Constrained clustering of winter precipitation in Greece

Eftychia Rousi[1], Christina Anagnostopoulou[1], Angelos Mimis[2], Marianthi Stamou[2]

[1]*Department of Meteorology and Climatology, Aristotle University of Thessaloniki, Thessaloniki, 54124, Greece, erousi@geo.auth.gr, chanag@geo.auth.gr*
[2]*Department of Economic and Regional Development, Panteion University of Social and Political Science, 136 Sygrou Av., Athens, 17671, Greece, mimis@panteion.gr, marianthi.stamou@panteion.g r*

Abstract
The aim of this paper is an objective clustering of winter precipitation in Greece. The methodology adopted contains a hierarchical constrained clustering technique, in which the data are aggregated into zones so that each area is assigned to only one zone and that the zones are contiguous internally and externally. In that way the internal connectivity of the zones is preserved and the homogeneity of the data within each region is optimized. The climatic parameter of precipitation was chosen because of its complex and highly spatial-dependent nature. The data consists of winter daily precipitation values obtained from a Regional Climate Model, the RACMO2/KNMI, for a period of 30 years, 1971-2000. The constrained clustering method is implemented by using 3 different linkages, single, complete and average, and for three different cluster numbers, 10, 20 and 30. The comparison of the results shows that the optimal linkage method is the complete, while 20 clusters seem to be good enough for the aims of the study.
Keywords: Constrained Clustering, Precipitation, Greece, Regional Climate Model

Introduction
The problem of defining the various climate zones has a wide range of uses (Iyigun et al., 2013). These include the redefinition of climate zones and rainfall regimes as a result of ongoing climate changes while at the same time examining the reasons that lead to those changes. Also these have a direct effect to hydrology and flora. So the regional water management as well the farming strategies are affected.
In this context, the famous classification system of Köppen – Geiger has emerged, which was originally published by Köppen in 1918. This system provides a set of rules applied to variables derived from long term values for temperature and precipitation. In these, with several rules at hand, various locations are classified into climate types (Cannon, 2012). This rule based approach has been adopted and extended by various researchers, as for example by Thornthwaite who by following manual classifications projected the various locations into climate regions which exhibit climate homogeneity.
With the widespread use of personal computers a different approach has emerged. In this, climate classification is performed by clustering algorithms based on the assumption that areas with similar values of variables characterizing climate, such as temperature or precipitation, can be classified in the same climate type. In this way, the climate types are directly defined by the data. In this methodology, usually a two step approach is adopted. Firstly, a principal component analysis (PCA), followed by clustering analysis (CA) (Fovell and Fovell, 1993; Cannon, 2012).
PCA is a transformation method finding projections of maximum variability. Thus, the n principal components contain the best n dimensional view of the data, helping us to explore the structure of the data. Therefore, having used the PCA with the multivariate data under hand, one can apply any of the CA techniques available to classify the areas into climate types. The most common method of CA in the literature is the agglomerative hierarchical clustering, which starts with n clusters and in each step merges the pair with the minimum separation distance ending into one cluster containing all the regions. Several decisions have to be made in using this CA, such as the number of clusters and the way the new distance is recomputed after a merge (resulting into various methods such as the complete linkage). The major drawback of this approach is that it explores a limited region of the solution space.

The described methodology has been applied to various studies targeting specific parts of the world or even for a whole continent (Cannon, 2012). In one of the first studies, Fovell and Fovell (1993) examined the climate zones of the USA by using temperature and precipitation data, consisting by 24 variables on lattices, for a 50 years period. Various numbers of clusters were illustrated and the shortcomings of this approach were the lack of robustness and the use of only land data. Similar studies exist for Turkey (Iyigun et al., 2013), Italy (Di Giuseppe et al., 2013) and Ireland (Pawitan and Huang, 2003). In the first study, Iyigun et al. (2013) used data for temperature, precipitation and humidity from 244 stations, for a period of 40 years. The 14 clusters concluded, captures the climate of Turkey in a realistic way. A functional clustering approach was adopted by Di Giuseppe et al. (2013) that used precipitation and temperature values, for a 30 years period, from 95 monitoring stations. They combined interpolation by penalized B-Splines and the k-medoid algorithm for classification. In that way, a small number of coefficients have captured the variability of the temporal pattern and their results have been compared with two step PCA and CA procedure.

