Research Article **Open Access**

Huseyin Zahit Selvi* and Burak Caglar

Using cluster analysis methods for multivariate mapping of traffic accidents

https://doi.org/10.1515/geo-2018-0060
Received January 25, 2018; accepted August 30, 2018

Abstract: Many factors affect the occurrence of traffic accidents. The classification and mapping of the different attributes of the resulting accident are important for the prevention of accidents. Multivariate mapping is the visual exploration of multiple attributes using a map or data reduction technique. More than one attribute can be visually explored and symbolized using numerous statistical classification systems or data reduction techniques. In this sense, clustering analysis methods can be used for multivariate mapping. This study aims to compare the multivariate maps produced by the K-means method, K-medoids method, and Agglomerative and Divisive Hierarchical Clustering (AGNES) method, which among clustering analysis methods, with real data. The results from the study will suggest which clustering methods should be preferred in terms of multivariate mapping. The results show that the K-medoids method is more appropriate in terms of clustering success. Moreover, the aim is to reveal spatial similarities in traffic accidents according to the results of traffic accidents that occur in different years. For this aim, multivariate maps created from traffic accident data of two different years in Turkey are used. The methods are compared, and the use of the maps produced with these methods for risk management and planning is discussed. Analysis of the maps reveals significant similarities for both years.

Keywords: traffic accidents; multivariate mapping; data mining; cluster analysis; visualization

1 Introduction

Casualties, injuries, and financial damages as a result of traffic accidents are among the most important problems of the world and also in Turkey. The increase in motor-vehicle ownership is very high in Turkey with 1,272,589 new vehicles registered just in 2015 [1]. When data of Turkey for the last 5 years are analysed, it is seen that there have been more than 1,000,000 traffic accidents, 145,000 of them ended up with death and injury, and nearly 1,060,000 of them resulted in financial damage. Due to these accidents 4000 people lose their life on average and nearly 250,000 people are injured.

Many factors such as driving mistakes, the number of vehicles, lack of infrastructure etc. influence the occurrence of traffic accidents. Many studies were conducted to determine the effects of various factors on traffic accidents. In this context: Bil et al. [2] used Kernel Density Estimation (KDE) to determine animal-vehicle collision hotspots; Lord and Mannering [3] assessed statistical analysis methods for crash-frequency data; Lin et al. [4] utilized M5P tree and hazard-based duration model for predicting urban freeway traffic accident durations; Yalcin and Duzgun [5] performed spatial analysis of two-wheeled vehicles traffic crashes; Erdogan [6] determined road mortality with geographically weighted regression analysis; Shi et al. [7] analysed traffic flow under accidents on highways using temporal data mining; and, Akoz and Karsligil [8] classified traffic events at intersections by using support vector machines and K-nearest neighbourhood algorithms. Clustering methods were also used in many studies in the spatial analysis of traffic accidents. Accident durations [9], driver risks [10, 11], risk factors affecting fatal bus accident severity [12], and rain-related fatal crashes [13] were examined using cluster-based analyses and Dogru and Subasi [14] compared clustering techniques for traffic accident detection. It is quite important to determine the similarity of traffic accidents by using more than one of the current traffic accident's attributes in order to detect precautions for traffic security. Mapping of traffic accidents according with common properties is also significant for estimating types and effects of future accidents. Geographic Information System (GIS) Software provides an opportunity to design thematic maps for this aim. Using GIS: Nokhandan et al. [15] studied how environmental factors impacted road accident frequency; Jackson and Sharif [13] researched

***Corresponding Author: Huseyin Zahit Selvi:** Necmettin Erbakan University Konya, Turkey, E-mail: hzselvi@konya.edu.tr; hzselvi@yahoo.com
Burak Caglar: Necmettin Erbakan University Konya, Turkey

rain-related fatal crashes and their correlation with rainfall; Erdogan et al. [16] and Erdogan [6] analysed traffic accidents statistics; and, Yalcin and Duzgun [5] analysed two-wheeled vehicles traffic crashes. In these studies, traffic accidents were classified according to one or two attributes, different thematic maps were designed using this classification and traffic accidents were analysed based on thematic maps. Unlike them, more than two attributes can be shown with multivariate maps. A thematic map that represents multiple related attributes is called a multivariate map. Multivariate mapping can be defined as a combination of more than one thematic map. Details of this topic are provided in section 2. Multivariate maps of occurred traffic accidents are more effective for determining common properties of traffic accidents and planning traffic systems. A classification method based on the clustering method in data mining can be used in order to represent multiple attributes in the same map [17, 18]. Unlike the above-mentioned studies, this study aims to classify traffic accidents by using clustering methods and to show them in multivariate maps. Thus, it aims to reveal the common features that cannot be determined in the classification made by considering one or two features of traffic accidents. It also aims to compare the multivariate maps produced by hierarchical and non-hierarchical clustering methods with real data, and to suggest which clustering methods should be preferred in terms of multivariate mapping. In this context, K-means, K-medoids, and Agglomerative and Divisive Hierarchical Clustering Methods, which among clustering methods, are examined in this study. These methods are used to extract profiles of traffic accidents in Turkey. Result maps produced with data of two different years are compared and it is revealed which one of these methods can be preferred in the sense of multivariate mapping.

