# A Decision Support System Using Demographic Issues: A Case Study in Turkey

Suat Secgin and Gokhan Dalkilic
Department of Computer Engineering Department, Dokuz Eylul University, Turkey

**Abstract**: *The demographic distribution of people by cities is an important parameter to address the people's behaviour. To distinguish people behaviour is useful for companies to understand the customer behaviour. In this article, a case study covering all 81 cities in Turkey and measuring 35 topics for each of them is handled. By using these topics and cities, it is investigated that how the cities are clustered. Because its efficiency, the Agglomerative hierarchical clustering and the K-medoids clustering methods in rapidminer data mining software are used to cluster the data. To measure the efficiency of the agglomerative clustering algorithm, the Cophenetic Correlation Coefficient (CPCC) is used. After clustering, the results are inserted into a geographic information system to depict the results in a Turkey map. The results show that, the cities distributed in the same geographical areas are in the same clusters with some exempts. On the other hand, some cities those are in different provinces show the same behaviour. The results of the study can also be used as a decision support system for a customer relations management.*

**Keywords**: *Agglomerative clustering, customer behaviour, data mining, decision support.*

*Revived July 24, 2014; accepted August 16, 2015*

## 1. Introduction

In countries, the demographic issues of cities play main roles in several counts such as political elections, economical status, customer and consumer behaviour, immigration and urbanization, etc., [3, 10, 16, 17, 19]. In addition to these, customer segmentation by using multiple data, customer differentiation using data mining algorithms and data mining usage in Geographical Information Systems (GIS) are the popular interdisciplinary research areas [1, 2, 7, 8, 11, 21]. Understanding the behaviour of the population of cities will help in deciding customer differentiation and retention for a company, to choose the right selling procedure [12].

The customer satisfaction issue of a company is the key parameter for retaining the customer, customer loyalty and preventing the churn [4, 13]. Keeping the customer perception high should be the focal point for a company to survive in the competitive environment. The state of the art companies are moving from acting as sales oriented to customer oriented [10].

In this article, by combining the CNBC-E magazine research [15] and the Customer Satisfaction Performance Scores (CSPS) for all the cities (totally 81) of Turkey, the Agglomerative hierarchical clustering algorithm is applied to investigate the cities in terms of some demographic issues and customer satisfaction. In the CNBC-E's habitableness research, 35 topics of different demographic issues such as air utilization rate, unemployment rate, number of pre-school students per classroom etc., are taken into consideration in order to, sort the cities. In addition to, the CNBC-E research results, the CSPS of a fixed line telecom operator in Turkey is added to rank the titles. All the data are conducted in the year of 2010. By using the CSPS apart from the CNBC-E data, the effect of the customer satisfaction ranks on the clustering is also investigated.

By clustering the cities depending on two different sources of data, a landscape of Turkey can be acquired. To cluster the cities, the Agglomerative hierarchical clustering and the K-medoids clustering algorithms are used. The Agglomerative clustering algorithm is selected because of its advantages such as nested clusters and flexibility. The K-medoids algorithm is used since the data being used in the article is a ranked table of the cities.

As the data mining software, the RapidMiner is used in the study. The MapInfo GIS software is used to convert the data to a map. As known, the Agglomerative clustering methods have three modes; single link, complete link and average link. After applying these modes, the efficiencies of the modes are calculated by using Cophenetic Correlation Coefficient (CPCC). Due to its efficiency, the average link mode is used to map the data and interpret the cluster results. As the measure type, the mixed measures and as the mixed measures parameter, the mixed Euclidian have been chosen.

Next step is calculating the accuracy and item distribution performance as well as optimizing the cluster count of the K-medoids. The relations among the cities spread in the same geographical areas are investigated. Using the maps for several levels, the status for the cities can be investigated by bottom-up

manner. After clustering the data, the results were used as the input to a geographical information system to visualize the results. In this respect, the study can be considered as a mixture of a geographic information system and a data mining system to present a decision support system.

## 2. Theoretical Background of Clustering Algorithms

In this section, theoretical background of the Agglomerative hierarchical clustering and K-medoids clustering algorithms in data mining for ranked data is given.

### 2.1. The Agglomerative Hierarchical Clustering Algorithm

There are two different hierarchical algorithms. The Divisive hierarchical clustering algorithm uses the top-down approach and the Agglomerative hierarchical clustering algorithm uses bottom-up approach [11, 18].

All Agglomerative hierarchical clustering algorithms have an initial state with each object as a separate cluster. Until only one cluster remaining, the clusters are successively combined to form new clusters [18]. In other words, we begin with each individual object and merge the two closest objects until all objects are aggregated into a single group [20]. The Agglomerative clustering algorithm flow chart is given in Figure 1 [20] and its pseudocode is given in Algorithm 1 [18].

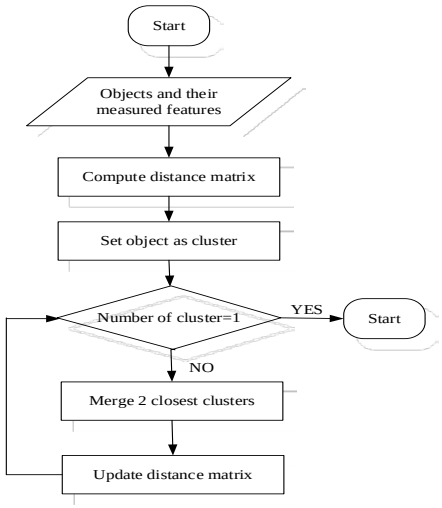


Figure 1. The flow chart of the Agglomerative hierarchical clustering algorithm.