The approach described treats the spatial problem of climate zones in an aspatial way, meaning that an area is included into a class regardless of its location. As a result, the existence of small patches of different classes within regions of a specific climate type that can be seen in Fovell and Fovell (1993) and an attempt to treat this was proposed by Pawitan and Huang (2003). In this paper, simulated rainfall data for 37 years and 1299 stations was used. This approach extended the previous ones by using a contiguous constraint on the clustering techniques and was compared to the classic hierarchical clustering methods. In this methodology, the areas with data (polygons) are aggregated into zones so that each area is assigned to only one zone and that the zones are contiguous internally and externally. In that way the internal connectivity of the zones is preserved, while merging the areas with the most similar characteristics and soretaining the homogeneity of the data within each region (Guo, 2008). Since Pawitan and Huang (2003) proposed a methodological approach several shortcomings were present. First the size constraint, in the hierarchical algorithm, was imposed indirectly, making its application inefficient for big data and secondly, their results were not compared to actual climate regions of Ireland.

In our approach, we are extending the preliminary study of Pawitan and Huang (2003) by applying a robust constrained hierarchical algorithm and examine the validity of the results in the wider area of Greece. Further we test the methodology in a regular grid of points and most importantly, as Fovell and Fovell (1993) encourages, data are not limited to the main land but climate characteristics in the Mediterranean Sea are also considered.

In order to examine our proposed methodology, the wider area of Greece is studied and classified into climate zones. The data consist of winter daily precipitation simulated values obtained from a Regional Climate Model, called RACMO2/KNMI, for a period of 30 years, 1971-2000. The results of the constrained clustering method are compared to those of current approaches adopted in literature.

Materials and Methods

The data used in this study are provided by the Regional Climate Model KNMI-RACMO2 of the Royal Netherlands Meteorological Institute (van Meijgaard et al., 2008). The model has a spatial resolution of 25km x 25km and it is driven by the General Circulation Model ECHAM5/MPI. A broader area around Greece has been chosen, composed by 1064 grid points. The data consist of winter daily precipitation values over Greece for the 30year period 1971-2000.

The problem of regionalization is computational intensive due to the contiguity constraint that is imposed. So, a number of methods have been developed over the years to tackle the problems that appear in a wide range of applications, ranging from political districting to map generalization. Early attempts to describe the methods devised have been made by Murtagh (1985) and Gordon (1996), but a full taxonomy has been defined by Duque et al. (2007). In these, the methods are divided into two classes, the algorithms that satisfy directly or indirectly the constraint. Here, the contiguity that is required to be satisfied in most of the

cases is not the only constraint considered, but the size or the composition of the regions as well.
The first class of methods is subdivided into three main categories. In the first, there are the exact optimization models, where n areas are aggregated into m spatially contiguous regions, while a predefined criterion is optimized. In the second, the heuristic models are included, which are used since 1960 and they are highly efficient, especially when a large number of areas is aggregated. The heuristic technique models that are included:
a) are based on the graph theory,
b) start with an initial solution and search for an improvement,
c) start with a seed and adds areas into regions without breaking the constraint,
d) are based on the hierarchical clustering adopted to capture the constraint.
In the third category, the mixed heuristic models are also included, which are a combination of the previous two.
In our case a 'flavor' of agglomerative hierarchical clustering is used (Mojena 1975; Murtagh 1985). In this method, at the beginning, each area is considered as a region and at each step (iteration) the closest regions, in terms of a given metric, are merged till only one region is left. Different techniques can be devised by following different strategies to define the new distance between the newly merged and the other regions. So, by merging region Ri and Rj , the newly defined region $R_i \cup R_j$will have distance in respect with the other regions Rk given by the Lance and Williams formula (1967):

$$d(R_i \cup R_j, R_k) = a_i \cdot d(R_i, R_k) + a_j \cdot d(R_j, R_k) + b \cdot d(R_i, R_j) + c \cdot |d(R_i, R_k) - d(R_j, R_k)| \quad (1)$$

