International Journal of
Applied Mathematics,
Electronics and Computers
ISSN: 2147-8228 http://ijamec.atscience.org

Advanced Technology and Science

Original Research Paper

Using K-Means and K-Medoids Methods for Multivariate Mapping

Huseyin Zahit Selvi *[1], Burak Caglar [1]

Accepted 3rd September 2016

Abstract: Multivariate mapping is the visual exploration of multiple attributes using a map or data reduction technique. The simultaneous display of sometimes 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. 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. k-means and k-medoids methods which are non-hierarchical clustering analysis methods were analyzed in this study. The aim of this study is to determine the success of the spatial analysis of the multivariate maps produced by these methods. For this aim, classes and multivariate maps created with these methods from traffic accident data of two different years in Turkey were presented. In addition usability of such maps in risk management and planning was discussed.

Keywords: *Multivariate mapping, Data mining, Cluster analysis, Visualization, Cartography*

1. Introduction

Multivariate mapping is the graphic display of more than one variable or attribute of geographic phenomena. The simultaneous display of sometimes 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 [1].

A fundamental issue in multivariate mapping is whether individual maps should be shown for each attribute or whether all attributes should be displayed on the same map ([2], p.327). Producing separate maps for each attributes would 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 more. In this sense, Trivariate Choropleth Map which is created by overlapping of two colored choropleth map [3-6], Multivariate Dot Maps method in which specific color or symbol is used for each attribute in the map [7], Multivariate Point Symbol methods which are used when multivariate data can be shown with point symbols [8-15], method in which combining different type of symbols is used for representing multivariate data [16] and method of separating different attributes from integral symbols [17] can be listed.

Different from the methods above, in order to represent many attributes in the same map, classification method based on clustering methods in data mining can be used as well. With the use of clustering methods, similar aspects of different spatial objects can be revealed by considering more than one attribute [18, 19]. In this sense, spatial analyses that would make important contributions for risk analysis, planning etc. can be done.

It is quite important to make use of more than one attributes of current spatial event and determine similarity of spatial events in order to detect future events. For this aim, different spatial and non-spatial analysis methods are used for determining common properties of the spatial events. In this context, [20], [21], [22] and [23] used different clustering methods for determining common properties of different spatial events. Mapping of spatial events according with common properties is also significant for estimating types and effects of future events. Multivariate maps of occurred spatial events are more effective on determining common properties of spatial events and planning investments according to these properties. The aim of this study is to determine the success of the spatial analysis of the multivariate maps produced by k-means and k-medoids clustering analysis methods. For this aim, k-means and k-medoids methods are used for designing multivariate mapping of traffic accidents in Turkey with data of two different years, and result maps are compared. Success of these methods is determined with comparison of maps designed in different years according to two different methods.

In this study, Cluster Analysis and k-means and k-medoids methods in the second section, application in the third section will be explained in details, results and suggestions will be shared in last section.

2. Cluster Analysis

Cluster analysis is the process of grouping information in a data set according to specific proximity criteria. Similarity of element in the same cluster should be high, similarity between clusters should be low [24]. In the process of classification, classes are determined before. In clustering method, classes are not determined before. Data are separated different classes according to the similarity of data obtained.

Cluster methods are classified in different ways in various resources. In general sense, cluster methods can be classified as hierarchical and non-hierarchical methods [2].

- Non-hierarchical Methods: In non-hierarchical methods, n objects are divided into k clusters according k number (k<n) given before. This method divides data in a way that there will be at least one object in each cluster and each object will be included at least in one cluster [25].
- Hierarchical Methods: Hierarchical clustering methods group data objects in tree structure. Hierarchical clustering methods are classified as agglomerative or divisive according to hierarchical division being bottom-up or top-down [26].

[1] *Necmettin Erbakan University Geomatic Engineering Department Division of Cartography, Konya, Turkey*
** Corresponding Author: Email: hzselvi@konya.edu.tr*
Note: This paper has been presented at the 3rd International Conference on Advanced Technology & Sciences (ICAT'16) held in Konya (Turkey), September 01-03, 2016.

In this study, k-means and k-medoids methods from non-hierarchical methods will be analysed.

2.1. 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. General logic of k-means algorithm is to divide a data set composed of n data object to k clusters determined depending on preliminary information and experience of researcher. The aim is to provide the intracluster similarity is high, but the intercluster similarity is low. Similarity of clusters is calculated with the mean value of objects.