This paper is divided into four sections. Following the introduction, the next section provides a brief overview of the multivariate mapping and clustering analysis methods used. Then, a detailed presentation of creating the multivariate maps is given. Finally, results and suggestions are shared in last section.

2 Material and Methods

2.1 Multivariate Mapping

Multivariate mapping is the graphic display of more than one variable or attribute of geographic phenomena. The simultaneous display of multiple features and their respective multivariate attributes allows for estimation of the degree or spatial pattern of cross-correlation between attributes. Multivariate mapping integrates computational, visual, and cartographic methods to develop a visual approach for exploring and understanding spatiotemporal and multivariate patterns [19].

A fundamental issue in multivariate mapping is whether individual maps are shown for each attribute or whether all attributes are displayed on the same map ([20] p.327). Producing separate maps for each attribute can make it difficult to compare two objects which have various attributes. Therefore, methods in which various attributes are shown in the same map are preferred. In this sense: a Trivariate Choropleth Map, which is created by overlapping two coloured choropleth maps [21–24]; the Multivariate Dot Maps method, in which a specific colour or symbol is used for each attribute in the map [25]; Multivariate Point Symbol methods, which are used when multivariate data can be shown with point symbols [26–33]; a method in which different types of symbols are combined is used to represent multivariate data [34]; and, a method of separating different attributes from integral symbols [35] can be listed.

Unlike the methods given above, in order to represent many attributes in the same map, a classification method based on a clustering method in data mining can be used as well ([20] p.344, [17, 18]). With the use of clustering methods, similar aspects of different spatial objects can be revealed by considering more than one attribute. In this sense, spatial analyses that would make important contributions for risk analysis, planning, etc. can be done.

2.2 Cluster Analysis

Cluster analysis is the process of grouping information in a data set according to specific proximity criteria. Similarity of elements in the same cluster should be high, similarity between clusters should be low [36]. In the process of classification, classes are determined before. In the clustering method, classes are not determined before. Data are divided into different classes according to the similarity of data.

Cluster methods are classified in different ways in various resources. In a general sense, cluster methods can be classified as hierarchical and non-hierarchical [20].

Non-hierarchical Methods: In non-hierarchical methods, n objects are divided into k clusters according to the k number ($k<n$) given before. This method divides data in a way such that there is at least one object in

each cluster and each object is included in at least in one cluster [37].

Hierarchical Methods: The hierarchical clustering method groups data objects in a tree structure. Hierarchical clustering methods are classified as agglomerative or divisive according to their hierarchical division being bottom-up or top-down [38].

In this study, the K-means method and K-medoids method, from non-hierarchical methods, and the Agglomerative and Divisive Hierarchical Clustering method, from hierarchical methods, are analysed.

2.3 K-Means Method

This algorithm, which was introduced by Mac Queen for the first time in 1967, is a cyclical algorithm in which clusters are continuously renewed until the most suitable solution is attained. The general logic of the K-means algorithm is to divide a data set composed of n data objects to k clusters determined depending on preliminary information and the experience of the researcher. The aim is for the intracluster similarity to be high, but the intercluster similarity to be low. Similarity of clusters is calculated with the mean value of objects.

The K-means procedure is summarized as below [37]:

Input:
k: the number of clusters,
D: a data set containing n objects.
Output: A set of k clusters.
Method:
arbitrarily choose k objects from D as the initial cluster centres;
(re)assign each object to the cluster to which the object is the most similar,
based on the distance between the object and the cluster mean;
update the cluster means, i.e., calculate the mean value of the objects for
each cluster;
repeat until there are no changes.

2.4 K- Medoids Method

In this algorithm, which was developed by Kauffman and Rousseeuw in 1990, instead of the mean value of each cluster, an object in each cluster is taken as a representative. This representative object, called a medoid, is meant to be the most centrally located object within the cluster [37]. After k medoids chosen for k-clusters are determined, each remaining object is clustered with the representative object to which it is the most similar [36].

Steps of K-medoids algorithm are summarized as follow [36]:

1. Determinate of k-cluster number.
2. Choice of k objects as initial medoids.
3. Assign the remaining objects to a cluster which has the most similar x medoid.
4. Calculate aim function (sum of distances of all objects to the closest medoid).
5. Arbitrary choice of y point which is not a medoid.
6. If change of x and y minimize the aim function, change the place of these two points (x and y).
7. The process is repeated between the $3^{rd}$and 6^{th} step until there is no change.

2.5 Agglomerative and Divisive Hierarchical Clustering Method

AGNES (AGglomerative NESting) follows the bottom-up strategy. In the beginning, each object is accepted as a separate cluster. In each step of the algorithm, similar clusters are agglomerated until they are a single cluster, or they enable expected properties. Most of the hierarchical clustering methods are included in this category. On the other hand, DIANA (DIvisive ANAlysis) follows the top-down strategy. In the beginning, the entire data object is accepted as one cluster. In each step of the algorithm, the most similar objects are merged together; a large cluster is divided into smaller clusters. This clustering method continues until each object composes a cluster on its own or any expected condition is enabled [38].

