Determining the Climate Zones of Turkey

by Center-Based Clustering Methods

Fidan M. Fahmi[*], Elçin Kartal[*], Cem İyigün[**], Murat Türkeş[***], Ceylan Yozgatlıgil[*], Vilda Purutcuoğlu[*], İnci Batmaz[*], Gülser Köksal[**]

[*] *Department of Statistics, Middle East Technical University, Ankara,Turkey,(e-mail: fidantelaferli@yahoo.com, kartalelcin@metu.edu.tr, ceylan@metu.edu.tr, vpurutcu@metu.edu.tr, ibatmaz@metu.edu.tr)*

[**] *Department of Industrial Engineering, Middle East Technical University, Ankara, Turkey, (e-mail:iyigun@ie.metu.edu.tr, koksal@ie.metu.edu.tr)*

[***] *Department of Geography, Onsekiz Mart University, Çanakkale, Turkey,(e-mail: murat.turkes@comu.edu.tr)*

Abstract: There is a growing evidence that the climate change has already been significant impacts on the world's physical, biological and human systems, and it is expected that these impacts will become more severe in the near future. Alterations in the weather patterns and the existence of extreme events can be considered as important indicators of this change. The validity of this reality can be judged by analyzing the climate data thoroughly. In this study, for determining the climate zones of Turkey, temperature measures obtained from the Turkish State Meteorological Service stations in the time period 1950-2006 are examined by two center-based clustering techniques which are k-means and fuzzy k-means. The clusters obtained from these methods are compared using objective criteria. They are also evaluated subjectively by the domain experts.

Keywords: Climate change, k-means, fuzzy k-means, subjective evaluation

1. INTRODUCTION

The increase in atmospheric concentrations of so-called greenhouse gasses such as carbon dioxide (CO_2) and methane (CH_4) has resulted in greenhouse effect and global warming. It is not surprising to see the effect of these changes also in Turkey. Recent analyses of Turkey's climate data reveal that minimum and maximum temperatures are increasing (Türkeş, 1996). On the contrary, observed temporal distribution patterns of winter rainfalls are significantly changing while moderate decreases in the total amount of rainfall have also been recorded. As a result of global warming, alterations in the weather patterns and the existence of extreme events can be considered as important indicators of the change in climate zones. To reduce the future vulnerability, it is important to examine the changes and new climate zones (if any) so that appropriate strategies can be developed and precautions can be taken accordingly.

In determining the climate zones, different methods and variables have been used in the literature. The most popular one is the classification method called Koeppen and Thorntwaite (Türkeş, 1996.) This method is a quantitative method that directly determines the climate types but the rules used in classification are subjectively extracted. Therefore, relatively objective clustering approaches can be applied instead for the same purpose (e.g. Ünal et al., 2003.) To exemplify, in the study of Ünal et al.'s, Ward's method is proposed to be used as a hierarchical clustering technique.

On the other hand, in literature, temperature and precipitation are treated as the main research variables in this kind of studies because of their influence on the distribution of flora and human activities. In our study, temperature measures are obtained from the Turkish State Meteorological Service stations in the time period 1950-2006 are examined by using two different center-based clustering techniques which are k-means and fuzzy k-means.

2. DATA

The data used in identifying the climate zones of Turkey are temperature measures including minimum and maximum temperature, average temperature, and average of maximum and minimum temperature, minimum and maximum of average temperature. This data set in which

values recored monthly in the time period 1950-2006 are obtained from the Turkish State Meteorological Service stations. Although 270 stations provide climate data, only 65 stations are used within 1950-2006 time period. The main reason of using this small number of stations is that stations with missing values for more than two successive years are removed from the analysis. Mean imputation is applied for the missing values recorded for months within a year. For each temperature variables, the number of stations (i.e. sample size) used are listed in Table 1.

3. THE METHODOLOGY

Clustering is the partitioning of a data set into subsets in such a way that the data in each subset share some common trait-often proximity-according to a distance measure. Assuming our data $D = \{x_{ij}\}$, where $i = 1,2,...,n, j = 1,2,...,p$, consist of p features measured on n independent observations. For a given data set, D and an integer K, $1 \le K \le N$, the clustering problem is to partition a data set into K disjoint clusters. In other words, $D = C_1 \cup C_2 \cup ... \cup C_K$, with $C_j \cap C_k = \phi$ if $j \ne k$ and $C_j \ne \phi$ for $j \in \{1,...,K\}$. Thus, each cluster is consisting of points that are similar in some sense, and points of different clusters are dissimilar. Here, similarity means to be close in the sense of a distance d(**x**, **y**) between any two points **x**, **y** $\in R^p$. Note here that the number of clusters denoted by K should be determined beforehand. However, determining the "right" number of clusters is an important issue to be dealt with in clustering. Also note that different distance measures such as Euclidean distance, Manhattan distance and Mahalanobis distance can be used to measure the similarity between any two points **x, y** $\in R^p$.

