DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

SUPPORTING MANAGERIAL DECISIONS USING DATA MINING: EVIDENCE FROM TELECOMMUNICATION INDUSTRY IN TURKEY

by
Suat SEÇGİN

February, 2016
İZMİR

SUPPORTING MANAGERIAL DECISIONS USING DATA MINING: EVIDENCE FROM TELECOMMUNICATION INDUSTRY IN TURKEY

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Computer Engineering Program

by
Suat SEÇGİN

February, 2016
İZMİR

Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"SUPPORTING MANAGERIAL DECISIONS USING DATA MINING: EVIDENCE FROM TELECOMMUNICATION INDUSTRY IN TURKEY"** completed by **SUAT SEÇGİN** under supervision of **ASSIST. PROF. DR. GÖKHAN DALKILIÇ** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Assist. Prof. Dr. Gökhan DALKILIÇ
Supervisor

Prof. Dr. Yalçın ÇEBİ
Thesis Committee Member

Assoc. Prof. Dr. Sabri ERDEM
Thesis Committee Member

Assoc. Prof. Dr. Onur Raif ÖNVURAL
Examining Committee Member

Assoc. Prof. Dr. Bekir Taner DİNÇER
Examining Committee Member

Prof. Dr.Ayşe OKUR
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to express my appreciation to Assistant Professor Dr. Gökhan Dalkılıç for his supportive guidance throughout this research.

I am grateful to members of the advisory committee Professor Dr. Yalçın Çebi and Associate Professor Dr. Sabri Erdem for their great support to this study.

I would also thank to Associate Prof. Dr. Güzin Özdağoğlu, for sharing her knowledge with me and for her support during the development of the data mining application.

Thanks to my colleagues in the company for which I am working for their great support to complete this thesis.

Suat SEÇGİN

SUPPORTING MANAGERIAL DECISIONS USING DATA MINING: EVIDENCE FROM TELECOMMUNICATION INDUSTRY IN TURKEY

ABSTRACT

Since 1970's, a lot of research and studies have been done on the customer behavior to achieve customer satisfaction and loyalty. The brand perception, service and product quality and after sales services are the main parameters that influences the customer satisfaction. The main part of the researches account for these topics.

The demographic distribution of people by cities is an important parameter to address the people's behavior. To distinguish people behavior is useful for companies to understand the customer behavior. A system depicting the customer behaviors in the geographically manner could be used by company's executives as a decision support system.

This study investigates the answer of how the demographic structure of the society affects the customer satisfaction. A fixed line telephone operator's customer satisfaction scores for 81 cities in Turkey and the research results of the habitability of Turkey is used to complete the investigation. The investigation is achieved on seven separate categories about the cities such as habitableness, economic development, educational status, health development, cultural-art level, security and city life. While examining the relations, Spearman's rank correlation coefficient is used. The results of the study showing the relations between the habitableness issues and customer satisfaction scores can be used by the executives in decision- making processes towards right actions for improving customer satisfaction and loyalty.

Additionally, a case study covering all cities in Turkey and measuring 35 items for each city is handled. The 35 items are the categories of seven main topics. By using these topics and the cities, geographically clustering of cities is investigated. The people living in a city are likely to be in the same cluster. After clustering and showing the results in a map depicting the cities, the graphics and tables can be used

as part of a decision support system utilizing data mining clustering techniques and geographic information systems.

In doing these, because of its efficiency, the agglomerative hierarchical clustering method and the k-medoids clustering method in Rapidminer data mining software are used. The cophenetic correlation coefficient is used to measure the efficiency of the agglomerative clustering algorithm. After clustering, the results are migrated into a geographic information system tool to depict the results in the map of Turkey. The results show that the cities located in the same geographical areas are in the same clusters showing similar patterns with some exemptions. On the other hand, some cities that are in different geographical areas could show the same characteristics.

Keywords: Customer behavior, Spearman rank correlation coefficient, decision support systems, data mining, clustering algorithms, geographic information systems

YÖNETSEL KARARLARIN VERİ MADENCİLİĞİ KULLANILARAK DESTEKLENMESİ: TÜRKİYE TELEKOMÜNİKASYON SEKTÖRÜNDEN BULGULAR

ÖZ

1970'lerden bu yana, müşteri memnuniyeti ve sadakatini elde edebilmek için, müşteri davranışları üzerine birçok araştırma yapılmıştır. Marka algısı, servis ve ürün kalitesi, satış sonrası hizmetler müşteri memnuniyetini etkileyen en önemli parametreler olarak alınabilir. Araştırmaların ana bölümleri bu başlıklar için yapılmaktadır.

Şehir bazlı demogratif dağılım, bu şehirlerde yaşayanların davranışlarını adreslemede önemli bir parametredir. İnsanların davranışlarını ayırt edebilmek müşteri davranışlarını anlamak açısından şirketler için yararlı bir uygulamadır. Coğrafik olarak müşteri davranışlarını gösteren bir sistem şirket yöneticileri tarafından bir karar destek sistemi olarak kullanılabilir.

Bu çalışmada, toplumun demografik yapısının müşteri memnuniyetini nasıl etkilediği incelenecektir. Bunun için, Türkiye'de bulunan 81 ilin, bir sabit hat operatörünün müşteri memnuniyeti skorları ve yine Türkiye için gerçekleştirilen yaşanabilirlik sonuçları kullanılacaktır. İnceleme, yaşanabilirlik, ekonomik gelişmişlik, eğitim durumu, sağlık gelişmişliği, kültür-sanat seviyesi, güvenlik ve şehir hayati başlıklarında 7 kategoride yapılacaktır. Şehirler ve bu başlıklar arasındaki ilişki incelenirken Spearman'ın sıra korelasyonu kullanılmıştır. Yaşanabilirlik parametreleri ve müşteri memnuniyeti skorları arasındaki ilişki, yöneticiler tarafından, müşteri memnuniyeti ve sadakati konularında doğru aksiyon alma karar süreçlerinde kullanılabilir.

Bunlara ilaveten, 81 il ve 35 alt başlığı kapsayan bir durum çalışması yapılacaktır. Yukarıda değinilen 7 başlığın 35 alt başlığı altında şehirlerin coğrafik olarak nasıl kümelendiği incelenecektir. Aynı kümede yer alan şehirlerde yaşayan insanlar aynı

şekilde davranacaktır. Böylece, kümeleme ve sonuçlarında şehirleri gösteren bir haritada çizilmesi sonrasında elde edilen grafik ve tablolar, veri madenciliği kümeleme tekniklerini ve coğrafi bilgi sistemlerini kullanan bir karar destek sistemi olarak kullanılabilir.

Bunu yaparken, derecelendirilmiş verilerdeki etkinliği nedeniyle, Rapidminer veri madenciliği kümeleme yazılımı üzerinde aglomeratif hiyerarşik kümeleme ve k-medoids kümeleme algoritmaları kullanılacaktır. Aglomeratif kümeleme algoritmasının etkinliği ise "cophenetic correlation coefficient" kullanılarak ölçülmüştür. Kümeleme algoritması sonuçları daha sonra bir coğrafi bilgi sistemine girilerek Türkiye haritası üzerinde gösterilmiştir. Sonuçlar göstermiştir ki, birkaç istisna ile aynı coğrafik bölgelerde bulunan iller, aynı kümelerde bulunma eğilimindedir. Bir başka açıdan ise, bir coğrafik bölgede bulunan bir il tamamen farklı bir coğrafyada bulunan bir il ile aynı davranışı gösterebilmektedir.

Anahtar kelimeler: Müşteri davranışları, Spearman rank korelasyonu, karar destek sistemleri, veri madenciliği, kümeleme algoritmaları, coğrafi bilgi sistemleri.

CONTENTS

Page

Ph.D. THESIS EXAMINATION RESULT FORM ii
ACKNOWLEDGEMENTS iii
ABSTRACT iv
ÖZ vi
LIST OF FIGURES xii
LIST OF TABLES xiiv

CHAPTER ONE - INTRODUCTION 1

1.1 Motivation 1
1.2 Overview 1
1.3 Thesis Organization 5

CHAPTER TWO - MIS METHODS, ALGORITHMS AND SOFTWARE USED IN THE STUDY 6

2.1 Data Mining 6
2.1.1 Data Clustering in Data Mining 8
2.1.2 K-Means Clustering Algorithm 9
2.1.3 X-Means Clustering Algorithm 10
2.1.4 K-medoids Clustering Algorithm 10
2.1.5 The Agglomerative Clustering Algorithm 11
2.2 Decision Support Systems 13
2.2.1 Model-Driven DSS 15
2.2.2 Data- Driven DSS 15
2.2.3 Communication - Driven DSS 16
2.2.4 Document - Driven DSS 17
2.2.5 Knowledge - Driven DSS 17

2.3 Statistical Coefficients Used in the Study 18
2.3.1 Spearman's Rank Correlation Coefficient 18
2.3.2 The Cophenetic Correlation Coefficient 20
2.4 Software Packages Used in the Study 21
2.4.1 Rapidminer Data Mining Software 21
2.4.2 SPSS (Statistical Package for Social Sciences) 22
2.4.3 MapInfo GIS Software 23

CHAPTER THREE - CORRELATION ANALYSIS FOR CNBC-E AND CSPS DATA AND CLUSTERING RESULTS 25

3.1 Data Used In the Study 25
3.1.1 The CSPS Data 25
3.1.2 The Habitableness Ranks Data 25
3.2 Correlation Analysis for CNBC-E and CSPS Data 29
3.2.1 Habitableness vs. Customer Satisfaction 29
3.2.2 Economic Development vs. Customer Satisfaction 30
3.2.3 Educational Status vs. Customer Satisfaction 30
3.2.4 Health Status vs. Customer Satisfaction 31
3.2.5 Culture-art Status vs. Customer Satisfaction 32
3.2.6 Security Status vs Customer Satisfaction 32
3.2.7 City Life Status vs. Customer Satisfaction 33
3.3 K-Means, X-Means and K-Medoids Clustering Results of CSPS 33
3.3.1 The Results of K-Means Clustering Algorithm 37
3.3.2 The Results of X-Means Clustering Algorithm 42
3.3.3 The Results of K-medoids Clustering Algorithm 47
3.3.4 Comparison of K-means, X-Means and K-medoids Results 52
3.3.5 Results for Decision Support System 54