A tree structure named as a dendrogram is used for stating the process of hierarchical clustering. The dendrogram shows how objects are grouped step by step (Figure 1).

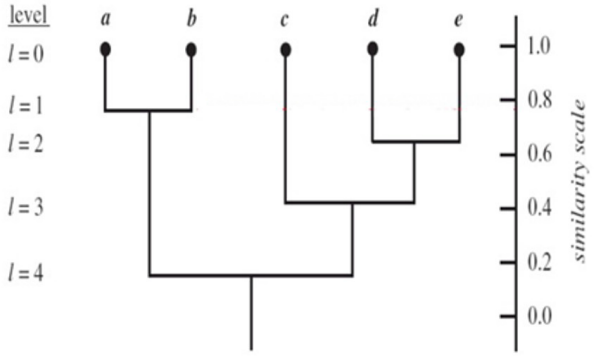

Figure 1: Dendrogram for Hierarchical Clustering of {a,b,c,d,e}

Although the hierarchical clustering method is considered to be simple, there are some difficulties in choosing agglomeration or division points. Choice of these points is quite important because further steps are carried out as new clusters are formed by agglomeration or by division of an object group. It is not possible to change previous processes or objects between clusters. Therefore, not taking agglomeration or division decisions in specific steps causes the formation of low qualified clusters.

3 Application

Clustering analysis was made with three different methods by using number of motor land vehicles based on city, number of traffic accidents resulting in death and injury, number of casualties, and number of injuries (four different values) for the years 2011 and 2012 prepared by Turkish Statistical Institute (TUIK) and multivariate maps were produced according to analysis results for determining the similarity of traffic data on a city basis in Turkey. Maps designed for both years with three different methods were compared, the success of multivariate mapping and clustering each method was evaluated.

In the application of clustering analysis methods, IBM SPSS (Statistical Package for the Social Sciences) developed by IBM Company, and RapidMiner software developed in Dortmund Technology University Artificial Mind Unit by Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer were used. Multivariate maps were designed by ArcGIS software developed by ESRI group.

3.1 Adapting the Data Set for the Process

The first step in cluster analysis is to standardize the data set, if necessary. Different units and enumeration units of varying size are possible in same data set [20]. For our data set, since all of the data have the same units, standardization is only necessary for enumeration. Because the number of motor land vehicles have larger values than the other raw data. The data corresponding to the number of motor land vehicles in each city were standardized by dividing by the total number of motor land vehicles in the whole country.

The second step of cluster analysis is to check the raw data regarding their correlation to each other, because it is possible to cluster any data set, even a set of random numbers, for multiple attributes. For this aim, correlation coefficients between the raw data were calculated according to equation 1:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (1)$$

Where r is the correlation coefficient of X and Y data, X_i and Y_i are raw data values, $\bar{X}$ and $\bar{Y}$ are the mean of attributes. Correlation coefficients for 2011 data were calculated as below:

Correlation between motor land vehicles based on city and number of traffic accidents resulting in death and injury, number of casualties, and number of injuries displayed with r_1, r_2, and r_3 respectively.

$r_1 = 0.96$

$r_2 = 0.83$

$r_3 = 0.94$

Correlation coefficients for 2012 data were also calculated as below:

$r_1 = 0.95$

$r_2 = 0.88$

$r_3 = 0.93$

As provided in Romesburg [39] which indicates the desirable correlation values of 0.8 or greater, estimated results reflect the true relations between the attributes of our data. After completing these steps, clustering processes were started.

3.2 Determination of k Cluster Number

3.2.1 Determination of k cluster number for K-means and K-medoids methods

In the K-means and K-medoids methods, the number of clusters is determined by the user. However, since the number of clusters is important in data mining, it is necessary to determine the number of clusters by some tests. In this study, Dunn Validity Index and Davies-Bouldin Validity Index tests were used in determining the number of clusters.

The Dunn validity index identifies clusters which are well separated and compact. The goal is therefore to maximize the inter-cluster distance while minimizing the intra-cluster distance. The Dunn validity index (D) is defined by equation 2.

$$D = \left(\begin{matrix} min \\ 1 \le i \le n \end{matrix}\right) \left\{ \left\{ \begin{matrix} min \\ \left(\begin{smallmatrix} 1 \le j \le n \\ i = j \end{smallmatrix}\right) \end{matrix} \frac{d(c_i, c_j)}{\left(\begin{smallmatrix} max(d'(c_k)) \\ 1 \le k \le n \end{smallmatrix}\right)} \right\} \right\} \quad (2)$$

In this equation $d(c_i, c_j)$ denotes the distance between ci and cj, $max(d'(c_k)$ is the furthest distance between

points of k cluster, and n is the number of clusters. If the Dunn validity index is large, it means that compact and well separated clusters exist [34].