*Algorithm* 1: *Agglomerative clustering algorithm.*

*Given:*
*A set of X objects* $\{X_1,\ldots, X_n\}$
*A distance function* $dis(c_1, c_2)$

1. *for i=1 to n*
   $C_i=\{x_i\}$
   *end for*
2. $C=\{c_1,\ldots, c_b\}$
3. *l=n+1*
4. *while C.size>1 do*
   a. $(c_{min1},c_{min2})$=*minimum* $dis(c_i,c_j)$ *for all* $c_i$, $c_j$ *in C*
   b. *remove* $c_{min1}$ *and* $c_{min2}$ *from C*
   c. *add* $\{c_{min1}, c_{min2}\}$*to C*
   d. *l=l+1*
   *end while*

The Agglomerative clustering has some advantages. First of all for discovery, different sized clusters in the tree have more importance. Second, clusters that are generated in early stages are nested inside the later stage generated clusters [6]. As well as hierarchical manner of nested clusters, the clusters explicitly separated from others can be taken as distinct clusters by investigating the clusters' result tree. These features of the Agglomerative clustering method are the main factors to use it in this article.

### 2.2. The K-medoids Clustering Algorithm

In contrast to taking the mean value of the objects in a cluster, in the K-medoids algorithm, an actual item in the cluster can be selected as the initial centroid of the cluster. After choosing the centroid, initial cluster figures can be depicted based on the distances between objects and centroids. The K-medoids algorithm also uses the principle of minimizing the sum of absolute-error criterion [9].

K-medoids clustering algorithm is as in Algorithm 2 [14]:

*Algorithm 2: K-medoids clustering algorithm.*

1. *Arbitrary selection of the k objects in the data D as medoid points (initial seeds).*
2. *After choosing of the k medoids, assign each data object in the given data set to most similar medoids based on the distances of Euclidean distance, Manhattan distance or Minskowski distance.*
3. *Non-representative object O' is randomly selected.*
4. *Swapping initial representative object's total cost (S) is computed to O'.*
5. *There will be new set of medoids when S<0, then swap the new medoid with the initial medoid.*
6. *All the steps between 2 to 5 are repeated until the point where there is no change in centroids (S>0).*

## 3. Applying the Agglomerative Clustering Algorithm and K-medoids Algorithm

In this section, after explaining the data used in the study, the processes that have been used in RapidMiner software are given.

### 3.1. The Source of the Data

In this study, for each city in Turkey, the annual average of the 2010 monthly CSPS of the fixed line operator and the research on the evaluation of livable cities in Turkey [15] are used for the inputs of the data mining algorithms. The 35 attributes for the cities in terms of the order of the cities are used in the

clustering algorithms as shown in Table 1. The attributes show the cities and the rows show the demographic topics. Every corresponding number of a city shows the order number of this city. For example, Ankara's rank is 48 for Unemployment Rate.

Table 1. Sample view of the data.

| Demographic Info/City Number | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Unemployment Rate | 48 | 58 | 30 | 12 | 65 |
| Amount of Tax per Capita | 3 | 14 | 32 | 22 | 2 |
| Deposit Amount per Capita | 2 | 5 | 22 | 24 | 1 |
| Public Expenditure per Capita | 18 | 13 | 63 | 23 | 16 |
| Number of Car per Adult | 1 | 9 | 10 | 54 | 3 |
| Number of House per Capita | 2 | 3 | 10 | 36 | 5 |
| Competitiveness | 2 | 6 | 23 | 16 | 1 |
| Average per Capita Expen. for Rental | 78 | 66 | 37 | 56 | 81 |
| Air Utilization Rate | 6 | 30 | 34 | 4 | 3 |
| Household per Capita Cons. Of Electricity | 6 | 18 | 42 | 12 | 3 |
| Rate of University Graduates | 1 | 2 | 5 | 7 | 4 |
| Literacy Rate | 4 | 3 | 15 | 40 | 20 |

For better understanding and simplicity, the city names are numbered as 1 to 5 for Ankara, Eskisehir, Isparta, Trabzon and Istanbul, respectively in Table 1. Full list is given in Table 2. Using the numbers rather than a name provides a clear view on the maps. Consequently, all of the ranking titles are given in Appendix A as well.

Table 2. City codes.