CHAPTER FOUR - CLUSTERING THE CITIES USING DEMOGRAPHIC ISSUES IN TURKEY 55

4.1 Applying the Agglomerative and K-medoids Algorithm to Data 56
4.2 The Main Process Design 57
4.3 The Results of Agglomerative and K-medoids Algorithms 59
4.4 The Map Views of The Clustering Results 63
4.5 Cluster Performance and Accuracy Analysis 66

CHAPTER FIVE - CONCLUSION 71

5.1 Conclusion 71
5.2 Future Works 73

REFERENCES 75

APPENDICES 83

LIST OF FIGURES

Page

Figure 2.1 Data mining as a step in the process of knowledge discovery 7
Figure 2.2 The flow chart of the Agglomerative hierarchical clustering algorithm.... 12
Figure 2.3 Components of a decision support system 14
Figure 2.4 The driven models of a decision support system..... 14
Figure 2.5 The cophenetic correlation coefficient operator connection 21
Figure 2.6 The screen view of Rapidminer..... 22
Figure 2.7 The screen view of SPSS..... 23
Figure 2.8 MapInfo GIS software..... 24
Figure 3.1 Cluster centroids for 2010 38
Figure 3.2 Cluster centroids for 2011 39
Figure 3.3 Cluster centroids for 2012 40
Figure 3.4 Cluster centroids for average values 41
Figure 3.5 Cluster centroids for 2010 43
Figure 3.6 Cluster centroids for 2011 44
Figure 3.7 Cluster centroids for 2012 45
Figure 3.8 Cluster centroids for average values 47
Figure 3.9 Cluster centroids for 2010 48
Figure 3.10 Cluster centroids for 2011..... 49
Figure 3.11 Cluster centroids for 2012..... 51
Figure 3.12 Cluster centroids for average scores 51
Figure 4.1 Main process diagram 58
Figure 4.2 Tree view of average link agglomerative clustering algorithm..... 59
Figure 4.3 Nested clusters for k=4 60
Figure 4.4 Cluster 147 (Tree depth is 2)..... 61
Figure 4.5 Cluster 149 (Tree depth is 2)..... 61
Figure 4.6Cluster 155 (Tree depth is 2) 62
Figure 4.7 Cluster 158 (Tree depth is 2)..... 62
Figure 4.8 K-Medoids Clusters for k=4..... 63
Figure 4.9 City Distribution for k=4 (Agglomerative clustering). 64
Figure 4.10 K-medoids results for the data (i.e., CNBC-E Data and CSPS) 64

Figure 4.11 Agglomerative algorithm results for CSPS... 65
Figure 4.12 K-Medoids results for CSPS 65
Figure 4.13 Main accuracy and performance measurement process 66
Figure 4.14 Validation sub processes 67
Figure 4.15 Optimization sub process 67
Figure 4.16 Distinctive parameters for clustering (for k=4) 69

LIST OF TABLES

Page

Table 2.1The major clustering methods.... 8
Table 2.2 Sample data for Spearman test.... 19
Table 3.1 General status.... 26
Table3.2 Economical status.... 26
Table3.3 Educational status.... 27
Table3.4 Health status.... 27
Table3.5 Culture-Art status.... 28
Table3.6 Security status.... 28
Table3.7 Urban life status.... 29
Table 3.8 The relation between the habitableness order of the cities and the customer satisfaction scores.... 29
Table 3.9 The relation between the economic development order of the cities and the customer satisfaction scores.... 30
Table 3.10 The relation between the educational status of the cities and the customer satisfaction scores.... 31
Table 3.11 The relation between the health status of the cities and the customer satisfaction scores.... 31
Table 3.12 The relation between the culture status of the cities and the customer satisfaction scores.... 32
Table 3.13 The relation between the security status of the cities and the customer satisfaction scores.... 32
Table 3.14 The relation between the city life status of the cities and the customer satisfaction scores.... 33
Table 3.15 The City k-means cluster list for 2010-2011-2012 and average's CSPS .34
Table 3.16 The City X-means cluster list for 2010-2011-2012 and average's CSPS 35
Table 3.17 City k-medoids cluster list for 2010-2011-2012 and average's CSPS 36
Table 3.18The Centroid table for 2010's CSPS.... 37
Table 3.19 The Centroid table for 2011's CSPS.... 38
Table 3.20 The Centroid table for 2012's CSPS.... 40
Table 3.21 The Centroid table for average CSPS.... 41

Table 3.22 The Centroid table for 2010's CSPS 42
Table 3.23 The Centroid table for 2011's CSPS 44
Table 3.24 The Centroid table for 2012's CSPS 45
Table 3.25 The centroid table for Average CSPS 46
Table 3.26 The centroid table for 2010's CSPS 48
Table 3.27 The centroid table for 2011's CSPS 49
Table 3.28 The centroid table for 2012's CSPS 50
Table 3.29 The centroid table for average CSPS 52
Table 3.30 The centroid distance table for 2010's CSPS 53
Table 3.31 The density table 53
Table 4.1 Sample view of the data used in the study 57
Table 4.2 Item distribution performance for k value optimization 67
Table 4.3 Accuracy, item distribution and class recall values for k=2, 3, 4, 5 68

CHAPTER ONE
INTRODUCTION

1.1 Motivation

In recent years, great amount of data has started to play important roles in various business areas. As the data size is growing, organization of the data by using various data mining techniques is gaining importance for the companies in the competitive world. The steps of data, information and knowledge take crucial roles especially in decision support processes.

To understand and estimate the customer behavior is important for companies that want to retain existing customer and to acquire new ones. To develop new products, new marketing methods or a technique for customer loyalty, a decision support system can support executives towards making right decisions for their company. In this respect, some statistical methods and data mining techniques can be used to build a decision support system to help with executive decisions.

1.2 Overview

The customer satisfaction issue has been gaining popularity in the companies. Since, customers are the core of a company; it has to be considered as the main focus. Customers need more and better service. Consequently, business owners require guidelines and frameworks to address customer requirements (Azad & Hashemi, 2013; Kuronen & Takala, 2013). Apart from service quality, there are significant and moderate to strong associations between satisfaction levels and firm's financial and market performance especially in today's competitive environment (Williams & Nauman, 2011; Deng et al., 2010). Accordingly, there are a lot of studies focusing on the customer behavior in various sectors (Gupta et al., 2012; Frank & Enkawa, 2008; Kazi, 2013; Tabaei et al., 2011; Siddiqi, 2011; Manani et al., 2013; Bayraktar et al., 2012; Khokhar et al., 2011; Raza et al., 2012). All these studies have focused on the issues such as service presentation, quality indices, prices, after sales support systems, complaint management, etc.

The customer satisfaction issue for a company is the key parameter for retaining customers, customer loyalty and preventing the churn (Kuo et al., 2009, Deng et al., 2010). Keeping the customer perception high should be the focuspoint for a company to survive in the competitive environment. The state of the art companies are moving from sales oriented models to customer-oriented models (Hossain & Leo, 2009; Yi & Gong, 2008). Changes in the World lead change in customer behavior. Initially the major feature of the economy was mass production and consumption of commodities. However, the modern economy is based on production and consumption of increasingly differentiated goods and services. To behave in customer centric issues, companies have started to measure the quality of the goods and services to understand customer behavior, and with that to serve better products and services (Suki, 2011; Rahman et al., 2010).

On the other hand, the customer satisfaction measurement of a company provides key decision support factor to the executives in taking actions to strengthen the weak points of customer relations. The executives have to consider the factors affecting customer perceptions such as channel structure, marketing strategy, human resources, etc. (Ramasubbu et al., 2008).

In countries, the demographic issues of cities play main roles in several aspects such as political elections, economical status, customer and consumer behavior, immigration, urbanization etc. (Cranshaw et al., 2012; Hossain & Leo, 2009; Musso, 2001; Pamuk, 2004; Schwarz, 2010). In addition to these, customer segmentation by using multiple types data, customer differentiation using data mining algorithms and data mining usage in geographical information systems are the popular interdisciplinary research areas (Anselin & Getis, 2010; Chiang, 2011; Garla et al., 2011; Gorsevski et al, 2012; Khan et al, 2014; Wu & Chou, 2011). To understand the behavior of the population of cities of a country will help in deciding customer differentiation, retention and choosing the right selling procedure for sales.

Initially, in addition to investigating the customer behavior on a monthly manner for years 2010, 2011 and 2012, the answer of how the demographic structure of the society affects the customer satisfaction will also be investigated. By using the customer satisfaction measurement results for 81 cities in Turkey for a fixed line operator for 2010 and 2011 (will be abbreviated CSPS standing for customer satisfaction performance scores) and the habitable cities ranks in Turkey (Mavi, 2010; Mavi, 2011), the relation among these measurements is investigated. The relation among researches are compared and investigated by using Spearman's rank correlation test on the software of the SPSS 15.0 (Muijs, 2004). The correlation tests are done on seven main topics (General, Economical, Education, Health, Security, Urban Life and Culture-Art rankings). The results show that there are statistical correlations between general, economic, educational and culture-art ranks and CSPS's results while there are no correlations between others.

Secondarily, as the input data, the monthly customer satisfaction performance scores of 2010, 2011 and 2012 for a fixed telephone line company in Turkey are used. By using k-means, x-means and k-medoids clustering algorithms on Rapidminer software, it is examined if there is any relation among the 81 cities in Turkey.

The x-means, the k-means and k-medoids clustering algorithms run on the Rapid Miner 5.3.13 are used to construct centroid tables, so as to gather similar characteristic of cities. The cluster number is determined as 7, measure type is determined as mixed measures and mixed measure type is determined as mixed Euclidean Distance for all algorithms in Rapidminer software. Additionally, max runs parameter is 10 and maximum optimization steps are chosen as 100.

After showing the relations between the habitableness issues and the customer satisfaction performance scores, by combining the CNBC-E magazine research (Mavi, 2011) and the customer satisfaction performance scores (CSPS) for the 81 cities of Turkey, the Agglomerative hierarchical clustering algorithm is applied to investigate the cities in Turkey in terms of some demographic issues and customer

satisfaction. In the CNBC-E's habitableness research, 34 topics of different demographic issues such as air utilization rate, unemployment rate, number of pre-school student per classroom etc. are taken into consideration in order to sort the cities in terms of these issues. In addition to the CNBC-E research results, the CSPS of a fixed line telecom operator in Turkey is added to ranking titles. All the data are conducted in the year of 2010. By using the CSPS apart from the CNBC-E data, how the customer satisfaction ranks affect the clustering is investigated as well.