The k-means procedure is summarized as below [25]:

Input:

k: the number of clusters,

D: a data set containing n objects.

Output: A set of k clusters.

Method:

(1) arbitrarily choose k objects from D as the initial cluster centers;

(2) (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;

(3) update the cluster means, i.e., calculate the mean value of the objects for each cluster;

(4) repeat until no change; (Fig.1)

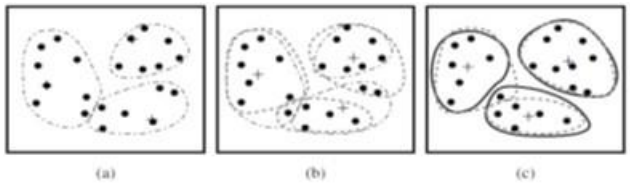

Fig. 1 Clustering with k-means algorithm [26]

2.2. K-Medoids Method

In this algorithm which was developed by Kauffman and Rousseeuw in 1990, instead of mean value of each cluster, an object in each cluster is taken as representative. This representative object, called a medoid, is meant to be the most centrally located object within the cluster [25]. 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 [27].

Steps of k-medoids algorithm are summarized as below [28]:

(1) Determination of k-cluster number.

(2) Choose of k objects as initial medoids

(3) Assign the remaining objects to cluster which has the most similar x medoid.

(4) Calculate aim function (sum of distances of all objects to the closest medoid)

(5) Arbitrarily choose of y point which is not medoid.

(6) If change of x and y would minimize aim function, change the place of these two points (x and y)

(7) The process is repeated between 3rd and 6th step until there is no change.

3. Application

Casualties, injuries and financial damages as a result of traffic accidents are among the most important problems of Turkey. When data of the last 5 years are analyzed, 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 results with financial damage. 4000 people lose their life on average in these accidents and nearly 250.000 people are injured. It is quite important to make use of more than one current traffic accidents attributes and determine similarity of traffic data on city basis in order to detect measures to be taken for traffic security in Turkey and future investments to be done. For this aim, in this study, clustering analysis will be made in 2 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 injuries (4 different values) for the years 2011 and 2012 prepared by Turkish Statistical Institute (TUIK) and multivariate maps will be produced according to analysis results. Maps designed for both years with 2 different methods were compared, it will be evaluated which method is suitable for the success of multivariate mapping and clustering.

In the application of clustering analysis methods, RapidMiner software developed in Dortmund Technology University Artificial Mind Unit by Ralf Klinkenberg, Ingo Mierswa and Simon Fischer was used. Multivariate maps were designed by ArcGIS software developed by ESRI group

3.1. Multivariate Map Design with k-means Method

RapidMiner software was used in application of k-means method. As a result of test made in this sense, k cluster number is 5, number of highest iteration that would be made during operation of algorithm for once is 100, and maximum cycle of algorithm is 35. The method was applied separately for 2011 and 2012 data. Centroid Tables of clusters generated as a result of clustering processes are given in Table 1 and Table 2.

Table 1: 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	Cluster 5
Number of Motor Land Vehicle	2147538.5	763392.7	452479.6	224009.6	63587.3
Number of Traffic Accidents with Death – Injury	12102.5	6210.7	3753	2057.4	653.1
Number of Death	195	147.7	104.8	68.8	24.2
Number of Injuries	19319.5	9957.7	6330.2	3865.2	1337.1

Table 2: 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	Cluster 5
Number of Motor Land Vehicle	2250907	803977.3	370049	141262.4	51499.9
Number of Traffic Accidents with Death – Injury	13427	7607.7	3544.2	1500.9	581.9
Number of Death	226	160.3	75.5	44	20.6
Number of Injuries	21119	11899	6015.5	2857.2	1155.7

With the help of classes obtained by using 4 different values (number of motor land vehicle, number of traffic accidents

resulting in death and injury, number of casualties and injuries) in clustering processes, multivariate maps showing similarity of traffic accidents on city basis for Turkey were designed (Fig.2).