There are many different clustering methods available in the literature. In this study, a well-known and easily applied ones, called center-based clustering techniques, were applied as a first attempt.

3.1 Center-based Clustering Methods

Center-based clustering algorithms construct clusters by using the distances of data points from the cluster centers. The best-known and most commonly used center-based algorithm is the k-means algorithm (Hartigan,1975), which minimizes the objective

$$\sum_{k=1}^{K} \sum_{x_i \in C_k} \left\| x_i - c_k \right\|^2 , \qquad (1)$$

where c_k is the centroid of the k^{th} cluster. The steps of the algorithm is as follows:

Step 0: Initialization: Given the data set *D* and an integer K, $2 \le K \le N$, select *K* initial centers $\{\mathbf{c}_k\}$.
Step 1: Compute the distances d($\mathbf{x}_i$, $\mathbf{c}_k$), i=1,...,N; k=1,...,K, between the data points and cluster centers, and partition the data set by assigning each data point to the cluster whose center is the nearest.
Step 2: Recompute the cluster centers as the cluster mean points.
Step 3: If the centers have not changed, stop; else go to Step 1.

A major challenge in cluster analysis is to determine the correct or natural number of clusters. It may be possible to provide insight into the number of clusters in the data by using the classical "elbow method". Fig.1 shows a typical graph of evaluation measure (SSE for our case) for k-means clustering applied for minimum temperature versus the number of clusters employed: the SSE (2) decreases monotonically as the number of clusters increases. We can try to find the natural number of clusters in a data set by looking for the number of clusters at which there is a knee in the plot (Tan et al., 2003).

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in C_k} d^2(c_i, x) , \qquad (2)$$

where c_i represents the center of the i^{th} cluster.

In Fig.1 there is a distinct knee in the SSE for k-means clustering applied for the minimum temperature when the number of cluster is equal to 4. For the other variables, the optimal number of clusters listed in Tab.1 are determined by using the same approach.

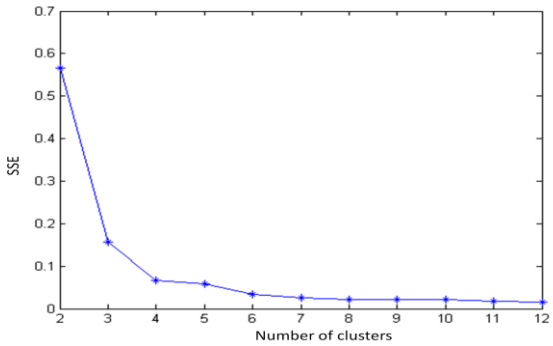

Fig. 1. Plot of SSE versus the number of clusters, K, for the minimum temperature.

Table 1. Variables studied and corresponding number of clusters and sample size

Temperature Variables	Number of Clusters	Sample Size
Average	4	61
Minimum	4	65
Maximum	5	65
Average Minimum	4	65
Average Maximum	4	65
Minimum Average	4	65
Maximum Average	5	65

Several variants of k-means algorithm have been reported in the literature (Anderberg, 1973). Some of them attempt to select a good initial partition so that the algorithm is

more likely to find the global minimum value. The k-means algorithm can be adapted to soft clustering. A well-known center-based algorithm for soft clustering is the so-called fuzzy k-means algorithm in which the objective function to be minimized is

$$\sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik}^{m} d_{ik}^{2} = \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik}^{m} \left\| x_i - v_k \right\|^2 \quad . \qquad (3)$$

Here, u_{ik}^{m} are the membership functions of $x_i \in C_k$, and typically satisfy (4); m>1 is a real number known as fuzzifier.

$$\sum_{k=1}^{K} u(x, C_k) = 1, \text{ and } u(x, C_k) \geq 0 \text{ for all } k = 1,..., K. \qquad (4)$$

The same optimal number of clusters which were determined for k-means clustering method by using elbow approach (in Tab.1) is used for fuzzy k-means clustering; therefore, the assignment of the stations to the clusters can be compared for both methods. The clustering results obtained by using both methods are displayed in graphs like in Fig. 2 and Fig. 3. By using these graphs, clusters are examined, and the stations that fall into different clusters from its neigbour are determined, and the reasons are discussed. Additionally, all of the graphs obtained for each variable are compared. Clustering methods were also applied to the multivariate data set containing all temperature variables. The optimal number of clusters for this application is found to be five by using the elbow method for the k-means clustering approach and this number of cluster is also used for fuzzy k-means clustering method. As a result, the graphs obtained by using k-means and fuzzy k-means clustering methods are displayed in Fig. 2 and Fig. 3, respectively.