where a_i, a_j , b and c are parameters whose values depend on the method (Mojena, 1975; Gordon 1996). So, for example, the average linkage is given for $a_i = a_j = 1/2$ and $b = c = 0$. A full description of the various methods available can be found in any standard textbook (e.g. Everitt et al. , 2001).
Having described the hierarchical clustering techniques, we have to incorporate explicitly the spatial contiguity constraint. For that purpose a "Sorted Dictionary" structure has been adopted where we only keep the pair of regions that are contiguous and are sorted by their relative distance. Following this approach, at each step of the algorithm, we know the pair of regions having the minimum distance while satisfying the contiguity constraint. After removing from the structure the pair with minimum distance, a new region is created by merging them and the structure is updated accordingly (by using Eq. 1).
In our implementation, the single, average and complete linkages are used by adopting the Lance and Williams formula. In the single linkage, the distance of the newly merged region $R_i \cup R_j$ with Rk is defined as the minimum distance between areas Ri ,Rj and Rk as seen in the following formula:

$$d(R_i \cup R_j, R_k) = \min(d(R_i, R_k), d(R_j, R_k)) \quad (2)$$

On the other hand, average linkage is the mean of all the distances among the areas of the regions and the complete linkage is given by the maximum distance.
It has been developed in Python 2.7 which can be easily imported as a script in ArcGIS.

Results and Discussion

Winter precipitation regime in Greece is governed by both topography and atmospheric circulation. Mediterranean depressions, mainly originating from the Gulfs of Genoa and Sidra, with their south-southwesterly flow and in combination with dynamic instability, cause high precipitation amounts on the windward regions of the western Greece mountain range and eastern Aegean Sea (Metaxas and Kallos, 1982). This is not the case for the lee areas of the eastern mainland, Crete and the southern Aegean islands (Cyclades), which are not significantly affected by these depressions (Bartzokas et al., 2003). Generally, precipitation decreases from the western to the central and eastern parts of the mainland and to the Aegean Sea, whereas it increases again in the islands of the eastern Aegean Sea and the western coasts of Asia Minor (Hatzianastasiou et al., 2008).

Winter daily precipitation values, as simulated by the KNMI model, are presented in the map of figure 1. The model succeeds in capturing the main characteristics of the precipitation regime in Greece. In more detail, as seen on the map of figure 1, the role of the Pindus mountain range is fundamental and the highest amounts of precipitation (around 600mm) in Greece occur on its west side. On the contrary, areas on the east side of Pindus mountains are much drier (100-200mm), since they are exposed to dry katabatic winds. One can readily identify the Olympus mountain that exhibits much higher precipitation amounts than its surroundings. Northern Greece has various precipitation regions, ranging from low to high precipitation amounts. For example, Thessaloniki area has an average of 150mm of winter rain, while areas in eastern Macedonia and western Thrace are characterized by higher values, reaching 450mm. The Aegean Sea is characterized by the lowest precipitation amounts, ranging from 88mm over the drier parts, to 250mm in its eastern part, where islands such as Lesvos, Chios and Samos have higher precipitation amounts. Crete island is divided in three parts, with its west part being the wettest and its eastern part the driest. These are the main characteristics of winter precipitation regime in Greece for the period 1971-2000, according to the KNMI model.

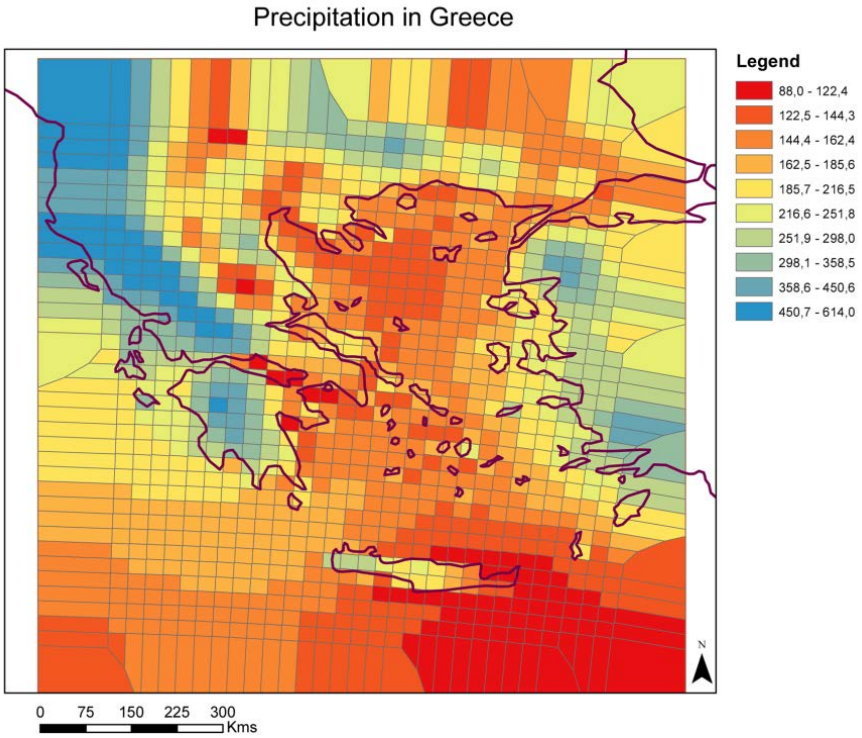

Figure 1. Choropleth map of winter daily precipitation values (in mm) in Greece based on KNMI simulated data for the period 1971-2000.

