodą Warda z warunkiem spójności*, „Acta Universitatis Lodzianensis. Folia Oeconomica”, no. 308, pp. 79–94.
- Sobolewski M., Sokołowski A. (2016), *Algorytm grupowania metodą k-średnich z warunkiem spójności*, 30th Taxonomic Conference “Data classification and analysis – theory and applications”, The book of abstracts, Bełchatów.

Metody grupowania hierarchicznego z warunkiem spójności terytorialnej

Streszczenie: Na konferencji „Ekonometria przestrzenna i regionalne analizy ekonomiczne”, która odbyła się w Łodzi w 2014 roku, zaproponowano wprowadzenie warunku spójności przestrzennej do metody Warda, stosowanej do grupowania jednostek administracyjnych. Na każdym etapie aglomeracji w zmodyfikowanej metodzie Warda uwzględniane są tylko te skupienia, które ze sobą sąsiadują. Niniejszy artykuł stanowi rozszerzenie tej koncepcji na inne metody grupowania hierarchicznego, w szczególności metodę prostych i zupełnych połączeń. Zwrócono uwagę na korzyści płynące z zastosowania metody grupowania z warunkiem spójności, akcentując jednak także pewne ograniczenia tych procedur. Wprowadzenie warunku ograniczającego podczas procedury grupowania hierarchicznego powoduje przede wszystkim zmniejszenie jednorodności wyodrębnianych skupień. Ograniczenia przestrzenne mogą też prowadzić do sytuacji, kiedy odległość między skupieniami łączonymi na późniejszym etapie jest mniejsza niż na etapie wcześniejszym (można tu mówić o graficznym „cofaniu się” dendrogramu). Od tej aberracji wolna jest metoda zupełnych połączeń, gdzie odległość między skupieniami jest wyznaczana jako maksimum odległości między ich elementami. Zmodyfikowany algorytm grupowania zaimplementowano jako rozszerzenie programu *STATISTICA*. Przykłady zastosowania metod grupowania hierarchicznego z warunkiem spójności dotyczą europejskiej przestrzeni regionalnej (w układzie NUTS-2) w latach poprzedzających i następujących po kryzysie finansowym 2008 roku. Celem analiz było wyodrębnienie spójnych przestrzennie obszarów, które charakteryzowałyby się podobną wrażliwością na zjawiska kryzysowe w obszarze rynku pracy.

Słowa kluczowe: grupowanie hierarchiczne, spójność terytorialna

JEL: C38, O18, R23

	© by the author, licensee Łódź University – Łódź University Press, Łódź, Poland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license CC-BY (http://creativecommons.org/licenses/by/3.0/)
	Received: 2016-12-28; verified: 2017-02-08. Accepted: 2017-08-03