| City | Code | City | Code | City | Code | City | Code |
|---|---|---|---|---|---|---|---|
| Ankara | 1 | Karabuk | 22 | Duzce | 43 | Osmaniye | 64 |
| Eskisehir | 2 | Usak | 23 | Tekirdag | 44 | Aksaray | 65 |
| Isparta | 3 | Kayseri | 24 | Bartin | 45 | Yozgat | 66 |
| Trabzon | 4 | Canakkale | 25 | Manisa | 46 | Kmaraş | 67 |
| Istanbul | 5 | Kocaeli | 26 | Tokat | 47 | Diyarbakir | 68 |
| Artvin | 6 | Erzurum | 27 | Nigde | 48 | Kars | 69 |
| Bolu | 7 | Malatya | 28 | Tunceli | 49 | Muş | 70 |
| Antalya | 8 | Giresun | 29 | Kirsehir | 50 | Şirnak | 71 |
| Izmir | 9 | Kutahya | 30 | Corum | 51 | Batman | 72 |
| Rize | 10 | Bayburt | 31 | Denizli | 52 | Hakkari | 73 |
| Kirikkale | 11 | Zonguldak | 32 | Bilecik | 53 | Bingol | 74 |
| Edirne | 12 | Cankiri | 33 | Amasya | 54 | Igdir | 75 |
| Adana | 13 | Mersin | 34 | Erzincan | 55 | Van | 76 |
| Sinop | 14 | Konya | 35 | Sakarya | 56 | Şanliurfa | 77 |
| Burdur | 15 | Aydin | 36 | Ordu | 57 | Siirt | 78 |
| Sivas | 16 | Kastamonu | 37 | Afyon | 58 | Mardin | 79 |
| Samsun | 17 | Bursa | 38 | Gaziantep | 59 | Bitlis | 80 |
| Karaman | 18 | Balikesir | 39 | Kilis | 60 | Agri | 81 |
| Nevsehir | 19 | Elazig | 40 | Hatay | 61 | | |
| Mugla | 20 | Gumuşhane | 41 | Adiyaman | 62 | | |
| Kirklareli | 21 | Yalova | 42 | Ardahan | 63 | | |

An independent company has determined the CSPS for the fixed operator by using following methods:

- The research has been performed by phone calls of experienced agents.
- 4898 phone inquiries (i.e., 31 calls for each) have been carried out for 158 offices in 81 cities in Turkey. The average results of the offices in related cities are used to calculate the customer satisfaction of a city.
- Desk waiting time, service quality, employee attitude and knowledge, physical status of an office are the key areas to be inquired.
- For every week, a new randomly chosen customer list has been used for the calls.

## 3.2. The Main Process

In this sub-section, the process design in RapidMiner is given. The steps of the process are shown in Figure 2:

- Multiply the data to get the inputs of the single link, complete link and average link Agglomerative algorithms and use data to similarity operator to calculate the performances of each algorithm.
- After applying the three Agglomerative algorithms, multiply the outputs to get the inputs of the CPCC to calculate the performances of the three Agglomerative clustering algorithms.
- The data to similarity operator is also multiplied to calculate each of the Agglomerative clustering algorithms (single, complete and average links).
- The outputs of the three Agglomerative algorithms are multiplied to show all the results at the same time. In other words, the clustering algorithm results, cluster performance results and data itself can be viewed at the result view of RapidMiner.
- Apply the K-medoids algorithm to compare with the results of the Agglomerative algorithm.

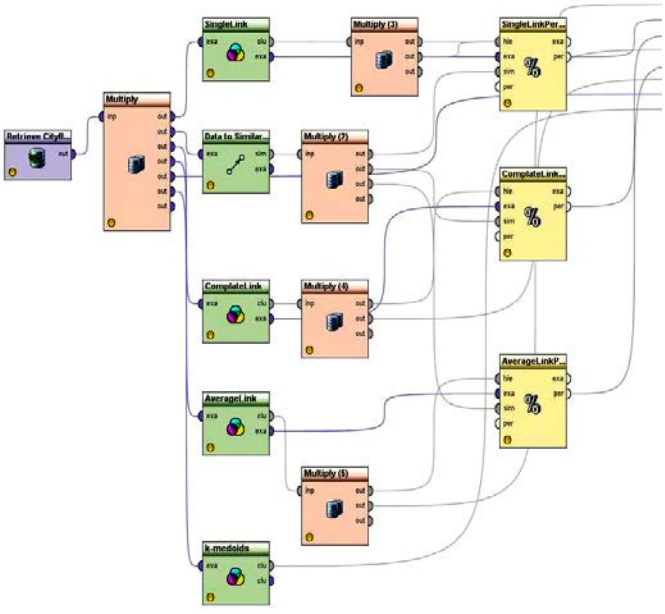


Figure 2. Main process diagram.

## 3.3. The Cophenetic Correlation Coefficient

The CPCC is an index that shows how good the clustering is performed. In hierarchical Agglomerative clustering algorithm, the distance matrix and the cophenetic matrix are needed to measure the efficiency level of the algorithm [20].

The cophenetic matrix is formed by the minimum merging distance and the lower triangular distance

matrix. The coefficient value close to 1 (100%) indicates the success of the clustering. In Figure 3, the operator connection of the CPCC in RapidMiner software [5] is given.

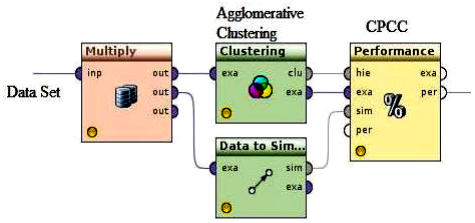


Figure 3. The CPCC operator connection in RapidMiner.

In our work, the CPCCs for single, complete and average link Agglomerative algorithms are 0.353, 0.589 and 0.644, respectively.

### 3.4. The Results of the Clustering Algorithms

In this sub-section, the results of the single, complete and average link Agglomerative algorithms and the results of the K-medoids algorithm are given. The Agglomerative clustering algorithm gives 162 clusters totally. As in this study average link Agglomerative algorithm result is 0.644 that is closest to 1, the average link Agglomerative clustering algorithm results are used. In Figure 4, the tree of the average link Agglomerative algorithm is depicted.