Similar to the Dunn validity index, the Davies-Bouldin validity index identifies clusters which are far from each other and compact. The Davies-Bouldin validity index (DB) is defined according to equation 3:

$$DB = \frac{1}{n}\sum_{i=1}^{n} max\left\{\frac{S_n(Q_i) + S_n(Q_j)}{S_n(Q_i, Q_j)}\right\} \quad (3)$$

Where $S_n(Q_i)$ is the average distance of the cluster elements to the cluster centre, and $S_n(Q_i, Q_j)$ is the distance between the two cluster centres. The low value of DB indicates that the clusters are homogeneous within themselves and the clusters are far from each other.

A high value of the Dunn validity index (D) and a low value of the Davies-Bouldin validity index (DB) indicates good clustering quality. In order to say that the value is high or low, it is necessary to cluster in at least two scenarios and calculate the index values for each scenario [34].

In this study, Dunn and Davies-Bouldin indices were calculated for k = 2, k = 3, k = 4 ..., k = 7 scenarios to determine the optimal k number for K-means and K-medoids methods. The Dunn and Davies-Bouldin indices calculated for the 2011 and 2012 data sets are shown in Table 1 and Table 2. When Table 1 and Table 2 are examined, it is considered that the k cluster number for K-means and K-medoids methods should be 4.

Table 1: Determination of k cluster number for K-means method

	2011 Data Set		2012 Data Set	
	Dunn	Davies-Bouldin	Dunn	Davies-Bouldin
k=2	0.027	0.764	0.245	0.543
k=3	0.035	0.657	0.087	0.660
k=4	0.040	0.642	0.088	0.589
k=5	0.034	0.695	0.028	0.659
k=6	0.024	0.751	0.036	0.694
k=7	0.024	0.721	0.036	0.623

3.2.2 Determination of k cluster number for AGNES method

The optimal number of clusters to be formed in the AGNES method is determined by using the tree structure called the dendrogram described in Chapter 2. The peak values on the dendrogram show the clusters. However, since the purpose of clustering is really homogeneous and different groups are formed, it is necessary to examine the dendrogram with an auxiliary axis and determine homogeneous groups. On the dendrogram, each node point intersected by the auxiliary axis shows a cluster. In this study, the number of clusters for 2011 and 2012 data was determined to be five with this method.

Table 2: Determination of k cluster number for K-medoids method

	2011 Data Set		2012 Data Set	
	Dunn	Davies-Bouldin	Dunn	Davies-Bouldin
k=2	0.005	1.064	0.008	1.250
k=3	0.009	0.756	0.017	0.769
k=4	0.009	0.746	0.025	0.775
k=5	0.023	0.776	0.008	0.807
k=6	0.017	0.875	0.008	0.943
k=7	0.006	0.911	0.008	0.841

One of the basic approaches used in determining the suitability of the clusters obtained by the AGNES method is to determine the Cophenetic Correlation Coefficient, which measures the correlation between raw resemblance coefficients and resemblance coefficients derived from the dendrogram. The Cophenetic Correlation Coefficient (CCC) is defined as in equation 4 [20].

$$CCC = \frac{\sum_{i<j}(x(i,j) - x)(t(i,j) - t)}{\sqrt{[\sum_{i<j}(x(i,j) - x)^2][\sum_{i<j}(t(i,j) - t)^2]}} \quad (4)$$

Where, $x(i,j)=|X_i-X_j|$ is the ordinary Euclidean distance between the i^{th} and j^{th} observations, $t(i,j)$ is the dendrogrammatic distance between the model points T_i and T_j, (this distance is the height of the node at which these two points are first joined together). Then, x is the average of the $x(i,j)$, and t is the average of the $t(i,j)$. Here, CCC has a value between [-1,1]. CCC values close to one indicate that clustering results are successful. In this application, when the number of clusters is taken as five, the Cophenetic Correlation Coefficient is calculated as 0.939683 and 0.943563 for 2011 and 2012 data, respectively. These results show that it is appropriate to determine the number of clusters to be five, with Romesburg [39] stating that the clustering results are acceptable when the Cophenetic value is 0.80 or greater.

3.3 Multivariate Map Design with K-means Method

RapidMiner software was used in application of the K-means method. As a result of tests made in this sense, the *k* cluster number was four, number of highest iteration was 100, and maximum cycle of algorithm was 35. The method was applied separately for 2011 and 2012 data. Centroid Tables of clusters generated as a result of clustering processes were given in Table 3 and Table 4.

With the help of classes obtained by using four different values (number of motor land vehicle, number of traffic accidents resulting in death and injury, number of casualties, and number of injuries) in the clustering processes, multivariate maps showing similarity of traffic accidents on city basis for Turkey were designed (Figure 2).