By clustering the cities depending on the CNBC-E research and the CSPS, a landscape of Turkey will beacquired. To cluster cities, the Agglomerative hierarchical clustering algorithm and the k-medoids 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 data to the map. As known, the agglomerative clustering methods have three modes such as single link, complete link and average link. After applying these modes, the efficiencies of the modes are calculated by using cophenetic correlation coefficient. Due to its efficiency, the average link mode is used to map the data and interpret the cluster results. As the measure types, the mixed measures and as the mixed measure parameter, the mixed Euclidian has been chosen. After applying the algorithms, the accuracy and item distribution performance are calculated as well as optimizing the cluster number of the k-medoids. It can be seen from the results that, there are explicit relations among the cities spread in the same geographical areas with some exemptions.

By using the maps for several levels, the status for the cities can be investigated by bottom-up manner. In this respect, the study can be considered a mixture of a geographic information system and a data mining system to form a decision support system. The graphics and maps can be used as a decision support system by the executives to address planning of new sales and customer experience strategies.

All results such as correlation tables, trees, clustering results and maps constitute a data driven decision support system. In which, new sales campaigns can be built, new approximation methods can be developed for the customer experience programs by using the outputs of the system.

1.3 Thesis Organization

The thesis is organized into five chapters. In this chapter, we have introduced an overviewabout what is done while motivating of research. The rest of the thesis is organized as follows.

In chapter 2, data mining algorithms, decision support systems, statistical correlation methods and software packages used in the study will be explained.

In chapter 3, the correlation analysis for the habitableness data and customer satisfaction scores will be investigated by using Spearman's rank correlation coefficient and the SPSS software. The results of k-means, x-means and k-medoids methods will be given graphically to show how the cities clustered in terms of geographic distributions.

In chapter 4, clustering the cities by using the agglomerative clustering methods will be studied. In this chapter, the clustering results will be entered to a geographic system to depict the clusters in a map.

In chapter 5, the conclusions and the future works of the thesis will be presented.

CHAPTER TWO
MIS METHODS, ALGORITHMS AND SOFTWARE USED IN THE STUDY

2.1 Data Mining

As long as computer and computerized systems grow, the datasize produced and stored by these systems will increase exponentially. The need to understand this "big data" set that has rich information about companies is a common in the areas of science, engineering and business. To pry out the hidden-knowledge from these complex data is becoming an important tool to compete in today's competitive world. Data mining is a series of processes based on computer methodology to discover knowledge from the data accumulated over the years (Kantardzic, 2011).

Knowledge Discovery from Data (KDD) is a term used in literature for the data mining in some cases. In other words, data mining takes an essential role in KDD process (Figure 2.1) (Han & Kamber, 2006). To find the patterns hidden in a warehouse system and interpret them to knowledge and information is the main goal of a data mining system (Zhao, 2013).

Data mining activities can be categorized in two ways; predictive and descriptive mining methods. In a predictive data mining technique, the system model is described by the data set, whereas in descriptive mining a complex information is produced by using data set (Kantardzic, 2011).

In Figure 2.1, a knowledge discovery process is shown. Cleaning, integration, selection, transformation and mining of the data are the basic steps of knowledge discovery. Consequently, pattern evaluation and knowledge presentation steps are executed respectively (Han & Kanber, 2006).

- Cleaning: inconsistent and noisy data are removed from the data set
- Integration: multiple data sources are unified.
- Selection: data to be used for analysis is selected.

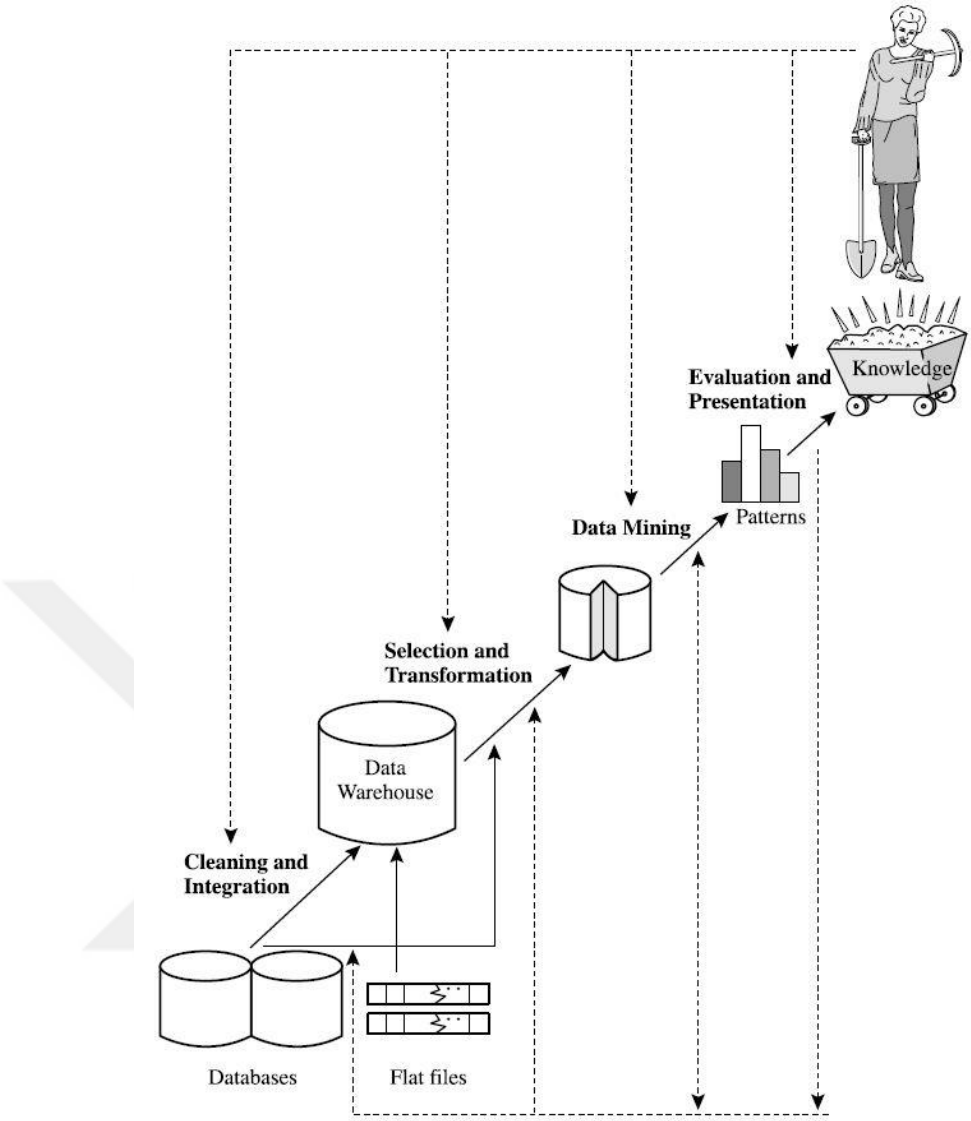


Figure 2.1 Data mining as a step in the process of knowledge discovery (Han & Kamber, 2006)

- Transformation: consolidation and transformation of data tasks are carried out in this step. The data is transformed into forms for mining operations
- Mining: main process where mining algorithms are applied to extract data patterns hidden in data.
- Evaluation of patterns: the extraordinary patterns are evaluated to reveal information.
- Presentation: In this stage, the knowledge is presented to users by using various reporting and presentation techniques.

The primary data mining tasks could be described as classification, regression, clustering, summarization dependency modeling and deviation detection (Kantardzic, 2011).

- Classification: Data is classified to several predefined sub-classes. This task is related to a predictive learning function to be discovered to form the predefined classes.
- Regression: In this stage, the predictive learning function is tried to be discovered. By using this function, data item is mapped to the real-value prediction variable.
- Clustering: Data is clustered in finite set of clusters. Clustering is a descriptive model of data mining.
- Summarization: In this descriptive method, the goal is to find a short version of the data to summarize it.
- Dependency modeling: To describe important dependencies between variables, the dependency modeling is carried out. The dependency modeling is also used to find dependencies between the values of a feature in a data set. Both model could be applied to data itself or a subset of the data.
- Change and Deviation Detection: To find out the extra-ordinary changes in a data set, the change and deviation detection tasksare applied to data.

2.1.1 Data Clustering in Data Mining

As its name implies, to describe the data, data is clustered into finite set of clusters. Clustering is a descriptive model of data mining.

Table 2.1 The major clustering methods (Han & Kanber, 2006)

Partitioning Methods	K-means, X-Means, K-medoids, CLARA, CLARANS, PAM, BICO
Hierarchical Methods	Agglomerative, Divisive, Chameleon, BIRCH, ROCK, CURE
Density Based Methods	DBSCAN, OPTICS, DENCLUE
Grid Based Methods	STING, WaveCluster
Model-Based Methods	EM, COBWEB, SOM

The major clustering methods can be classified as partitioning, hierarchical, density based, grid based and model based (Han & Kanber, 2006). A summary of these methods is given in Table 2.1.

2.1.2 K-Means Clustering Algorithm

Clustering analysis depends mainly on distance metrics and is considered as a unsupervised learning method. Unlike supervised data mining methods such as classification, no predefined classes and training data set are used in unsupervised learning methods to cluster data. In other words, in clustering, learning is done by observation rather than examples (Kantardzic, 2011).

The k-means algorithm is one of the well-known and widely used classical partitional methods. The algorithm takes k value as the number of clusters and the data set to be clustered. Initially, k items from the data set D is chosen arbitrarily as the center of the cluster. Each item in the cluster is clustered based on the distances to the cluster centers. After that, the mean value of each cluster is recalculated based on the current items in the cluster to redistribute to the clusters based on the distances between the item and the new cluster's centroid until the square-error criterion function converges (Formula 2.1).

$$E = \sum_{i=1}^{k} \sum_{p \epsilon Ci}^{n} | p - m_i |^2 \qquad (2.1)$$

Where, E is the sum of the square error for all items in the data set; p is the point in the space representing the given item and m_i is the mean of the cluster C_i (Han & Kamber, 2006). K-means provides high intracluster similarity and low intercluster results.