On the next step, the maps that resulted from the implementation of the hierarchical constrained clustering method on the same data are presented. The daily precipitation simulated values are point data and in order to use them in the analysis, a Dirichlet tessellation was applied to define the neighborhood structure.
Figure 2 presents the maps for 10, 20 and 30 clusters based on single linkage constrained clustering. It is obvious that this kind of linkage does not provide good results for the precipitation regime in Greece. Although the number of clusters is 10, 20 and 30, just a few of them are represented in the maps. Especially the map of 10 clusters fails almost completely to capture any of the precipitation variability. Maps of 20 and 30 clusters only succeed in capturing the precipitation variability around Pindus mountain range.

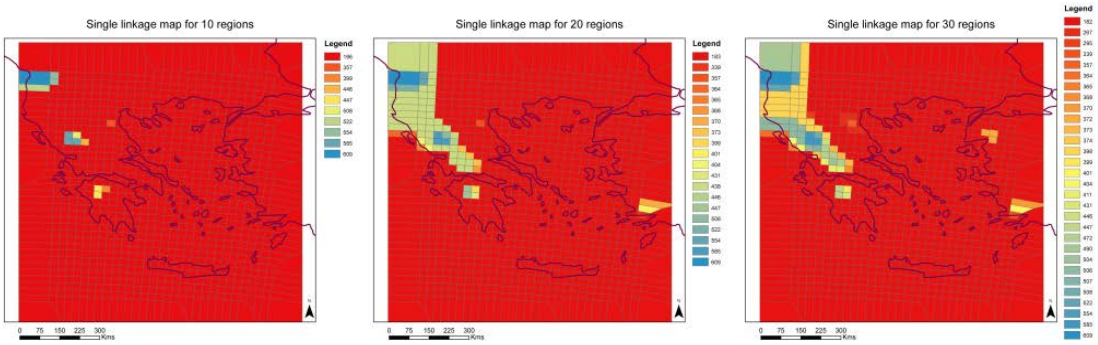

Figure 2. Single linkage constrained clustering of winter daily precipitation values in Greece for 10, 20 and 30 clusters. The average precipitation (in mm) of all grid points grouped in each cluster is shown in the legend.

Results of the average linkage for 10, 20 and 30 clusters are presented in the three maps of figure 3. This linkage provides a much better classification of the precipitation in Greece, especially for 20 and 30 clusters. In these corresponding two maps, the methodology captures not only the high precipitation amounts in the windward areas of Pindus mountains, but it also distinguishes the Olympus peak, the eastern Macedonia-western Thrace region and the increase of precipitation in the eastern Aegean and Asia Minor. The 10 cluster classification is very different than the 20 and 30 ones and it is obviously not as good. On the contrary, there are no significant differences between the latter ones. This probably indicates that 20 clusters are enough in this case, since we do not gain significantly more detail with the use of 30.

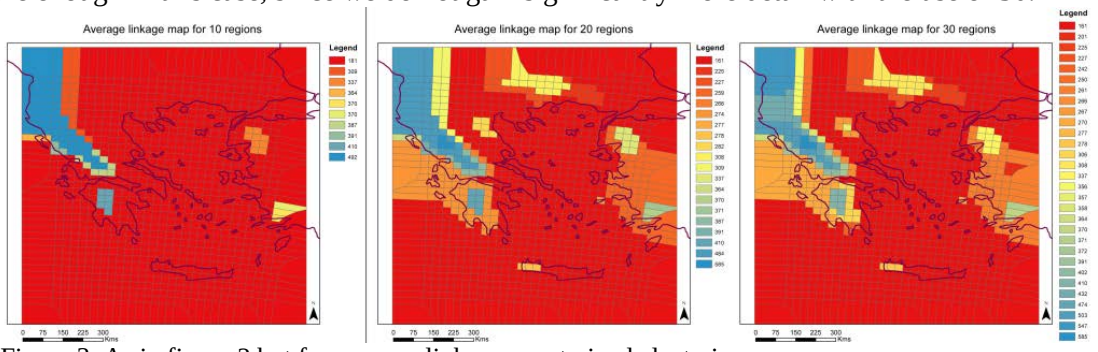

Figure 3. As in figure 2 but for average linkage constrained clustering.