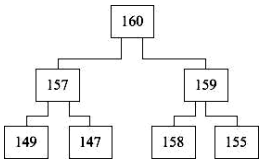


Figure 4. Tree view of the average link agglomerative clustering algorithm.

In the K-medoids algorithm, as a result of the parameter optimization, the cluster number k is chosen as 4. The detailed explanation for this process is given in section 4.

In Figure 4, the clustering mechanism of the data can be seen from the tree in detail. Since, our data are acquired by using city orders, the Agglomerative and the K-medoids algorithms are used to cluster. The Agglomerative clustering result can be seen from Figure 5 as well. For the k value of 8, the result is given in Figure 19 of Appendix C.

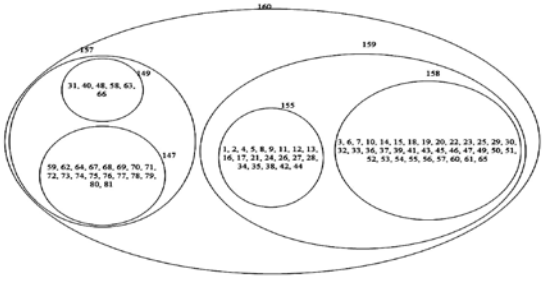

Figure 5. Nested clusters for k=4.

By using the tree view of the Agglomerative clustering, we can investigate the city distribution in detail. Some examples of the results are given in the following;

- City 22 (Karabuk) and city 30 (Kutahya) are the cities that have minimum distances in other words maximum similarity.
- City 15 (Burdur) and city 54 (Amasya) are the cities that have minimum distance in other words maximum similarity.
- When going up the tree, cities 22, 30, 15 and 54 are participated to the same cluster (Cluster number 104).

By using the tree view and the map view of the clusters, the detailed analysis of how the cities are combined in terms of the clusters can be conducted.

In Figure 5, the clusters for k=4 are depicted for the Agglomerative clustering. The cluster branches are also given as tree representation in Figures 15, 16, 17 and 18, respectively in Appendix B. The tree views of the clusters provide an opportunity of viewing data distribution as well.

In Figure 6, the K-medoids algorithm clusters are given to compare with the Agglomerative clustering results. These results are also compared in section 4 by using the map views.

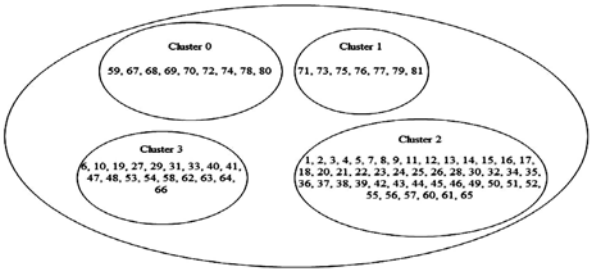


Figure 6. K-Medoids Clusters for k=4.

## 4. The Map Views of the Clustering Results and Performance Analysis

Map views for the results of the Agglomerative and K-medoids clustering algorithms are given in this section before analysing cluster performance and accuracy.

### 4.1. The Map Views of the Clustering Results

In this sub-section, the results of the Agglomerative and K-medoids clustering are shown in the maps to visualize the data. The maps are used to investigate the data in terms of geographical issues.

As seen from Figure 5, when the depth of the result tree is 2, there will be four clusters. The distribution of the cities is depicted in Figure 7.

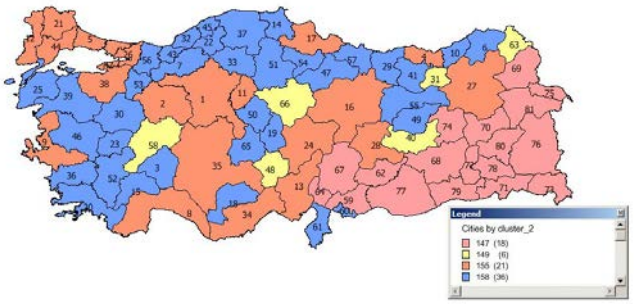

Figure 7. The city distribution for k=4 (Agglomerative Clustering).

When the cluster count k is chosen as 4, as can be seen from Figure 7, the cities in the same geographical areas are grouped into the same cluster with some exemptions. In other words, the people living in the cities of the same areas are behaving in the same manner in terms of our 35 parameters mentioned in Appendix A. Again, cities 58 (Afyonkarahisar), 48 (Nigde), 66 (Yozgat), 40 (Elazig), 31 (Bayburt) and 63 (Ardahan) are cut loose from the groups that are geographically dispersed into the same areas. The cities in cluster 149 exhibit a different behaviour compared to the others. This situation can be seen from Figure 8 (in K-medoids) as well. The maps of Agglomerative clustering for the k values 8, 16, 27, and 40 are given in Figures 20, 21, 22, and 23, respectively in Appendix C.