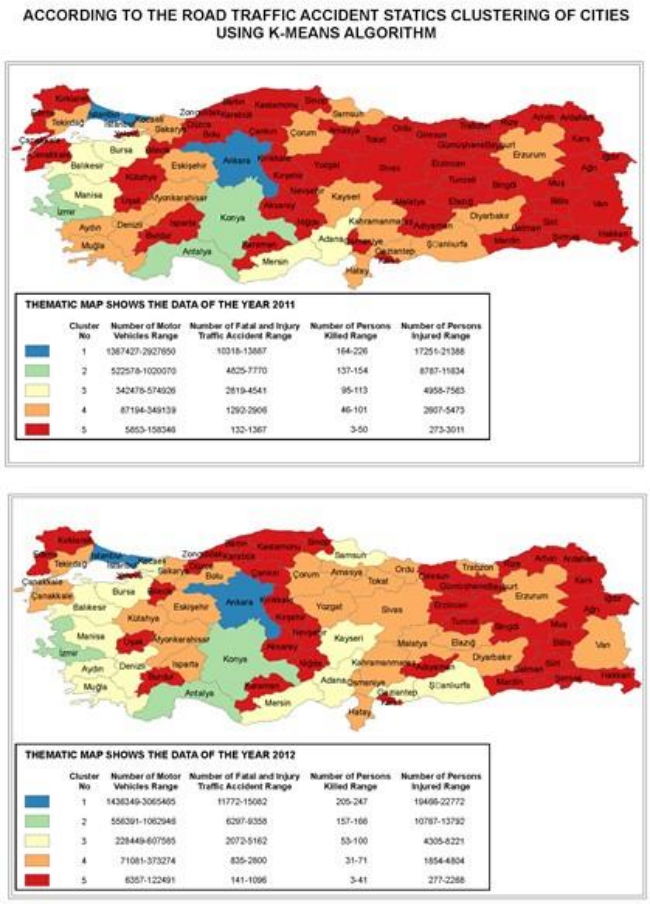

Fig. 2: Multivariate maps designed with k-means method for the years 2011 (above) and 2012 (below)

3.2. Multivariate Map Design with K-Medoids Method

RapidMiner software was also used in application of k-medoids method. Different from k-means algorithm, k-medoids process operator was used instead of k-means operator. In this scope, k cluster number was again taken as 5, and maximum cycle of algorithm was taken as 35. Centroid Tables of clusters generated as a result of clustering processes are given in Table 3 and Table 4.

Table 3: 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**	**Cluster 5**
Number of Motor Land Vehicle	1771715.7	443392.2	230032	91045.2	30875.1
Number of Traffic Accidents with Death – Injury	10658.3	3745.8	2111.1	899.8	362.9
Number of Death	175.7	112.9	66.7	33.7	13
Number of Injuries	16757.7	6460.4	3901.4	1836.2	754.8

Table 4: 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**	**Cluster 5**
Number of Motor Land Vehicle	1854920	423655.3	164997.7	78071.2	30854.4
Number of Traffic Accidents with Death – Injury	12070.7	4104.9	1730.2	831.5	402.4
Number of Death	203.3	87.4	49.3	31.1	10.8
Number of Injuries	18676.7	6842.9	3288.4	1657.2	773.2

Again with the help of classes obtained by using 4 different values in clustering processes with k-medoids method, multivariate maps were designed for the years 2011 and 2012 (Fig. 3).

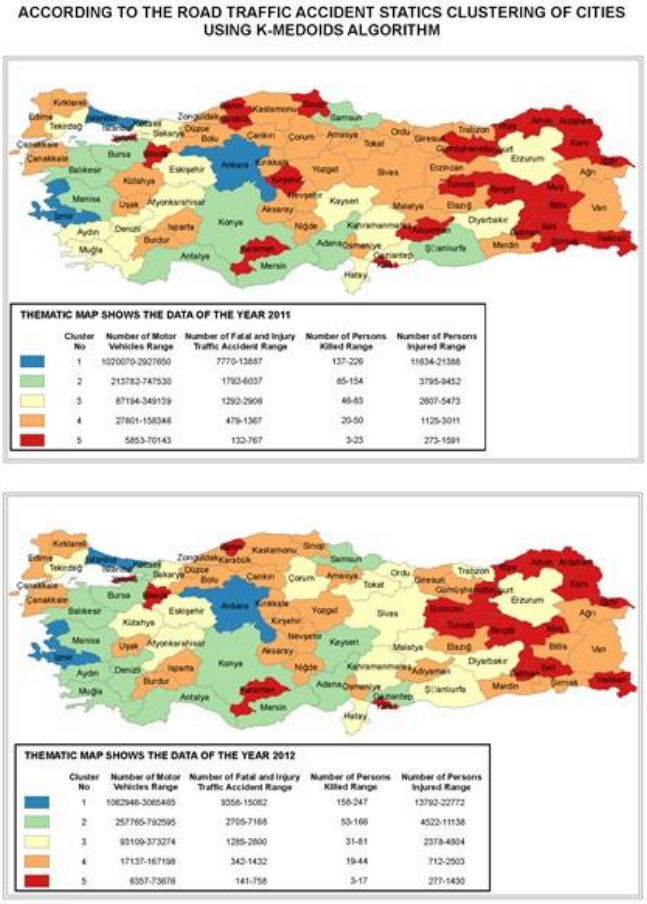

Fig. 3 Multivariate maps designed with K-medoids for the years 2011 (above) and 2012 (below)