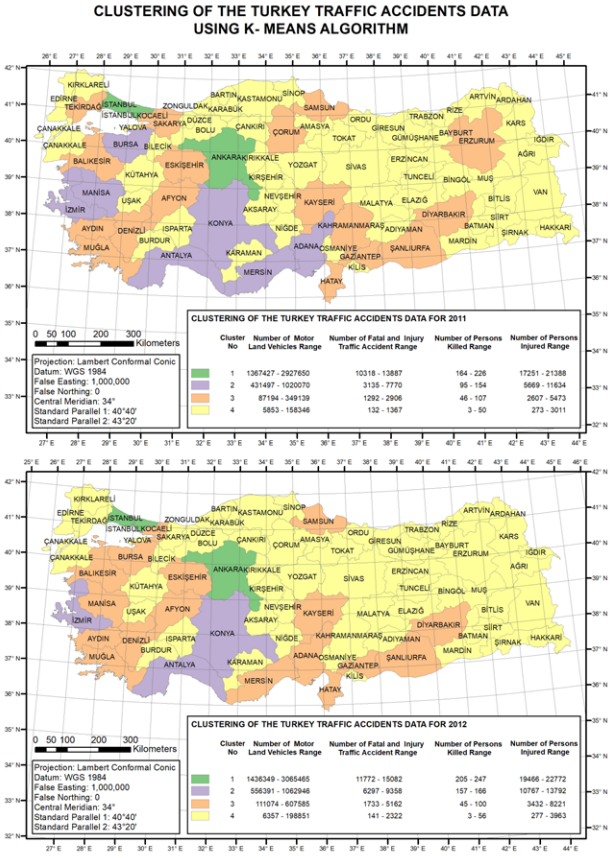

Figure 2: Multivariate maps designed with the K-means method for the years 2011 (above) and 2012 (below)

3.4 Multivariate Map Design with K-medoids Method

RapidMiner software was also used in the application of the K-medoids method. Different from the K-means algorithm, the K-medoids process operator was used instead of the K-means operator. In this scope, *k* cluster number was again taken as four, and the maximum cycle of algorithm was taken as 35. Centroid Tables of clusters generated as a result of clustering processes are given in Table 5 and Table 6.

Again, with the help of classes obtained by using four different values in clustering processes with the K-medoids method, multivariate maps were designed for the years 2011 and 2012 (Figure 3).

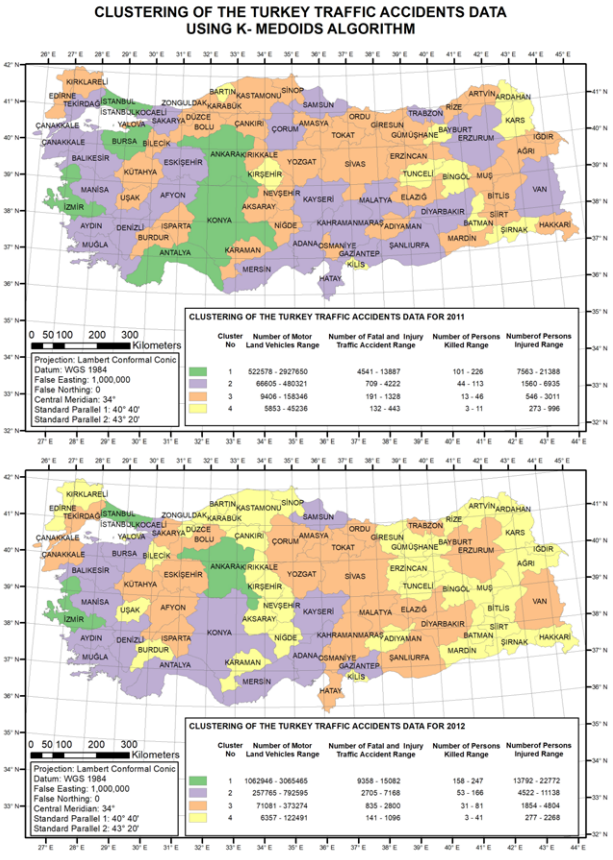

Figure 3: Multivariate maps designed with K-medoids for the years 2011 (above) and 2012 (below)

Table 3: Centroid values of clusters generated with K-means method for 2011 data

K-MEANS CENTROID TABLE (2011)				
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Number of Motor Land Vehicles	2,147,539	601,443	230,591	63,587
Number of Traffic Accidents with Death – Injury	12,103	4,940	2,100	653
Number of Deaths	195	123	71	24
Number of Injuries	19,320	8,081	3,926	1,337

Table 4: Centroid values of clusters generated with K-means method for 2012 data

K-MEANS CENTROID TABLE (2012)				
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Number of Motor Land Vehicles	2,250,907	803,977	333,858	75,324
Number of Traffic Accidents with Death – Injury	13,427	7,608	3,211	835
Number of Deaths	226	160	71	27
Number of Injuries	21,119	11,899	5,523	1,631

Table 5: Centroid values of clusters generated with K-medoids method for 2011 data

K-MEDOIDS CENTROID TABLE (2011)				
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Number of Motor Land Vehicles	1,193,364	238,043	68,183	25,096
Number of Traffic Accidents with Death – Injury	7,896	2,144	697	297
Number of Deaths	176	71	26	7
Number of Injuries	16,758	3,944	1,426	637

Table 6: Centroid values of clusters generated with K-medoids method for 2012 data

K-MEDOIDS CENTROID TABLE (2012)				
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Number of Motor Land Vehicles	1,854,920	423,655	144,895	51,500
Number of Traffic Accidents with Death – Injury	12,071	4,105	1,525	582
Number of Deaths	203	87	46	21
Number of Injuries	18,677	6,843	2,918	1,156