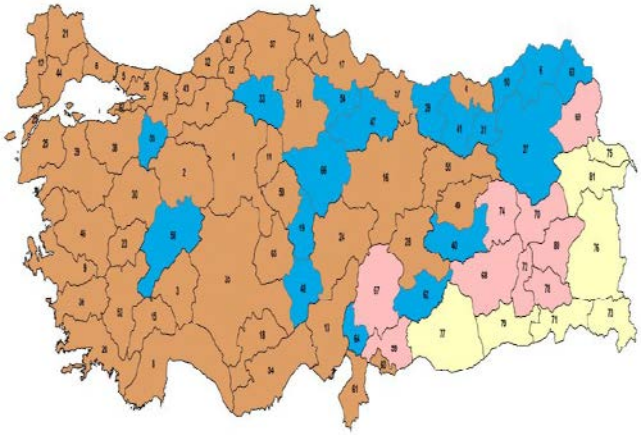

Figure 8. K-medoids results for the data (CNBC-E Data and CSPS).

To find out the effect of the customer satisfaction to the clustering, the Agglomerative and K-medoids algorithms are applied to the CSPS and the cities. As mentioned beforehand, the CNBC-E data is organized by sorted city scores. To get a better performance, in this case, the actual points of cities are used rather than sorted scores.

By comparing Figure 9 (Agglomerative clustering), and Figure 10 (K-medoids clustering), it can be said that there is no explicit effect of customer satisfaction to clusters. This situation is also proved in section 4.2.

Figure 9. Agglomerative clustering algorithm results for CSPS.

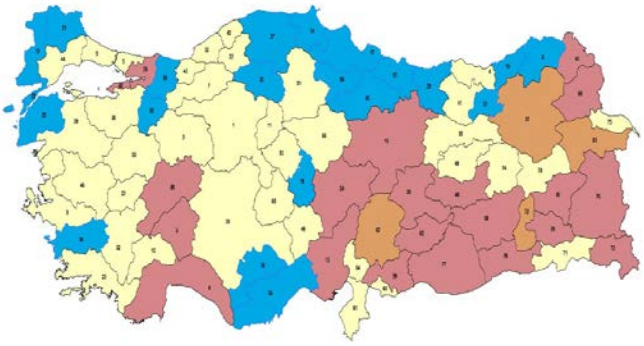

Figure 10. K-medoids results for CSPS.

## 4.2. Cluster Performance and Accuracy Analysis

In cluster performance and accuracy analysis, the main process depicted in Figure 11 is used. With the set role operator, the cluster names (i.e., cluster 0, cluster 1 etc.,) are assigned to the class labels. In this way, the city that has certain social parameters belongs to which cluster is investigated.

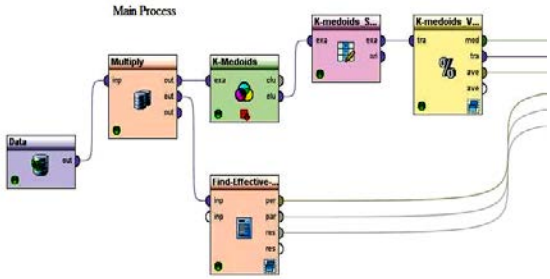


Figure 11. Main accuracy and performance measurement process.

After applying the set role operator, the X-validation is used to estimate how accurately a model will perform in practice. Since, the data used in the article is small sized; there is no need to split data. Thus, the cross validation operator is used to estimate the statistical performance. For the criterions of the cluster performance, the accuracy and the correlation performance parameters are selected in RapidMiner. The sub-process schema of the X-validation process is given in Figure 12. The decision tree is used since there are too many parameters in the data table such as 81 cities and 35 social items.

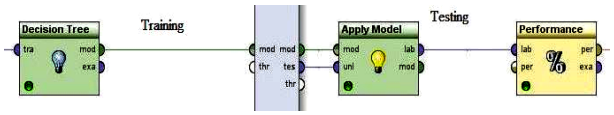


Figure 12. Validation sub processes.

In Figure 13, the optimization sub-process is depicted. This process is used to find the most effective k value. Item distribution performances for the cluster counts from 2 to 10 are given in Table 3.

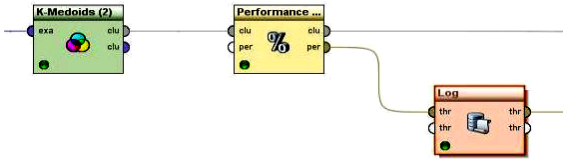


Figure 13. Optimization sub process.

Table 3. Item distribution performance for k value optimization.

| Cluster Count | Item Distribution Performance |
|---|---|
| **2** | 0.9912 |
| **3** | 0.9907 |
| **4** | 0.9733 |
| **5** | 0.9660 |
| **6** | 0.9640 |
| **7** | 0.9645 |
| **8** | 0.9579 |
| **9** | 0.9651 |
| **10** | 0.9588 |

The item distribution performances for k values are higher than 90%. So, to find the optimal k value the accuracy values should be checked. When comparing the accuracy levels, the optimal k value is found as 2 with the accuracy level of 86.25%. To create differences among cities, the optimal k value is selected as 4 by checking the accuracy levels and class recall values as in shown Table 4. Additionally, in the agglomerative clustering section, the optimal tree level was selected as 2 for tree depth. This also means that there are 4 clusters to investigate the cities in terms of geographic distribution.