The algorithm starts with determining the center of each cluster. Then it assigns each object to the cluster by considering similarities calculated using distances. In the next step, centroids positions are re-calculated. Finally, step 2 and 3 are repeated until no change is found in the clusters.

2.1.3 X-Means Clustering Algorithm

In k-means algorithm, the initial selection of centroids is done randomly. In k-medoids algorithm, the initial representative of a cluster is selected from the actual data. However, in x-means algorithm, cluster location space and number of clusters are searched. The aim of the searching process is to optimize the Bayesian Information Criterion which is used to select model among finite set of models (BIC) or the Akaike Information Criterion which is used to measure quality of statistical methods (AIC) (Pelleg & Moore, 2000).

The algorithm has three main steps (Kumar & Wasan, 2010);

- Step 1: Apply the k-means algorithm to data for k clusters until it convergences, where k is equal to lower bound supplied by the user. This step is called as Improve-Params
- Step 2: This step begins by dividing each cluster center into two nodes in opposite directions along a randomly selected vector. Then the k-means algorithm is applied to these two nodes of the cluster. The decision between the nodes of each center and itself is carried out by comparing BIC-values of the two structures. This step is called Improve-structure.
- Step 3: if $k>=k_{max}$ (upper value) stop and report to best scoring model otherwise jump to step 1.

2.1.4 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 is selected as the initial centroid of a 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 (Han &Kamber, 2006).

K-medoids clustering algorithm is as follows (Wyld et al., 2011):

(1) Selection of the k objects in data D is arbitrarily done as medoids points (initial seeds).

(2) After choosing of the k medoids, each data object in the data set is assigned to most similar medoids. Similarity is calculated by using the Euclidian, Manhattan or Minskowski distances.

(3) Randomly select non-representative object ‘O′.

(4) Compute total cost, S of swapping initial representative object (i.e. medoids) to O′.

(5) If S<0, then swap initial medoids with the new one (if S<0 then there will be new set of medoids).

(6) Repeat steps 2 to 5 until S>0 (i.e. no change in centroids)

2.1.5 The Agglomerative Clustering Algorithm

In hierarchical clustering algorithm, a decomposition process is applied to the data set. (Han & Kanber, 2006). A hierarchical algorithm can be divided into two sub-algorithms. The Divisive hierarchical clustering algorithm uses the top-down approach and the Agglomerative hierarchical clustering algorithm uses bottom-up approach (Hamilton, 2012; Kaur & Gupta, 2014). In the agglomerative clustering algorithm objects or groups that has maximum similarity are merged to one other. This process is performed until all groups are merged to just one (Han & Kanber, 2006).

```
Given:
A set of X objects {X1,…,Xn)
A distance function dis(c1, c2)
 for i=1 to n
    Ci={xi}
 end for
 C={c1,…,cb}
 l=n+1
 while C.size>1 do
  (cmin1,cmin2)=minimum dis(ci,cj) for all ci, cj in C
  remove cmin1 and cmin2 from C
  add {cmin1, cmin2}to C
  l=l+1
end while
```

All agglomerative hierarchical clustering algorithms have an initial state with each object as a separate group. Then, these objects are sequentially combined to upper ones by using similarity measurements among them until a specified termination

condition is matched or just one group is remained (Hamilton, 2012). The Agglomerative clustering algorithm is given in Figure 2.2. The nested nature of the agglomerative clustering algorithm in the form of a tree can be valuable for knowledge discovery (Hamilton, 2012; Teknomo, 2012).

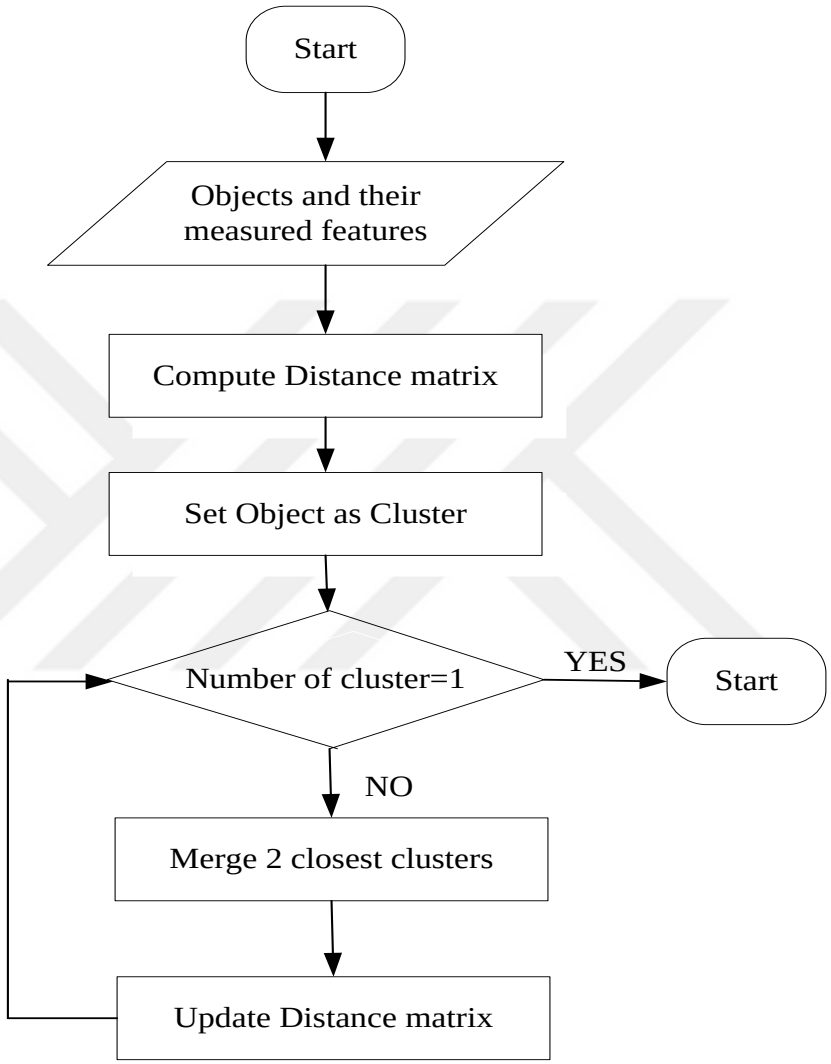


Figure 2.2 The flow chart of the agglomerative hierarchical clustering algorithm (Teknomo, 2012)

As well as hierarchical manner of nested clusters, the clusters are explicitly separated than others by investigating the cluster's result tree. These features of the Agglomerative clustering methods are the main factor to use it in this article. On the other hand, other hierarchical clustering algorithms are designed to address some special cases as in the following;

- Chameleon: discovers dynamic modeling by using sparse graphs (Karypis et al., 1999).
- BIRCH: designed for large amount data, minimizes I/O costs and can handle noise (Zhang et al., 1996)
- ROCK: designed for Boolean and categorical data (Guha et al., 1999)
- CURE: can handle large amount of non-spherical data, wide variances and is robust to outliers (Guha et al., 1998)

As can be seen from the list, Chameleon, BIRCH, ROCK and CURE algorithms are designed to be used for some special cases explained above. Since, our data set is a simple and ranked data set, the agglomerative clustering algorithm is used to cluster the data.

2.2 Decision Support Systems

Beginning from the 1960's, use of computerized quantitative models has been researched by the scientists to support in decision making and planning (Frada & Holsapple, 2008). A decision support system (DSS) is designed to support the managerial decision making processes for non-structured management problems. It depends on computer based information system and it is also a highly flexible, adaptable and interactive information system (Averweg & Richard, 2012).

As one of the core parts of information management system, a decision support system (DSS) can support the executives in decision-making processes. The output reports of the system are used in companies for strategic planning, sales operations, financial planning, etc. Especially in recent year, companies have been gathering great amount of data about their organizations such as financial information, sales operations, customer information etc. A DSS takes a role in a decision making stage by using the collected data. To do this a DSS uses databases, text sources and data warehouses. The output of a DSS could be a graph or a table depicting some relations among the parameters. A conceptual view of a DSS are given in Figure 2.3 (Averweg & Richard, 2012).

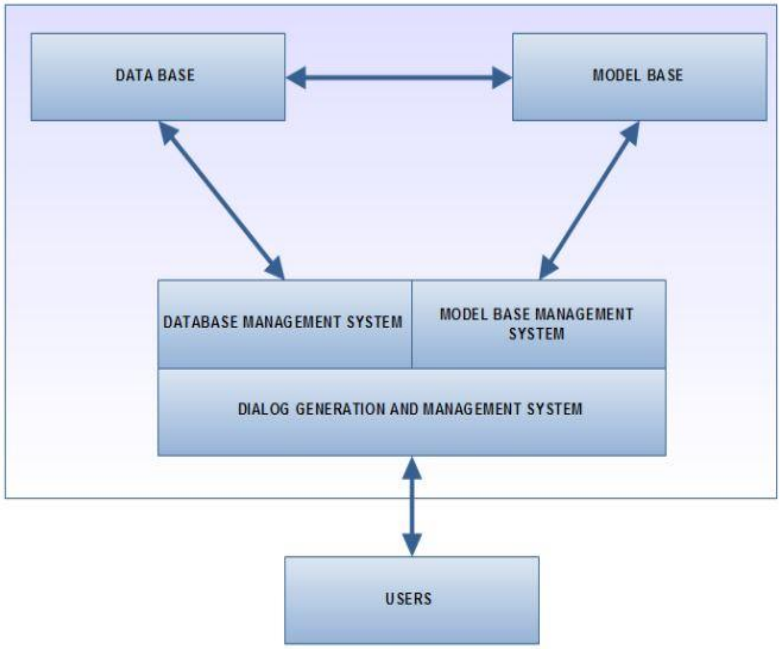


Figure 2.3 Components of a decision support system (Averweg & Richard, 2012)

Dominant DSS Component	**Target Users:** **Internal →External**	**Purpose:** **General → Specific**	**Deployment** **Technology**
Communications **Communications-Driven DSS**	Internal teams, now expanding to external partners	Conduct a meeting or Help users collaborate	Web or Client/ Server
Database **Data-Driven DSS**	Managers, staff, now suppliers	Query a Data Warehouse	Main Frame, Client/Server, Web
Document base **Document-Driven DSS**	Internal users, but the user group is expanding	Search Web pages or Find documents	Web or Client/ Server
Knowledge base **Knowledge-Driven DSS**	Internal users, now customers	Management Advice or Choose products	Client/Server, Web, Stand-alone PC
Models **Model-Driven DSS**	Managers and staff, now customers	Crew Scheduling or Decision Analysis	Stand-alone PC or Client/Server or Web

Figure 2.4 The driven models of a decision support system (Power, 2001)

A DSS falls into five categories: Model driven, data driven, communication driven, document driven and knowledge driven (Frada et al., 2008). In Figure 2.4, these models are compared. The driven models are explained in the following sub-chapters.