3.5 Multivariate Map Design with AGNES Hierarchical Clustering Method

In application of AGNES method, Euclidean distances were used in determination of similarity of *i* and *j* elements of the cluster. Euclidean distance of two elements was calculated with the equation:

$$d_{ij} = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2} \quad (5)$$

In determination of the most suitable cluster number for AGNES method, a tree structure called a dendrogram, which was explained in Section 2, was used. Large jumps on the dendrogram showed the clusters. However, since the aim is to generate homogeneity and to be different from the other groups in clustering, it is necessary to analyse the dendrogram with an auxiliary axis and determine the homogeneous groups. SPSS software was used in an AGNES Hierarchical Clustering process. A dissimilarity matrix was used in the clustering process, while the dendrogram was used in determination of cluster numbers, and cluster elements were attained with the help of this software. Each node of the tree structure (dendrogram) that intersects an

auxiliary axis represents a cluster. In this way, five clusters were determined for both 2011 and 2012 data. Centroid Tables of clusters generated as a result of the clustering process were given in Table 7 and Table 8.

Again, with the help of classes obtained by using four different values in clustering processes with AGNES method, multivariate maps were designed for the years 2011 and 2012 (Figure 4).

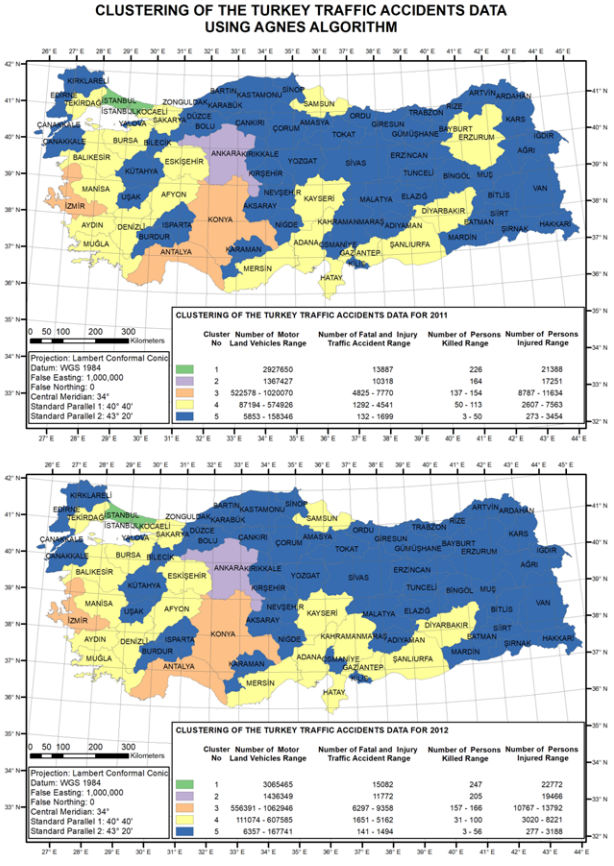

Figure 4: Multivariate maps designed with the AGNES method for the years of 2011 (above) and 2012 (below)

4 Results

In the scope of multivariate mapping, more than one attribute can be displayed in separate maps or in the same map. Comparing two objects which have various attributes with separate maps is more difficult. Therefore, methods in which various attributes are shown in the same map are preferred more. One of the methods in which various attributes are displayed in the same map is to generate thematic map classes by determining the effect of different attributes with clustering analysis. In this sense, in this study, considering traffic accidents in 2011 and 2012 in Turkey and by using four parameters (number of vehicles in traffic, number of traffic accidents resulting in death and injuries, number of casualties, and number of injuries), multivariate maps were designed with three different cluster analysis methods.

As with all data mining methods, the purpose of clustering methods is to uncover results that are not normally observed in the data set. In this study, if the thematic maps were generated using a single variable, the resulting maps showed significant differences from the multivariate maps (Figure 2, 3, and 4) designed by the clustering of four variables. For example, considering the number of motor vehicles in traffic, there are 1,020,070 vehicles in Izmir, 1,367,000 vehicles in Ankara, and 522,578 vehicles in Konya in 2011 [1]. According to this result, it can be considered that Izmir and Ankara are in similar risk groups. However, when Figure 2 and Figure 4 are examined, it is seen that Izmir is in the same cluster as Konya when four variables are taken into consideration. This situation shows that although Izmir has about twice the number of motor vehicles in Konya, the results are similar to Konya when the results of traffic accidents are also taken into account. A similar situation applies to 2012-year data. In different cities such as Erzurum and Diyarbakır, risk groups for both 2011 and 2012 were also changed in a similar way.

When the multivariate maps (Figure 2, 3, and 4), designed using the three methods, are examined, it is seen that risk groups for more than 50 provinces (Konya, Antalya, Eskisehir, Afyon, Sivas, Tokat, etc.) are similar for 2011 and 2012. This situation is very important in terms of forecasting accidents and traffic planning.

When the designed maps (Figure 4), clusters, and cluster elements determined on the dendrogram were analysed, the 2011 and 2012 maps designed with AGNES method were the same except for one city (K.maraş). This result showed that multivariate maps designed with the AGNES method were quite important in the sense of risk management. This is because risk regions predicted with 2011 data were confirmed in the 2012 data.