Table 4. Accuracy, item distribution and class recall values for k=2, 3, 4, 5 (#: not available for this item).

| K Value | Accuracy (%) | Item Distribution Performance (%) | Class Recall (0, 1, 2, 3, 4 Respectively) (%) |
|---|---|---|---|
| 2 | 86.25 | 99.12 | 50.00, 92.75, #, #, # |
| 3 | 79.17 | 99.07 | 57.14, 92.06, 18.18, #, # |
| 4 | 77.92 | 97.33 | 40.00, 20, 95.8, 20, # |
| 5 | 74.17 | 96.60 | 46.15, 66.67, 90.57, 0, 66.67, # |

After selecting k value as 4, the most distinctive items are found as household per capita consumption of electricity, forest area ratio, rate of primary school students per teacher and rate of traffic accident per vehicle.

As seen from Figure 14, the first distinctive item for clustering is the household per capita consumption of electricity value. The cities having household per capita consumption of electricity higher than 78.00 is assigned to cluster 0. For the remaining cities, the parameter of the forest area ratio is the next item for determining the cluster to which the city is assigned. If the forest area ratio value is higher than 78.50, then the city is assigned to cluster 0, otherwise the rate of the primary school students per teacher value will be considered for the next distinctive parameter. In the same way, if the rate of the primary school students per teacher value is higher than 76.50, the city is assigned to the cluster 1. Otherwise, the rate of the traffic accident per vehicle is considered. As can be seen from Figure 6, 9 cities are assigned to cluster 0 in this way. From Figure 6, other items of the clusters can be found.

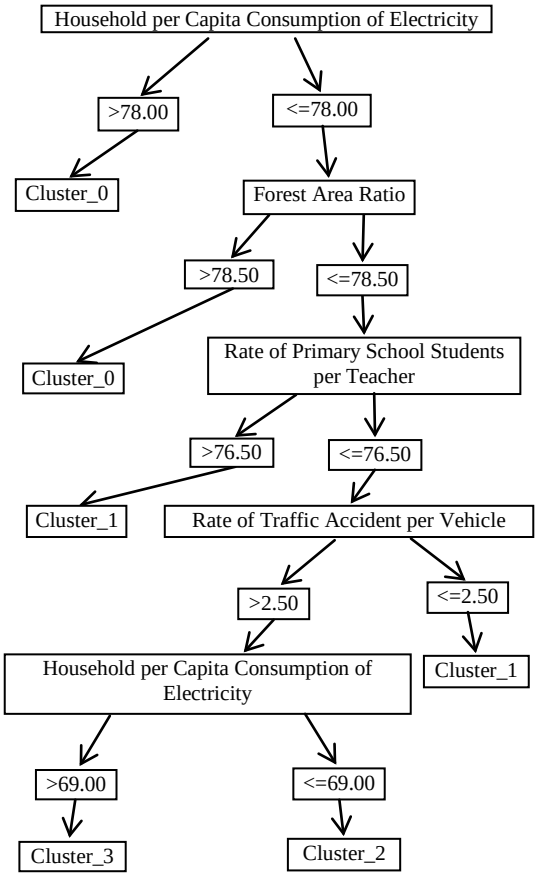


Figure 14. Distinctive parameters for clustering (for k=4).

If the rate of the traffic accident per vehicle value is higher than 2.50, the household per capita consumption of electricity value is reconsidered. If this value is higher than 69.00, the city is assigned to cluster 3 and if the value is less than or equal to 69.00, then the city is assigned to cluster 2. Finally, when the rate of traffic accident value is less than or equal to 2.50, the city is assigned to cluster 1. Additionally, the CSPS value is not a distinctive parameter for clustering.

Since, the optimum k value is computed as 4, there are four distinctive parameters. It is obvious that, the distinctive parameters will change if the cluster count k is changed.

## 5. Conclusions

The customer behaviour is the key parameter for the companies to organize the right campaigns or to

achieve better customer loyalty. In this article, by using a case study in Turkey, how the demographic issues affect the clustering of cities is investigated. As well as 35 demographic topics, the CSPS of a fixed line telecom operator in Turkey is added to the investigation. It was proved that:

- The cities that are geographically in the same areas tend to be in the same cluster.
- Bayburt and Ardahan are the two cities that are detached from others explicitly.
- The cities in eastern Anatolia region and southeastern Anatolia region tend to be in the same cluster.
- The cities in the black sea region tend to be in the same cluster.
- The cities 58 (Afyonkarahisar), 48 (Nigde), 66 (Yozgat), 40 (Elazig), 31 (Bayburt) and 63 (Ardahan) are cut loose from the groups that are geographically dispersed in to the same areas.
- The CPCC for single link Agglomerative clustering is 0.353, for complete link Agglomerative clustering it is 0.589 and for average link clustering algorithm it is 0.644. So, the average link Agglomerative clustering algorithm was used to examine the results.
- By using the tree view of the Agglomerative clustering, we can investigate the city distribution in detail as mentioned in section 3.4.
- The optimal k value for the K-medoids algorithm is 4.
- The clustering results can be used as a decision support system for the companies.

Since, the demographic issues of the people living in a country make difference among the cities; the question of how cities differ from others or are similar to others is a key area for a company's executives. Understanding the customer behaviour is very important to provide the services needed by the customers. In this article, a decision support system consisting of GIS and data mining techniques is designed as a case study for Turkey to depict a picture showing the customer behaviour in terms of the demographic issue of the cities in Turkey. It has also been shown that the people living in the same geographic areas are behaving in similar manners. So, when designing a customer experience program, the results of the article can be used easily.

## References

[1] Anselin L. and Getis A., "Spatial Statistical Analysis and Geographic Information Systems," *The Annals of Regional Science*, vol. 26, no. 1, pp. 19-33, 1992.