Finally, the complete linkage was tested, again for the three different number of clusters and the results are shown in the maps of figure 4. This method seems to be the best choice, since the regions provided show the greater resemblance to the spatial distribution of the precipitation as it is presented in the choropleth map of figure 1. In this case, even the 10 cluster solution gives a good image, although, 20 and 30 clusters are clearly more representative. As in the average linkage, 20 and 30 clusters do not have significant differences, even though they are a bit larger here. Still, a 20 cluster solution seems to be the optimum choice, in order to maintain an amount of regions that is easy to be handled and analyzed. Table 1 presents the characteristics of the 20 clusters/regions shown in the central map of figure 4. In particular, each cluster is described by the number of grid points that are grouped in it, the average precipitation of all grid points grouped in it (same as the precipitation amounts in the corresponding map legend), and the minimum and maximum precipitation in it. As it is obvious in the map as well, there is a small number of clusters that contains the majority of the grid points. The largest region is cluster number 3, containing more than half of the grid points (equal to 592), having an average precipitation of 150mm and covering a great part of the Aegean Sea, one of the driest regions in Greece. Following, clusters 5 and 6 contain around 100 grid points each and are characterized by 208mm and 234mm of average winter precipitation respectively. Cluster number 5 covers the eastern Aegean and western Turkey, while cluster number 6 is found in northern and central Greece, covering part of Macedonia and Thessaly regions. Every other cluster represents a number of

grid points lower than 70. There are even some cluster containing less than 10 grid points (Tab. 1). The case is about very small areas that have significantly different amount or precipitation compared to their surrounding, so they cannot be grouped with them, such as the Olympus mountain, or the mountainous region in central Peloponnese.

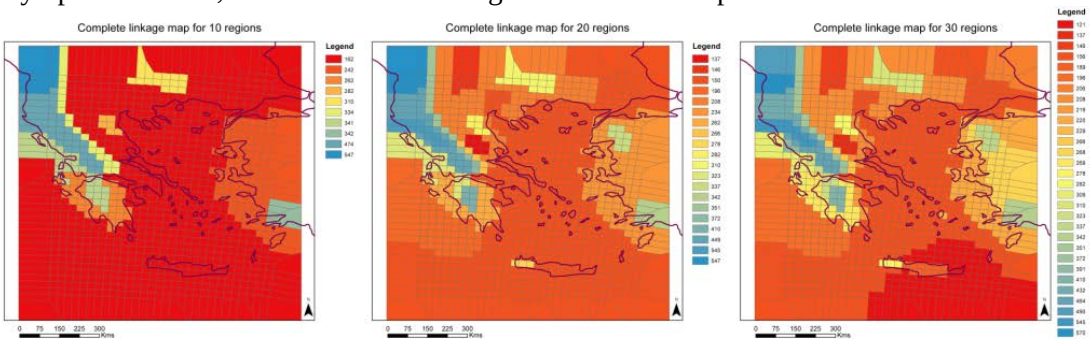

Figure 4. As in figure 2 but for complete linkage constrained clustering.

Table 1. Characteristics of the 20 cluster complete linkage solution.

cluster	number of grid points	average precipitation (mm)	minimum precipitation (mm)	maximum precipitation (mm)
1	7	137	101	162
2	10	146	119	169
3	592	150	88	232
4	69	196	169	257
5	102	208	158	279
6	112	234	176	312
7	38	262	201	329
8	5	266	231	292
9	3	278	270	286
10	6	282	254	357
11	16	310	283	368
12	26	323	271	405
13	9	337	313	383
14	5	342	298	402
15	11	351	295	410
16	4	372	350	397
17	7	410	363	456
18	26	449	373	508
19	9	545	506	589
20	7	547	476	613

Conclusions

A hierarchical constrained clustering was performed in order to classify winter precipitation in Greece. The main conclusions are:

- Three different methods of linkage, single linkage, average linkage and complete linkage were applied on winter daily Greek precipitation data.
- Of the three different linkage methods tested, the single linkage was the one that performed poorly, in accordance with Hands and Everitt (1987), who made a comparative study of 5 hierarchical clustering techniques. This is probably due to the nature of the precipitation data. According to Ferreira and Hitchcock (2009) that also compared different clustering methods, the superiority of one method over others was

not uniform, but rather depended greatly on the form of the data. Nevertheless, single linkage seems to be the worst choice in most cases.