K-means and K-medoids non-hierarchical clustering algorithms divide n objects into k clusters according to k input parameters. They form the same cluster if objects resemble each other but not with the objects in other clusters. The greatest problem in applying these algorithms is the determination of the k cluster number. This can be determined with some of the calculation methods which were explained in Section 3. Better clustering results were obtained with k=4 cluster number for data sets used in the

Table 7: Centroid values of clusters generated with AGNES method for 2011 data

AGNES CENTROID TABLE (2011)					
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Number of Motor Land Vehicles	2,927,650	1,367,427	763,393	289,375	66,371
Number of Traffic Accidents with Death – Injury	13,887	10,318	6,211	2535,3	684
Number of Deaths	226	164	148	80	25
Number of Injuries	21,388	17,251	9,958	4,554	1,402

Table 8: Centroid values of clusters generated with AGNES method for 2012 data

AGNES CENTROID TABLE (2012)					
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Number of Motor Land Vehicles	3,065,465	1,436,349	803,977	300,970	69,078
Number of Traffic Accidents with Death – Injury	15,082	11,772	7,608	2,955	761
Number of Deaths	247	205	160	66	26
Number of Injuries	22,772	19,466	11,899	5,130	1,498

study. Although the clustering success for both algorithms was similar, when centroid tables of clusters formed with both methods (Table 3-6) were observed, it was detected that clusters are separated better in the K-medoids algorithm. Since the aim is to provide high intracluster similarity and low similarity between different clusters, it can be said that the K-medoids method gives better results for these data.

5 Conclusion

It is possible to make analyses on the spatial data with the thematic maps produced by utilizing Geographic Information Systems. Generally, one or two spatial data are visualized and analysed in thematic maps. However, as in this study, it is possible to make analyses by taking advantage of more spatial data with multivariate maps produced using data mining methods. In this study, traffic accidents are grouped spatially using four criteria according to the results of traffic accidents. In this way, the aim was to estimate the results of traffic accidents that may occur in the future and to plan accordingly.

The purpose of the clustering methods is to reveal similar data in the data set. Analysis results based on the designed maps (Figure 2, 3, and 4) exhibited that metropolitan cities (such as Istanbul, Ankara, Izmir), were usually in the same cluster, and same risk level when four attributes were considered. Additionally, cities which are tourist provinces and the main corridor provinces of Turkey (such as Antalya, Konya, etc.), were determined to be the similar cluster. This situation was also true for both years. In addition, it is observed that Tunceli, Batman, Bayburt, Mardin, etc., which have low population densities and are not found on the main road routes, were in the same risk level in terms of the nature of traffic accidents.

With this study it was shown that, by using clustering methods, similar aspects of different spatial objects can be presented by considering more than one attribute. It is thought that by using multivariate maps designed with clustering methods, spatial analyses which have important contributions for practices such as risk management, planning, etc., can be made. In this context, multivariate maps can also be used for determining the reasons for traffic accidents, common features of traffic accidents, estimating effects of future traffic accidents, planning traffic systems, and providing traffic safety.

Acknowledgement: We would like to thank to reviewers for their reviews, critical comments and helpful suggestions, which improved the manuscript greatly.

References

[1] TUIK 2015 http://www.tuik.gov.tr/ (accessed 25.12.2017)
[2] Bil M., Andrasik R., Svoboda, T., and Sedonik, J. The KDE+ software: a tool for effective identification and ranking of animal-vehicle collision hotspots along networks. Landscape Ecol., 2016, 31, 231–237.