2.2.1 Model-Driven DSS

"Model-driven DSS use algebraic, decision analytic, financial, simulation and optimization models to provide decision support" (Power & Sharda, 2007). These systems use accounting and financial models, representational models and optimization models (Power, 2004). As well as forming a "what if" analysis, in this model, system parameters can be altered or manipulated by users, to examine output sensitivity.

The issues that differentiate the model-driven DSS from others are given in (Power & Ramesh, 2007):

- Easy to use user interfaces such as dashboards to a non-technical users
- A specific DSS is designed to be used in similar decision situations

Several types of quantitative models (behavioral models for multi-agent simulations, algebraic and differential equation models, multi-criteria and multi-attribute models, decision trees and decision matrix etc.) are supported in the model-driven decision support system (Power & Ramesh, 2007). Models in a model-driven DSS should be understood by a decision maker to modify the model.

2.2.2 Data- Driven DSS

Business intelligence (BI) is a data-driven DSS that mainly provides historical database querying and produce periodic summary reports. In other words, in data-driven DSS, the main tasks are accessing and altering of company's historical data and external data (Power, 2008). In a data-driven model, data is processed by using database queries, data mining tools and OLAP (Bhargaya et al., 2007).

After introduction of online analytical processing (OLAP) software in 1990s, data driven DSS had gained advanced capabilities. In a successful data driven model, easy and rapid access to well organized big data system is an essential issue (Power, 2008). OLAP provides the tools and instruments to achieve these goals of the data driven DSS.

An OLAP system could be characterized by the features of multi-dimensional conceptual view, link to a variety of data sources, easy to access and understand, support for multiusers, intuitive data manipulation, flexible reporting and analytic capabilities (Codd et al., 1993).

The main features provided by a data driven DSS could be are summarized as ad-hoc data and retrieval, excel integration, data management and summarization, statistical analysis, create data displays, alerts and triggers, design, report design, report generation, report storage, view production reports and view predefined data displays (Power, 2008).

In this study, a RapidMiner process is used to cluster data. The data used in the study is in a form that, 35 titles for the cities is clearly organized. Once the main process is designed, the input file can be manipulated for different cases of interest. Since, well-organized data is mandatory for a data-driven DSS, the DSS designed in the study can be considered as a data-driven DSS. Additionally, some statistical methods such as Spearman's rank correlation coefficient and the cophenetic correlation coefficient are used.

2.2.3 Communication - Driven DSS

Communication-driven DSS is comprised of electronically data sharing systems. The participants of the communication-driven DSS may be different locations in different time intervals. The model provides access to relevant information and tools to multiple decision makers. In this sense, a communication-driven DSS can support more than one person working on a shared task (Bhargava et al., 2007).

Examples for the systems are Microsoft NetMeeting, various video conferencing systems, bulletin boards, forums and whiteboards. In this respect, a communication driven DSS runs on networks and communication infrastructures such as LAN, WAN, Internet, VPN etc. to provide collaboration, communication and decision making (Gachet, 2000; Power, 2007).

2.2.4 Document - Driven DSS

The need for retrieving and managing unstructured data and web pages reveals the document-driven DSS. To achieve complete document retrieval, the model depends on some storage and processing technologies.

By growing web technologies in recent year, a great amount of data such as images, sounds, hypertext and videos has been gathered. So, the document driven DSS uses the source of this big data. With the document driven DSS, various product catalogs, company policies, customer behaviors etc. can be addressed. The web forms as well as internal data are main inputs for the document driven DSS. The document driven DSS can also be used in search engine optimization, advertising procedures and various marketing policies. A search engine could be considered as a strong decision support tool in a document driven DSS (Power, 2001).

2.2.5 Knowledge - Driven DSS

The knowledge driven DSS could be considered as an intelligent decision support system (Power, 2001). A related concept with knowledge driven DSS is Data Mining. Revealing meaningful data and information by using data mining techniques gives the opportunity to managers to make critical and strategic decisions. The sequence of data, information and knowledge is the key concept of knowledge driven DSS.

A knowledge-driven DSS is a knowledge driven decision support system, which has problem solving expertise. The knowledge driven DSS has outputs of suggestions and recommendations. These outputs could be based on several criteria's and requires human computer interaction (Suresh et al., 2012). To find the

knowledge hidden in data set, a knowledge driven DSS receive help from some analytical tools such as data mining techniques. In other words, data mining techniques are often integrated with the knowledge driven DSS. The model also has self-learning capacity and can identify relations between data items. Additionally, the knowledge driven DSS can perform heuristic functions. All these features make the model be an intelligent system. A knowledge driven DSS should has a well-defined knowledge base (Hamad & Qader, 2014).

2.3 Statistical Coefficients Used in the Study

In this chapter the Spearman's rank correlation coefficient and the cophenetic correlation coefficient are explained respectively.

2.3.1 Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient is used in statistics to measure statistical dependence between two variables in nonparametric way. The coefficient is showed as ρ(rho) or r_s. The Spearman's rank correlation coefficient examines the relationship or correlation between two variables using monotonic function (Gautheir, 2001).

The correlation of +1 or -1 take places when no repeated data values exist. In this case, the both of the variables are a perfect monotone function of each other (Spearman, 1904).

Spearman's coefficient can be used for both continuous and discrete variables including ordinal variables (Zar, 1972; Kendall, 1962).

"The idea behind the rank correlation coefficient is simple. Each variable is ranked separately from lowest to highest (e.g. 1, 2, 3 etc.) and the difference between ranks for each data pair is recorded. If the data set is correlated, then the sum of the square of the difference between ranks will be small. The magnitude of the sum is related to the significance of the correlation" (Gautheir, 2001).

Spearman's rank correlation coefficient ρ, is a special situation of the Pearson product-moment correlation coefficient. There are two methods to calculate Spearman's rank correlation coefficient;

First, if you have no tied ranks (Where, d_i is difference in paired ranks)

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{2.2}$$

Second, if you have tied ranks (Where, "i" is paired score)

$$\rho = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}} \tag{2.3}$$

When calculating Spearman's correlation coefficient, scores of each variable are ranked separately and difference between two ranks are calculated for each pair of ranks. Identical values (rank ties or value duplicates) are assigned a rank equal to the average of their positions in the ascending order of the values. In the Table 2.2, there is an example to calculate Spearman's rank correlation coefficient.

Table 2.2 Sample data for Spearman test

Data1	**Data2**	**Rank1**	**Rank2**	**d**
6	2	5	1	4
3	5	2.5	4	-2
1	9	1	5	-4
5	3	4	2	2
3	4	2.5	3	-1

In this example, each (Table 2.2) variable ranked separately and their rank differences are calculated in a new column. The lowest value for the first variable is 1 and the rank value of this value assigned as 1. Next lowest value will have a rank 2 and so on. There are two same values in data1 which is 3 and their positions are 2

and 3. So, mean value of their positions are assigned as the rank of each 3. The same procedure will be followed to calculate the ranks of Data2. After calculating difference values for each pair of ranks, it is straightforward to calculate Spearman's correlation coefficient using the formula in 2.2.

2.3.2 The Cophenetic Correlation Coefficient

The cophenetic correlation coefficient 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 (Teknomo, 2012).

Suppose that the original data $\{X_i\}$ have been modeled using a cluster method to produce a dendrogram $\{T_i\}$; that is, a simplified model in which data that are "close" have been grouped into a hierarchical tree. Define the following distance measures.

- $x(i, j) = |X_i - X_j|$, the ordinary Euclidean distance between the i^{th} and j^{th} observations.
- $t(i, j)$ = the dendo-grammatic 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, letting $\bar{x}$ be the average of the $x(i, j)$, and letting $\bar{t}$ be the average of the $t(i, j)$, the cophenetic correlation coefficient c is given by formula 2.4.

$$c = \frac{\sum_{i<j}(x(i,j)-\bar{x})(t(i,j)-\bar{t})}{\sqrt{[\sum_{i<j}(x(i,j)-\bar{x})^2][(t(i,j)-\bar{t})^2]}} \quad (2.4)$$

In obtaining cophenetic matrix, the lower triangular distance matrix with the minimum merging distance is used. The cophenetic correlation coefficient states correlation between the cophenetic matrix and distance matrix . When the coefficient is close to 1 (%100), this states that the clustering is perfect fit. In Figure 2.5, the operator connection of the cophenetic correlation coefficient in Rapidminer software is given (Dvoroznak, 2014).