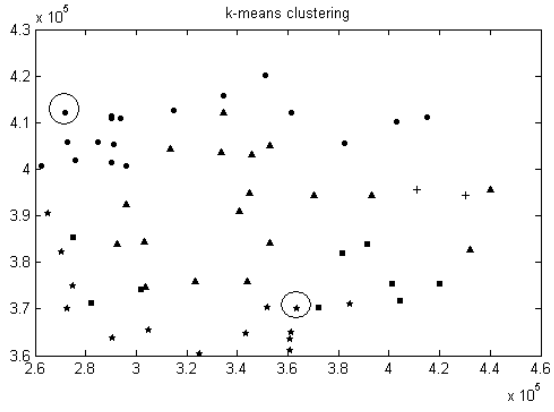

Fig. 2. k-means clustering graph.

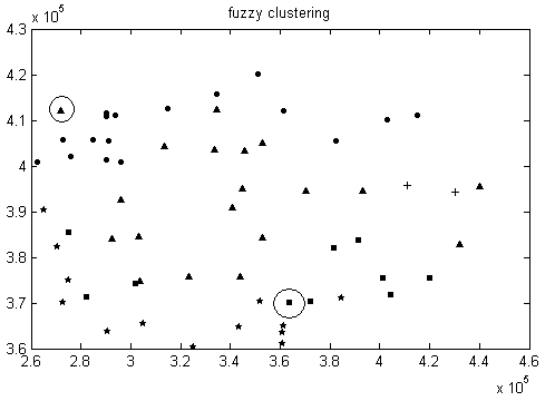

Fig. 3. Fuzzy k-means clustering graph.

In both graphs, similar partitioning are obtained except two stations indicated by circles. This in not surprising, however, since the membership values of these two stations (i.e. Lüleburgaz and İslahiye) for two clusters are very close to each others.

4. CONCLUSION AND FUTURE STUDIES

There have been seven recognized climate zones of Turkey for many years. These include Black sea, Marmara, Aegean, Mediterranean, central Anatolian, eastern Anatolian and southeastern Anatolian regions. Note here that in addition to climate differences, social and economic variables are also considered in identifying these zones. In this study, however, the climate zones of Turkey are re-examined by using a matematical methodology of center-based clustering analysis. As a result of applying k-means and fuzzy k-means clustering methods on Turkish data, different clusterings are obtained for each temperature variable. Moreover, when the multivariate cluster analysis is applied, five clusters are obtained over Turkey, and the partition of regions are found similar by two methods studied, except two stations. As can be seen on the graph, the Aegean region has the same properties as the Mediterranean and Southeastern Anatolian regions. In addition, in Marmara region, Black Sea's climate effect become dominant, except Trakya station, which is similar to central Anatolian region. Moreover, two stations have different climate characteristics within Eastern Anatolian.

Because the clustering of the temperature variables does not consider the spatial properties of the stations, resulting clusters may not represent the correct partitioning. To increase the efficiency of the method, bisecting k-means which is less susceptible to initialization problems will be applied. This method is an extension of the k-means algorithm that is based on simple idea: to obtain *K* clusters, split the set of all points into two clusters, select one of these clusters to split, and so on, until *K* clusters have been produced (Tan et al., 2003). Actually, our main goal is to cluster the stations by considering their spatio and temporal properties. In this way, as an initial steps, bisecting and hierarchical clustering will be applied, and compared with k-means and fuzzy k-means results.

REFERENCES

Anderberg, M.R., *Cluster Analysis for Cluster Applications*. Academic Press, Inc., New York, NY, 1973.

Hartigan, J. *Clustering Algorithms*. Wiley and Sons, Inc., New York, N.Y., 1975.

Tan,P.N.,Steinbach,M.,Kumar,V. (2006). *Introduction to Data Mining*. Pearson Education, Inc.,Boston.

Türkeş, M. 1996. Spatial and temporal analysis of annual rainfall variations in Turkey. *International Journal of Climatology*, 16, 1057–1076.

Türkeş, M. 1996. Observed changes in maximum and minimum temperatures in Turkey. *International Journal of Climatology*, 16, 463-477.

Ünal, Y., Kindap, T. ve Karaca, M., 2003. Redefining the climate zones of Turkey using cluster analysis. *International Journal of Climatology*, 23, 1045-1055.