4. Conclusion

In the scope of multivariate mapping, more than one attribute can be displayed in separate maps or in the same map. Designing separate maps for each attribute would 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 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, number of vehicles in

traffic in these years, number of traffic accidents resulting in death and injuries, number of casualties and injuries parameters, multivariate maps were designed with 2 different cluster analysis methods.

K-means and k-medoids non-hierarchical clustering algorithms divide n objects into k clusters according to k input parameter. 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 determination of k cluster number. This can be determined with some of the practices experiences. When k=5 cluster number is given for data sets used in the study, it is observed that better clustering results were obtained. Although clustering success of both algorithms are similar, when centroid tables of clusters formed with both methods (Table 1-4) are observed, it was detected that clusters are separated better in k-medoids algorithm. Since the aim is to provide high intracluster similarity and low similarity between different clusters, it can be said that k-medoids method gives better results for these data. When the maps designed using k-medoids method are examined, it is observed that 61 of 81 cities are located in the same cluster for two years. On the other hand, using k-means method 58 of 81 cities are in the same cluster. These results exhibit that foresights using these methods with 2011 data are highly consistent with 2012 data.

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 attributes. 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.

References

[1] Buckley A., Multivariate mapping, In Encyclopedia of Geographic Information Science edited by Kemp K., 2008, 300-303.

[2] Slocum T.A., McMaster R.B., Kessler F.C. and Howard H.H., Thematic Cartography and Geovisualization, Pearson Education Inc. Third Edition, USA, 2009.

[3] 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.

[4] Metternicht G. and Stott J., Trivariate Spectral Encoding: A Prototype System for Automated Selection of Colours for Soil Maps Based on Soil Textural Composition, in Proceedings of the 21st International Cartographic Conference, Durban, CD, 2003.

[5] Byron J. R., Spectral Encoding of Soil Texture: A New Visualization Method, in GIS/LIS Proceedings, Phoenix, Airz., 1994, 125-132.

[6] Interrante V., Harnessing Natural Textures for Multivariate Visualization, IEEE Computer Graphics and Applications, 2000, 20(6), 6-11.

[7] Jenks G. F., Pointillism as a Cartographic Technique, The Professional Geographer, 1953, 5, 4-6.

[8] Cox D.J., The Art of Scientific Visualization, Academic Computing, 1990, 4, 20-22, 32-34, 36-38.

[9] Ellson R., Visualization at Work, Academic Computing, 1990, 4(6), 26-28,54-56.

[10] Dorling D., The visualization of local urban change across Britain, Environment and Planning B: Planning and Design, 1995, 22, 269 -290.

[11] 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.

[12] 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.

[13] Miller J.R., Attribute Blocks: Visualizing Multiple Continuously Defined Attributes, IEEE Computer Graphics and Applications, 2007, 27(3), 57-69.

[14] 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.

[15] 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.

[16] DiBiase D., Designing Animated Maps for A Multimedia Encyclopedia, Cartographic Perspectives, 1994, 19, 3-7.

[17] Nelson E.S., Designing Effective Bivariate Symbols: The Influence of Perceptual Grouping Processes, Cartography and Geographic Information Science, 2000, 27(4), 261-78.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Martinussen L. M., Møller M., 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.

[24] Dinçer E.Ş., Veri Madenciliğinde K-means Algoritması ve Tıp Alanında Uygulanması, M. Eng. Thesis, Kocaeli University Institute of Science, Kocaeli, 2006 (in Turkish).

[25] 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.

[26] Han J. and Kamber M., Data Mining: Concepts and Techniques, San Francisco, 2006

[27] Silahtaroğlu, G.. Veri Madenciliği (Kavram ve Algoritmaları). Papatya Publishing, 2013, İstanbul, (in Turkish).

[28] Akın, Y.K. Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi. PhD diss., Marmara University Institute of Social Science, 2008,İSTANBUL, (in Turkish).