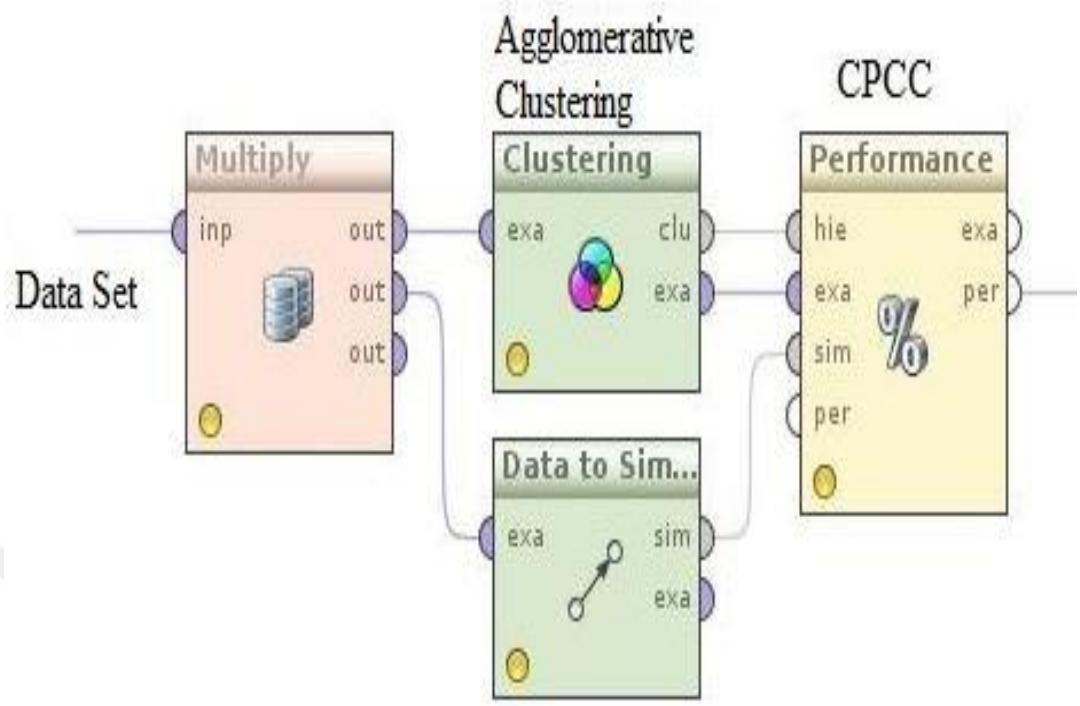


Figure 2.5 The cophenetic correlation coefficient operator connection in Rapidminer

2.4 Software Packages Used in the Study

In this chapter, the software packages used in the study will be examined.

2.4.1 Rapidminer Data Mining Software

Rapidminer is a software tool designed for business analytics and data mining. With the software, a various data mining and business analytics methods such as machine learning, text mining and data mining activities can be done easily. It provides a graphical user interface and drag-and-drop facility to users. A text body, an excel table or a database can be used to an input file for the software. The software has a broad library supporting the data mining algorithms such as regression analysis, clustering and classification methods, support vector machines etc.

It is used for research and training in data mining activities. Rapidminer also provides visualization, optimization and validation operations to users (Land & Fischer, 2012). A sample screen view of the software package is given in Figure 2.6.

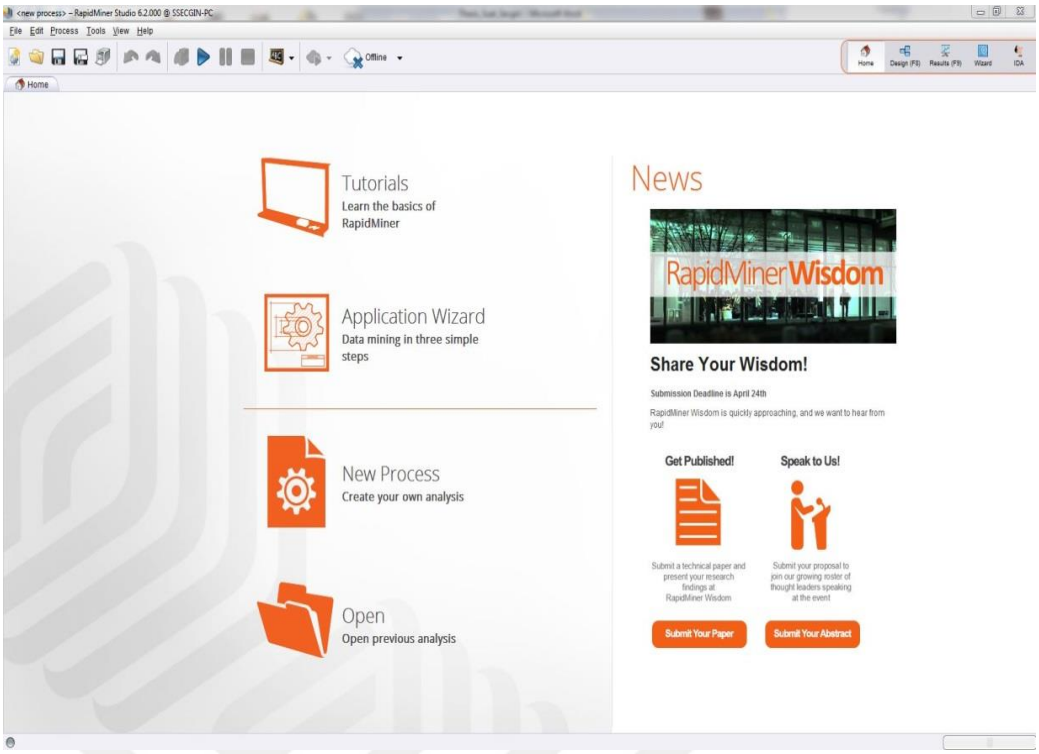


Figure 2.6 The screen view of Rapidminer

2.4.2 SPSS (Statistical Package for Social Sciences)

The Statistical package for social sciences abbreviated as SPSS is a software package in which the statistical methods can be used. It was first released 1968 (Field, 2013). After acquired by IBM in 2009, the software has been officially called as IBM SPSS Statistics. A sample screen view of SPSS is given in Figure 2.7.

SPSS is especially used in the areas of marketing, health, government, data mining, education and training. Data management and documentations are also the key parts of the software.

Statistics included in the base software are descriptive statistics (ratio statistics, cross tabulation, explore and frequencies), identifying group predictions (factor analysis, cluster analysis), bivariate statistics (correlation analysis, nonparametric tests, means, T-test etc.) and numerical outcomes predictions linear regressions).

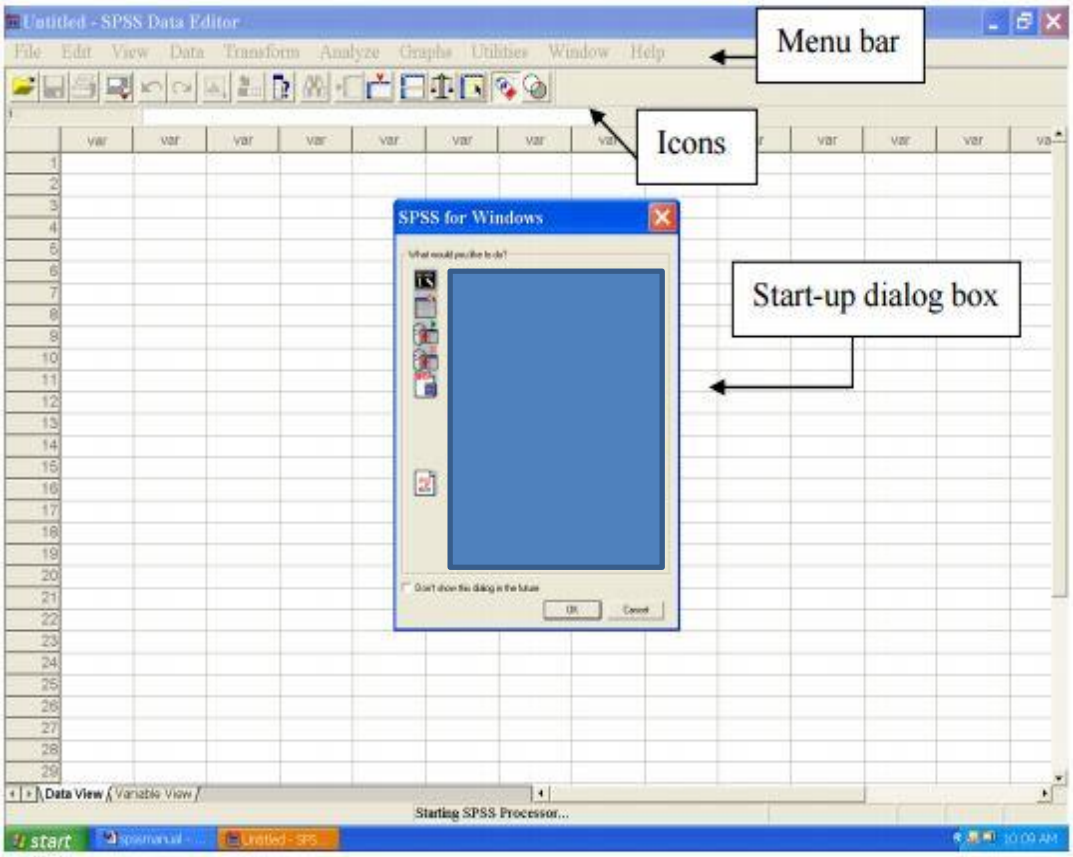


Figure 2.7 The screen view of SPSS

2.4.3 MapInfo GIS Software

MapInfo Professional is a software tool that has designed for examining geographical data. Desktop geographic information system software allows users to analyze, interpret and visualize geographic data. It is produced by Pitney Bowes Software Corporation.

It also allows users to explore spatial data by using data sets and to create geographic maps showing relations among data. A sample screen view of MapInfo is given in the Figure 2.8.

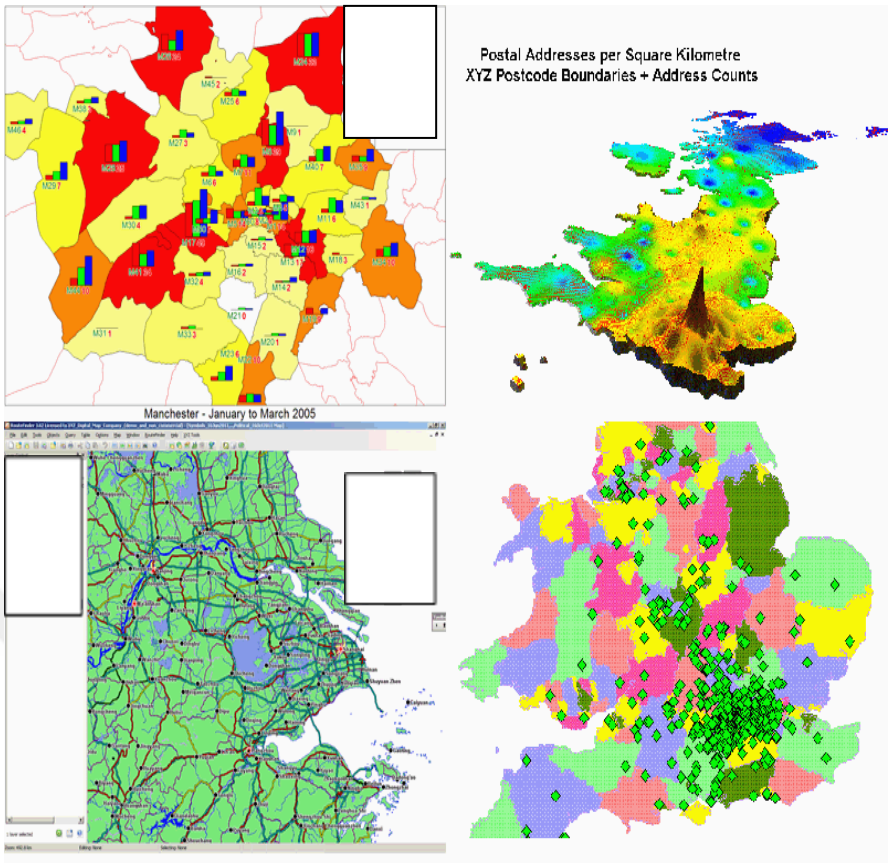


Figure 2.8 MapInfo GIS software

CHAPTER THREE
CORRELATION ANALYSIS FOR CNBC-E AND CSPS DATA AND CLUSTERING RESULTS

3.1 Data Used In the Study

In this study, two categorized data are examined; the monthly average customer satisfaction scores for 2010, 2011 and 2012 of the fixed line operator and the habitable cities evaluation rank in Turkey research of the CNBC-E Business Magazine in 2010 and 2011.