[2] Chiang W., "To Mine Association Rules of Customer Values via a Data Mining Procedure with Improved Model: An Empirical Case Study," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1716-1722, 2011.

[3] Cranshaw J., Schwartz R., Hong J., and Sadeh N., "The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City," *in Proceeding of 6th International AAAI Conference on Weblogs and Social Media*, Dublin, pp. 58-65, 2012.

[4] Deng Z., Lu Y., Wei K., and Zhang J., "Understanding Customer Satisfaction and Loyalty: An Empirical Study of Mobile Instant Messages in China," *International Journal of Information Management*, vol. 30, no. 4, pp. 289-300, 2010.

[5] Dvoroznak M., *http://korek.name/web/moje-tvorba/rapidminer-clustering_performance_plugin-average_silhouette-cophenetic_coefficient*, Last Visited 2014.

[6] Felici G. and Vercellis C., *Mathematical Methods for Knowledge Discovery and Data Mining*, Idea Group Reference, 2007.

[7] Garla S., Chakraborty G., and Gaeth G., "Comparison of K-means, Normal Mixtures and Probabilistic-D Clustering for B2B Segmentation using Customers' Perceptions," *in Proceeding of the SAS Global Forum*, Las Vegas, pp. 1-8, 2012.

[8] Gorsevski P., Donevska K., Mitrovski C., and Frizado J., "Integrating Multi-criteria Evaluation Techniques with Geographic Information Systems for Landfill Site Selection: A Case Study Using Ordered Weighted Average," *Waste Management*, vol. 32, no. 2, pp. 287-296, 2012.

[9] Han J., Kamber M., and Pei J., *Data Mining: Concepts and Techniques*, Morgan Kaufmann Press, 2006.

[10] Hossain M. and Leo S., "Customer Perception on Service Quality in Retail Banking in Middle East: The Case of Qatar," *International Journal of Islamic and Middle Eastern Finance and Management*, vol. 2, no. 4, pp. 338-350, 2009.

[11] Kaur J. and Gupta G., "Optimized Clustering Algorithm with Hybrid K-Means and Hierarchical Algorithms," *International Journal for Multi Disciplinary Engineering and Business Management*, vol. 2, no. 1, pp. 4-7, 2014.

[12] Khan K., Baharudin B., and Khan, A. "Identifying Product Features from Customer Reviews Using Hybrid Dependency Patterns," *The International Arab Journal of Information Technology*, vol. 11, no. 3, pp. 281-286, 2014.

[13] Kuo Y., Wu C., and Deng W., "The Relationships Among Service Quality, Perceived Value, Customer Satisfaction, and Post-purchase Intention in Mobile Value-added Services," *Computers in Human Behavior*, vol. 25, no. 4, pp. 887-896, 2009.

[14] Madhulatha T., "Comparison between K-Means and K-Medoids Clustering Algorithms,"

*Advances in Computing and Information Technology Communications in Computer and Information Science*, vol. 198, pp. 472-481, 2011.
[15] Mavi B., "Research on Livable Cities in Turkey," *CNBC-E Business Magazine*, vol. 9, pp. 64-98 2011.
[16] Musso J., "The Political Economy of City Formation in California: Limits to Tiebout Sorting," *Social Science Quarterly*, vol. 82, no. 1, pp. 139-153, 2001.
[17] Pamuk A., "Geography of Immigrant Clusters in Global Cities: A Case Study of San Francisco," *International Journal of Urban and Regional Research*, vol. 28, no. 2, pp. 287-307, 2004.
[18] Prabhu S. and Venkatesan N., *Data Mining and Warehousing*, New Age International Publishers, 2006.
[19] Schwarz N., "Urban Form Revisited-Selecting Indicators for Characterizing European Cities," *Landscape and Urban Planning*, vol. 96, no. 1, pp. 29-47, 2010.
[20] Teknomo K., Hierarchical Clustering Tutorial, *http://people.revoledu.com/kardi/tutorial/Clustering/index.html*, Last Visited 2014.
[21] Wu R. and Chou P., "Customer Segmentation of Multiple Category Data in E-commerce Using a Soft-Clustering Approach," *Electronic Commerce Research and Applications*, vol. 10, no. 3, pp. 331-341, 2011.

**Suat Secgin** gained his BSc degree from Dokuz Eylul University at the department of Electrical and Electronics Engineering in 1992. He also gained his MSc degree from the same university's Computer Engineering Department with the thesis of Mobile Networks and Data Access Strategies. Currently he is a Phd student in the Dokuz Eylul University Computer Engineering department. He is a member of Electrical Engineering Camperships and also has been working for Turk Telekom. Some of his research areas is traffic engineering in packet based networks, wireless networking and data mining.

**Gokhan Dalkilic** received BS degree in Computer Engineering from Ege University, Izmir, Turkey, in 1997, MS degrees in Computer Science from University of Southern California, Los Angeles, USA, in 1999, and from Ege University International Computing Institute, Izmir, Turkey, in 2001, and Ph.D. degree in Computer Engineering from Dokuz Eylul University, Izmir, Turkey, in 2004. He had been a visiting lecturer in University of Central Florida, Orlando, USA from January 2003 to December 2003. He has been an Assistant Professor of the Department of Computer Engineering of Dokuz Eylul University, Izmir, Turkey since 2004. His research areas are cryptography, statistical language processing and computer networks. His fields of studies are lightweight authentication, cryptography, and NLP. He has over 50 papers and four books to his name.