- Average and complete linkage both performed well, with the latter proving to be more detailed and its spatial resolution presents many similarities to the original data.
- The number of clusters (zones) was tested, too. The 10 cluster solution is clearly not suitable for the precipitation data since it does not capture any of the data variability. On the contrary, 20 clusters could be considered good while 30 clusters do not add any value to the classification results. According to Fovell and Fovell (1993), cluster analysis represents a compromise between specificity and generality. Each merger unavoidably results in loss of precision and detail, but this is justifiable as long as the ability to interpret and generalize is enhanced. So, this is what a researcher should have in mind when defining the number of clusters.
- Finally, the results demonstrate that the complete linkage method with 20 clusters could be the best choice for clustering modeled precipitation data in Greece. Future work could test the hierarchical constrained clustering on a combination of meteorological parameters that affect Greek climate.

References

Bartzokas A., Lolis C.J. and Metaxas D.A., 2003. The 850 hPa relative vorticity centres of action for winter precipitation in the Greek area. International Journal of Climatology, 23, 813-828.

Cannon A.J., 2012. Köppen versus the computer: comparing Köppen-Geiger and multivariate regression tree climate classifications in terms of climate homogeneity. Hydrology and Earth Systems Science, 16, 217–229.

Di Giuseppe E., Lasinio G.J., Esposito S. and Pasqui M., 2013. Functional clustering for Italian climate zones identification. Theoretical and Applied Climatology, 114, 39-54.

Duque J.C., Ramos, R. and Surinach J., 2007. Supervised Regionalization Methods: A Survey. International Regional Science Review, 30(3), 195-220.

Everitt, B.S., Landau, S. and Leese M., 2001. Cluster Analysis, Fourth edition, Arnold, London.

Ferreira L. and Hitchcock B.D., 2009. A comparison of hierarchical methods for clustering functional data. Communications in Statistics - Simulation and Computation, 38:9, 1925-1949.

Fovell R.G. and Fovell M-Y C., 1993. Climate zones of the conterminous United States Defined using cluster analysis. Journal of Climate, 6, 2103-2135.

Gordon A.D. 1996. A survey of constrained clustering. Computational Statistics and Data Analysis, 21, 17-29.

Guo D.,2008. Regionalization with dynamically constrained agglomerative clustering and partitioning(REDCAP). International Journal of Geographical Information Science, 22, 7, 801-823.

Hands S. and Everitt B., (1987). A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques. Multivariate Behavioral Research, 22, 235–243.

Hatzianastassiou N., Katsoulis B., Pnevmatikos J. and Antakis V., 2008. Spatial and temporal variation of precipitation in Greece and surrounding regions based on global precipitation climatology project data. Journal of Climate, 21, 1349-1370.

Iyigun C., Turkes M., Batmaz I, Yozgatligil C., Purutcuoglu V., Koc E.K. and Ozturk M.Z., 2013. Clustering current climate regions of Turkey by using a multivariate statistical model. 114, 95-106.

Lance G.N. and Williams W.T., 1967. A general theory of classificatory sorting strategies, I. Hierarchical Systems. The Computer Journal, 9, 373-380.

Metaxas DA, Kallos G., 1982. High rainfall amounts over W Greek mainland during December and January. In Proceedings of the 1st Hellenic–British Climatological Meeting, Hellenic Meteorological Society, Athens, 5–11 September, 1980; 125–137.

Mojena R., 1975. Hierarchical grouping methods and stopping rules: An evaluation. The Computer Journal, 20(4), 359-363.
Murtagh F., 1985. A Survey of Algorithms for Contiguity-constrained Clustering and Related Problems. The Computer Journal, 28(1), 82-88.
Pawitan Y. and Huang J., 2003. Constrained clustering of irregularly sampled spatial data. Journal of Statistical Computation and Simulation, 73:12, 853-865.
van Meijgaard E et al., 2008. The KNMI regional atmospheric climate model RACMO version 2.1. Tech. Rep. TR-302, Royal Netherlands Meteorological Institute