3.1.1 The CSPS Data

In the study, the 2010, 2011 and 2012 monthly customer satisfaction scores of the fixed line operator for 81 cities in Turkey are used for data mining algorithms. The customer satisfaction performance scores (CSPS) has been determined by an independent research company for the operator as follows:

- The computer aided telephone inquiry (CATI) method has been used. The phone inquiries have carried out by experienced CATI agents.
- The 4898 phone inquiries (i.e. 31 calls for each) have carried out for 158 offices in 81 cities in Turkey. The customer satisfaction level of a city has been calculated by using the average results of the offices in related cities.
- Desk waiting time, service quality, employee attitude and knowledge, physical status of an office etc. are the key areas to be inquired.
- To call in every week, a new customer list has been used. The preceding week list is used in the current week. In terms of services they took in the offices, the satisfaction levels of the customers in the list have been inquired. The customers in the list have been casually chosen.

3.1.2 The Habitableness Ranks Data

In the general assessment, the sub-topics (i.e. economy, education, health, culture-art, security and city life) are summated in the rates shown on Table 3.1. After summation, the general order has been calculated for every city in the Turkey. The

general order of a city is calculated by using the rates for each subtopic.

Table 3.1 General status

Topic	Weight
Economy	25%
Education	20%
Health	20%
Culture-Art	8%
Security	15%
Urban Life	12%

In economical assessment, the parameters that are unemployment rate, amount of tax per capita, deposit amount per capita, public expenditure per capita, number of cars per adult, number of residential per capita, competitiveness, average expenditure for rental per capita, usage rate of airlines and household electricity consumption per capita have been taken in to consideration in the rates shown on Table 3.2.

Table 3.2 Economical status

Topic	Weight
Unemployment Rate	10%
Amount of Tax per Capita	15%
Deposit Amount per Capita	15%
Public Expenditure per Capita	10%
Number of Cars per Adult	5%
Number of Residential per Capita	5%
Competitiveness	15%
Average Expenditure for Rental per Capita	10%
Usage Rate of Airlines	5%
Household Electricity Consumption per Capita	10%

In educational assessment, the parameters that are number of graduate students per adult population rate, literacy rate, number of students per teachers in preschool

education, number of classrooms per students in preschool education, number of students per teachers in primary education, number of classrooms per students in primary education, number of students per teachers in secondary education and number of classrooms per students in secondary education have been taken in to consideration in the rates shown on Table 3.3.

Table 3.3 Educational status

Topic	Weight
Number of Graduate Students per Adult Population Rate	40%
Literacy Rate	15%
Number of Students per Teachers in Preschool Education	5%
Number of Classrooms per Students in Preschool Education	10%
Number of Students per Teachers in Primary Education	10%
Number of Classrooms per Students in Primary Education	5%
Number of Students per Teachers in Secondary Education	10%
Number of Classrooms per Students in Secondary Education	5%

In health assessment, the parameters that are number of doctors per capita and number of hospital beds per capita have been taken in to consideration in the rates of 50% as shown on Table 3.4.

Table 3.4 Health status

Topic	Weight
Number of Doctors per Capita	50%
Number of Hospital Beds per Capita	50%

In culture-art status assessment, the parameters that are population rate per number of libraries and art-works, number of museum visitors per population, theatre audience rate of population, theatre seating capacity rate of population, cinema audience of population rate, cinema seating capacity of population rate, opera and bullet audience of population rate and opera and bullet seating capacity of population

rate have been taken in to consideration in the rates shown on Table 3.5.

Table 3.5 Culture-Art status

Topic	Weight
Population Rate per Number of Libraries and Art-works	20%
Number of Museum Visitors per Population	20%
Theatre Audience Rate of Population	10%
Theatre Seating Capacity Rate of Population	10%
Cinema Audience of Population Rate	10%
Cinema Seating Capacity of Population Rate	10%
Opera and Ballet Audience of Population Rate	10%
Opera and Ballet Seating Capacity of Population Rate	10%

In security assessment, the parameters that are crime rate and earthquake risk have been taken in to consideration in the rates of 50% as shown on Table 3.6.

Table 3.6 Security status

Topic	Weight
Crime Rate	50%
Earthquake Risk	50%

In urban life status assessment, the parameters that are traffic accident rate per vehicle, forested land rate, air quality, urban area and population rate per shopping malls, urban area and population rate per five stars hotels and population rate per licensed sportsmen and sportswomen have been taken in to consideration in the rates shown on Table 3.7.

By using the weights below, the habitableness rank (general order) is calculated.

Table 3.7 Urban life status

Topic	Weight
Traffic Accident Rate per Vehicle	20%
Forested Land Rate	20%
Air Quality	20%
Urban Area and Population Rate per Shopping Malls	20%
Urban Area and Population Rate per Five Stars Hotels	10%
Population Rate per Licensed Sportsmen and Sportswomen	10%

3.2 Correlation Analysis for CNBC-E and CSPS Data

On the CNBC-E Business research, seven factors (General, Economical, Education, Health, Security, Urban Life and Culture-Art rankings) are handled. The relations between these factors and the general customer satisfaction scores of Turkey's cities are investigated by using Spearman's rank correlation test on the SPSS software. The main goal is here is to find and examine the effects of the community structure on the choices and behaviors of the customers for the years 2010 and 2011. The first row of the cell refers the 2010 comparison results and the second refers the 2011 comparison results on the tables.

3.2.1 Habitableness vs. Customer Satisfaction

Initially, the relation between the general order of cities in terms of habitable issues (general rank) and the customer satisfaction scores is examined (Table 3.8).

Table 3.8 The relation between the habitableness order of the cities and the customer satisfaction scores

	Satisfaction Scores	Habitableness Rank	p-value
Satisfaction Scores	1	-0.338	0.002
Habitableness Rank	-0.338	1	

As can be seen, there is an inverse proportion between two variables (habitableness and customer satisfaction) with the significance level of 0.05 and p value < 0.05. In other words, the more the habitableness order is, the less the customer satisfaction score is. This is an expected result, since as long as the habitableness level is higher, customer expectations will increased. Consequently, it would be more difficult to satisfy the customers expecting better services and products. When measuring and comparing the customer satisfaction levels for each city, this matter should be taken in to consideration.

3.2.2 Economic Development vs. Customer Satisfaction

The relation between economical order of the cities and the customer satisfaction scores is examined on the following. The test results of the SPSS are given on Table 3.9.

Table 3.9 The relation between the economic development order of the cities and the customer satisfaction scores

	Satisfaction Scores	**Economic Development**	**p-value**
Satisfaction Scores	1	-0.359	0.001
Economic Dev.	-0.359	1	

In according to the test, there is an inverse proportion between two variables (economic development and customer satisfaction) with the significance levelof 0.05 and p-value < 0.05. In other words, the more the economic development order is, the less the customer satisfaction score is. As in the general ranking of the cities, it can be said that, as long as the economic development level is higher, customer expectations will increase. So this result is also expected.

3.2.3 Educational Status vs. Customer Satisfaction

The relation between educational status of the cities and the customer satisfaction scores is examined on the following. After providingthe data to SPSS, the results are given on the following (Table 3.10).

There is an inverse proportion between two variables (educational status and customer satisfaction) with the significance level of 0.05 and p value < 0.05. In other words, the better the educational status of the cities is, the less the customer satisfaction score is. It can be said that, as long as the educational grade of a customer grows, the number of customers that are conscious increases. Therefore, it becomes hard to meet these customer needs.

Table 3.10 The relation between the educational status of the cities and the CSPS

	Satisfaction Scores	**Educational Status**	**p-value**
Satisfaction Scores	1	-0.358	0.001
Educational Status	-0.358	1	

3.2.4 Health Status vs. Customer Satisfaction

The relation between health status of the cities and the customer satisfaction scores is examined on the following and the results of Spearman's correlation test are given on the following (Table 3.11).

Table 3.11 The relation between the health status of the cities and the customer satisfaction scores

	Satisfaction Scores	**Health Status**	**p-value**
Satisfaction Scores	1	-0.011	0.925
Health Status	-0.011	1	

No relation is found between two variables (health development and customer satisfaction) as the significance levelis 0.05 and p value not less than 0.05. In other words, there is no meaningful relation between the health developments of the cities and the customer satisfaction scores. Actually, it can be expected that, the more the habitableness order is, the more health status rank is. In the same way, the better the health status is, the less the customer satisfaction level is. The reason is that it used just two parameters (i.e. number of doctors per capita and number of hospital beds per capita) to calculate the health status rank.

3.2.5 Culture-art Status vs. Customer Satisfaction

The relation between cultural-art level of the cities and the customer satisfaction scores is examined on the following and the results are given on the Table 3.12.

Table 3.12 The relation between the culture status of the cities and the customer satisfaction scores

	Satisfaction Scores	**Culture-Art**	**p-value**
Satisfaction Scores	1	-0.358	0.001
Culture-Art	-0.358	1	

There is a relation between two variables (cultural-art development and customer satisfaction) as significance levelis 0.05 and p value less than 0.05. In other words, there is meaningful positive relation between the cultural-art developments of the cities and the customer satisfaction scores. This is also an expected result. Since, the more the habitableness order is, the culture-art level of a city is higher with some exceptions. So, it could be said that, the more the culture-art level is, the less the customer satisfaction score is.