## Appendix

### A. The CNBC-E Data

The total numbers of 35 rows are shown in the following. The missing values for some cities were replaced with the average number of other cities. The ranking values of the data constitute the table that is being used.

1. Unemployment rate.
2. Amount of tax per capita.
3. Deposit amount per capita.
4. Public expenditure per capita.
5. Number of cars per adult.
6. Number of house per capita.
7. Competitiveness.
8. Average per capita expenditure for rental.
9. Air utilization rate.
10. Household per capita consumption of electricity.
11. Rate of university graduates.
12. Literacy rate.
13. Rate of pre-school students per teacher.
14. Number of pre-school students per classroom.
15. Rate of primary school students per teacher.
16. Number of primary school students per classroom.
17. Rate of secondary school students per teacher.
18. Number of secondary school students per classroom.
19. Number of people per doctor.
20. Number of hospital beds per capita.
21. Crime rate.
22. Earthquake risk.
23. Rate of traffic accident per vehicle.
24. Forest area ratio.
25. Air quality.
26. Divorce ratio.
27. Rate of shopping centers per urban area and population.
28. Rate of 5 stars hotels per urban area and population.
29. Rate of licensed sportsmen per population.
30. Rate of number of library and art work per population.
31. Rate of number of visitors to museums per population.
32. Rate of theatre audience per population.
33. Theatre seat capacity rate per population.

34. Rate of cinema audience per population.
35. Customer satisfaction performance order.

## B. Tree View of the Clusters for Different K Values


Figure 15. Cluster 147 (Tree depth is 2).

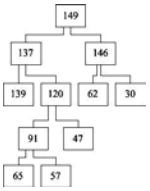


Figure 16. Cluster 149 (Tree depth is 2).

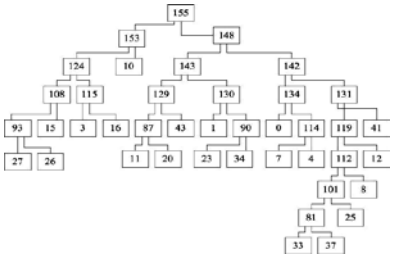


Figure 17. Cluster 155 (Tree depth is 2).

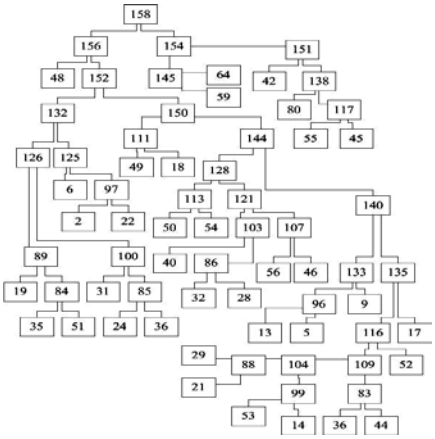


Figure 18. Cluster 158 (Tree depth is 2).

## C. Further Clusters

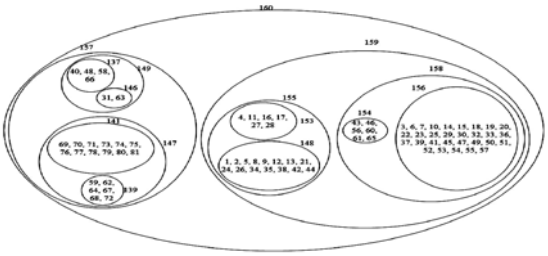

Figure 19. Nested clusters for k=8.

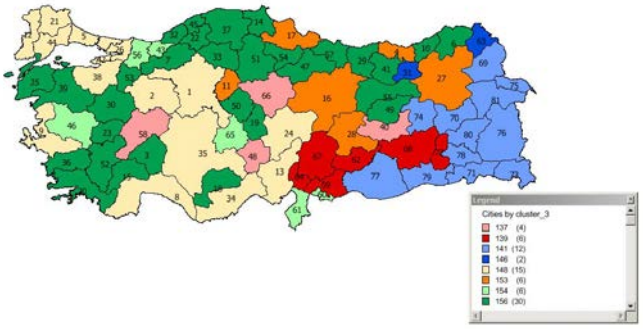


Figure 20. The city distribution for k=8.


Figure 21. The city distribution for k=16.


Figure 22. The city distribution for k=27.

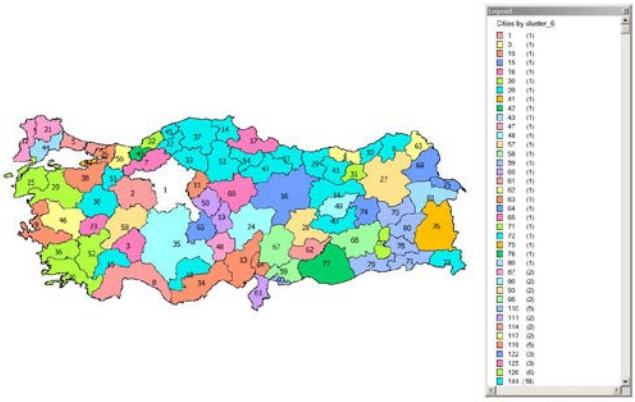


Figure 23. The city distribution for k=40.