[3] Lord D., and Mannering F. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A, 2010, 44, 291–305.
[4] Lin L., Wang Q., Sadek A.W. A combined M5P tree and hazard-based duration model for predicting urban freeway traffic accident durations. Accident Analysis and Prevention, 2016, 91, 114–126.
[5] Yalcin G., and Duzgun H.S. Spatial analysis of two-wheeled vehicles traffic crashes: Osmaniye in Turkey. KSCE Journal of Civil Engineering, 2015, 19(7), 2225-2232.
[6] Erdogan S. Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. Journal of Safety Research, 2009, 40, 341–351.
[7] Shi A., Tao Z., Xinming Z., and Jian W. "Evolution of traffic flow analysis under accidents on highways using temporal data mining." 2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications, Zhangjiajie, Hunan, 2014.
[8] Akoz Ö., Karsligil M. E. Traffic event classification at intersections based on the severity of abnormality. Machine Vision and Applications, 2014, 25, 613–632.
[9] Weng J., Qiao W., Qu X., and Yan X. Cluster-based lognormal distribution model for accident duration, Transportmetrica A: Transport Science, 2015, 11(4), 345-363.
[10] Guo, F. and Fang Y. "Individual driver risk analysis using naturalistic driving data." 3rd International Conference on Road Safety and Simulation, September 14-16, 2011, Indianapolis, USA.
[11] Martinussen, L.M., Møller M., and Prato C.G. Assessing the relationship between the Driver Behavior Questionnaire and the Driver Skill Inventory: Revealing sub-groups of drivers. Transportation Research Part F, 2014 26, 82–91.
[12] Feng S., Li Z., Ci Y., and Zhang G. Risk factors affecting fatal bus accident severity: Their impact on different types of bus drivers. Accident Analysis and Prevention, 2016, 86, 29–39.
[13] Jackson T. and Sharif H.O. Rainfall impacts on traffic safety: rain-related fatal crashes in Texas, Geomatics, Natural Hazards and Risk, 2016, 7(2), 843-860.
[14] Dogru N. and Subasi A. Comparison of clustering techniques for traffic accident detection. Turkish Journal of Electrical Engineering & Computer Sciences, 2015, 23, 2124-2137
[15] Nokhandan M.H., Bazrafshan J., Ghorbani K. A quantitative analysis of risk based on climatic factors on the roads in Iran. Meteorol Appl. 2008, 15:347357.
[16] Erdogan S., Yilmaz I., Baybura T., Gullu M. Geographical information systems aided traffic accident analysis system case study: City of Afyonkarahisar," Accident Analysis & Prevention, 2008, 40(1), 174-181.
[17] Murray A.T. and Grubesic, T.H. "Exploring spatial patterns of crime using non-hierarchical cluster analysis." *In Crime modeling and mapping using geospatial technologies,* 2013, 105-124, Springer Netherlands.
[18] Grubesic T.H., Wei R. and Murray, A.T. Spatial clustering overview and comparison: accuracy, sensitivity, and computational expense. Annals of the Association of American Geographers, 2014, 104(6), 1134-1156.
[19] Buckley A. "Multivariate mapping." In *Encyclopedia of Geographic Information Science* edited by Kemp K., 2008, 300-303.
[20] Slocum T.A., McMaster R.B., Kessler F.C. and Howard. H.H. "Thematic Cartography and Geovisualization." Pearson Education Inc. Third Edition, USA, 2009.
[21] Brewer C.A. "Color use guidelines for mapping and visualization." In *Visualization in Modern Cartography* edited by MacEachren A.M. and Taylor D.R.F., 1994, 123-147.
[22] Metternicht G. and Stott. J. "Trivariate spectral encoding: A prototype system for automated selection of colours for soil maps based on soil textural composition." *Proceedings of the 21st International Cartographic Conference*, Durban, CD, 2003.
[23] Byron J.R. "Spectral encoding of soil texture: A new visualization method." *GIS/LIS Proceedings*, Phoenix, Airz., 1994, 125-132.
[24] Interrante V. 2000. "Harnessing natural textures for multivariate visualization." *IEEE Computer Graphics and Applications* 20(6), 6-11.
[25] Jenks G.F. Pointillism as a cartographic technique. The Professional Geographer, 1953, 5, 4–6.
[26] Cox D.J. The art of scientific visualization. Academic Computing. 1990, 4, 20-22, 32-34, 36-38.
[27] Ellson R. Visualization at work. Academic Computing, 1990, 4(6), 26-28, 54-56.
[28] Dorling D. The visualization of local urban change across Britain. Environment and Planning B: Planning and Design, *1995*, 22, 269-290.
[29] Grinstein, G., Sieg J.C.J., Smith S. and Williams M.G. Visualization for knowledge discovery. International Journal of Intelligent Systems, 1992, 7, 637–648.
[30] Healey, C.G. and Enns J.T. Large datasets at a glance: Combining textures and colors in scientific visualization. IEEE Transactions on Visualization and Computer Graphics, 1999, 5(2), 145-167.
[31] Miller J.R. Attribute blocks: Visualizing multiple continuously defined attributes. IEEE Computer Graphics and Applications, 2007, 27(3), 57-69.
[32] Zhang X. and Pazner M. The icon imagemap technique for multivariate geospatial data visualization: Approach and software system. Cartography and Geographic Information Science, 2004, 31(1), 29-41.
[33] Nelson E. S. and Gilmartin P. P. "An evaluation of multivariate, quantitative point symbols for maps." In *Cartographic Design: Theoretical and Practical Perspectives* edited by C. H. Wood and C. P. Keller, 1996, 191–203.
[34] DiBiase D. Designing animated maps for a multimedia encyclopedia. Cartographic Perspectives, 1994, 19, 3-7.
[35] Nelson E.S. Designing effective bivariate symbols: The influence of perceptual grouping processes. Cartography and Geographic Information Science, 2000, 27(4), 261-78.
[36] Everitt B.S., Landau S., Leese M., and Stahl D. *Cluster Analysis*. Chichester, West Sussex, U.K: Wiley. ISBN 9780470749913. 2011.
[37] Han J., Lee J.G. and Kamber M. "An overview of clustering methods in geographic data analysis." In *Geographic Data Mining and Knowledge Discovery* edited by Miller H.J. and Han H., Taylor & Francis Group, LLC. 2009.
[38] Han J. and Kamber M. "Data Mining: Concepts and Techniques." San Francisco, 2006.
[39] Romesburg H.C. Cluster Analysis for Researchers. Belmont, CA: Lifetime Learning Publications. 1984.