3.2.6 Security Status vs Customer Satisfaction

The relation between security and the customer satisfaction scores is examined on the following (Table 3.13).

Table 3.13 The relation between the security status of the cities and the customer satisfaction scores

	Satisfaction Scores	**Security**	**p-value**
Satisfaction Scores	1	0.152	0.180
Security	0.152	1	

There is no relation between two variables (security status and customer satisfaction) as the significance level is 0.05 and p-value not less than 0.05. In other words, there is no meaningful relation between the security developments of the cities and the customer satisfaction scores.

3.2.7 City Life Status vs. Customer Satisfaction

The relation between city life and the customer satisfaction scores is examined on the following. After loading the data to SPSS, the results are given on the following (Table 3.14).

Table 3.14 The relation between the city life status of the cities and the customer satisfaction scores

	Satisfaction Scores	**City Life**	**p-value**
Satisfaction Scores	1	-0.072	0.258
City Life	-0.072	1	

No relation is found between two variables (city life and customer satisfaction) as the significance levelis 0.05 and p value not less than 0.05. In other words, there is no meaningful relation between the city life and the customer satisfaction scores. Actually, it could be expected that there should be an inverse relation between the city life level and the customer satisfaction scores as in general order or economic development order.

3.3 K-Means, X-Means and K-Medoids Clustering Results of CSPS

In this section, customer satisfaction performance scores (CSPS) of the 81 cities in Turkey is examined for the years 2010, 2011, 2012 and for the average of these years. By using k-means, x-means and x-medoids clustering algorithms, monthly and seasonal customer behaviors are investigated for the three years. The results show us a great picture of customer behaviors to maximize the telecom users' loyalty and the efficiency of customer retention operations. The executives could use these results in strategic planning in order to retain and satisfy their customers.

The CSPS's of 2010, 2011, 2012 and the average scores will be examined on k-means, x-means and k-medoids. After getting results, the results are compared. In the algorithm k values is considered as 7, since there are 7 separate geographical districts in Turkey. By doing this, the relations and similarities among the clusters and the districts are searched. The clustering results are given in Tables 3.15, 3.16 and 3.17.

Table 3.15 The City k-means cluster list for 2010-2011-2012 and average's CSPS

	2010		2011		2012		Average	
Cl.	**Cities**	**Count**	**Cities**	**Count**	**Cities**	**Count**	**Cities**	**Count**
0	16, 28, 46, 50 63, 65	8	13, 15, 16, 24 28, 34, 35, 44	8	1, 3, 6, 7, 8 17, 19, 20, 27 33, 37, 40, 48 53, 56, 59, 62 68, 72, 73,78	21	10, 11, 13, 15 15, 16, 28, 34 35, 43, 46, 50 61, 63, 65, 70 74, 80	17
1	18, 26, 27, 36, 48, 69, 79	7	14, 17, 18, 26, 42, 45, 72	7	9, 38, 76	3	14, 18, 26, 31 36, 42, 45, 60	8
2	32, 35, 38, 54, 57, 71, 77	7	6, 10, 11, 22, 23, 48, 48, 50 58, 61, 63, 70 74, 80	14	2, 5, 25, 29, 36, 45, 47, 54, 67, 77	10	2, 8, 27, 29, 30, 47, 54, 59, 69, 77	10
3	1, 2, 3, 8, 17 19, 21, 25, 33 39, 45, 47, 52 55, 59, 64, 68 72, 75, 78	20	1, 2, 5, 19, 21 29, 32, 33, 47 52, 53, 55, 57 59, 68, 69, 71 78, 79	20	10, 11, 12, 13 15, 16, 22, 28 34, 35, 39, 41 43, 44, 46, 49 50, 58, 61, 63 64,65,70,74,79	25	1, 3, 5, 17, 19 21, 32, 33, 51 52, 55, 57, 68 71, 72, 78	16
4	10, 15, 23, 24 30, 34, 43, 53 73, 76, 80	11	4, 31, 36, 60	4	21, 23, 24, 51, 52, 55, 57, 66, 69, 71, 75, 80	12	4	1
5	5, 6, 7, 11, 12 13, 20, 22, 37 40, 41, 44, 49 51, 56, 58, 60 61, 62, 66, 67	21	3, 7, 8, 12, 25 30, 37, 38, 40 41, 46, 49, 51 56, 62, 64, 65 66, 67, 73	20	30, 42	2	6,7,12,20,22, 23,24, 25, 37, 38,39,40, 41, 44, 48, 49,53 56,58,62,64 66,67,73,75,79	26
6	4, 9, 14, 29, 31, 42	6	9, 27, 39, 54, 75, 76, 77	7	4, 14, 18, 26, 21, 32, 60	7	9, 76	2

Table 3.16 The City x-means cluster list for 2010-2011-2012 and average's CSPS

	2010		2011		2012		Average	
Cl.	**Cities**	**Count**	**Cities**	**Count**	**Cities**	**Count**	**Cities**	**Count**
0	32, 35, 54, 71, 77	5	28, 49, 70	3	1, 3, 21, 23, 24, 33, 51,52 57, 71, 78	11	1, 3, 5, 17, 19 21, 32, 33, 51 52, 55, 57,68 71, 72, 78	16
1	1, 9, 14, 31, 39, 52, 55, 64, 68	9	14, 17, 18, 26, 69, 72	6	30, 38, 66, 69, 76	5	7, 20, 24, 25, 30, 37, 38, 39 40, 41, 48, 53 56, 62, 64, 66 67, 73, 75, 79	20
2	2, 3, 8, 18, 19 21, 26, 27, 33, 36, 45, 47, 57, 72, 75 78, 79	17	3, 7, 8, 12, 20 22, 25, 30, 37 38, 40, 41, 46 48, 50, 56,58 62, 64, 65,66 67,73, 75	24	2, 5, 29, 36, 54, 55, 67, 77	8	2, 8, 27, 29, 47, 54, 59, 69, 77	9
3	4, 29, 42, 69	4	9, 27, 39, 44, 54, 76, 77, 78	8	9, 13, 16, 70	4	9, 76	2
4	5, 6, 7, 11, 12 13, 17, 20, 22 25, 30, 37, 38 40, 41, 44, 49 51, 53, 56, 58 59, 60, 62, 66 67, 73, 76	28	6, 10, 11, 13, 15, 16, 23, 24, 34, 35 43, 51, 61, 63, 74, 80	16	6, 7, 8, 12, 20 22, 25, 37, 40 41, 42, 48, 49 53, 56, 58, 59 60, 61, 62, 65 68, 72, 73	23	6, 10, 11, 12, 13, 22, 23, 35, 44, 46 49, 58, 61, 65, 80	15
5	10, 15, 23, 24 34, 43, 48, 50 61, 74, 80	11	4, 31, 36, 42, 45, 60	6	4, 14, 17, 18, 19, 26, 27, 31 32,45,47, 60	12	4, 14, 18, 26, 31, 36, 42, 45, 60	9
6	16, 28, 46, 63, 65 70	6	1,2,5,19,21,29 32, 33, 47,52 53,55,57,59, 68 71, 79	17	10, 11, 15,28 34, 35, 39, 43 44, 46, 50, 63 64, 74, 75,79 80	17	15, 16, 28, 34, 43 50, 63, 70, 74	9

Table 3.17 The City k-medoids cluster list for 2010-2011-2012 and average's CSPS

	2010		2011		2012		Average	
Cl.	**Cities**	**Count**	**Cities**	**Count**	**Cities**	**Count**	**Cities**	**Count**
0	7, 8,18,19, 21 22, 25, 27 32, 36,37, 38 38, 40, 51, 55,59, 64, 98, 71 ,75	20	50, 70	2	5, 30, 36, 42, 47, 77	6	16, 28, 50, 70	4
1	1, 2, 3, 4, 9 14, 29, 31, 33 42,45, 52, 69, 72, 78	15	39, 71, 78	3	1, 2, 3, 6, 7,8, 17,18,19,20,21 22, 23, 25, 27 29, 32, 33, 37, 39, 40, 48, 51, 54, 55, 56, 59, 62, 64, 67,68, 69, 76, 78	34	1, 2, 3, 5, 17 19, 21, 27, 29 32, 33, 47, 52, 68, 69, 71 76, 77, 78	19
2	5, 13, 53, 30, 44, 57, 62, 67, 76	9	4,14,17,18,29 31, 32, 33,36 42, 45, 47,55 59, 68, 69, 72	17	71	1	9, 20, 22, 24, 44, 49, 67	7
3	6, 10, 11, 12, 15, 20, 24, 34 41, 43, 48,49 49, 50, 53, 56, 58, 61, 63, 66,73, 74, 80	22	26, 54, 57, 79	4	79	1	7, 8, 23, 25, 30, 37, 38, 39 40, 41, 48, 51 53, 54, 55, 56,57, 59, 62, 64,72, 73, 75	23
4	17, 26, 39, 47, 60, 79	6	5, 9, 27, 60, 75, 76, 77	7	9, 10, 11, 12, 13, 15, 16,28 34, 35, 41,43 46, 49, 50, 63, 65, 70, 74	19	6, 10, 11, 12, 13, 15, 34,35 43, 46, 58,61 63, 65, 66, 74, 80	17
5	35, 54, 77	3	1,2,3,7,8,12,13 19, 20, 21,25,30 30,37,38, 40,41, 46,49,51, 52, 53, 56,58, 62, 64, 65 66, 67, 73	29	4, 14, 26, 31, 45, 60	6	4, 14, 18, 26, 31, 36, 42, 45, 60	9
6	16, 28, 46, 65, 70	5	6, 10, 11, 15, 16, 22, 23, 24 28, 34,35, 43, 44, 48, 61, 63,74, 80	18	24, 38, 44, 52 53, 57, 58, 61 66, 72, 73, 75 80	13	79	1

3.3.1 The Results of K-Means Clustering Algorithm

In this section, the results of K-means algorithm are given. The algorithm is applied to the data for the years 2010, 2011, 2012 and the average. After getting centroid tables, monthly changes of the customer satisfaction of the cities are commented. In the algorithm, k is taken as 7, as there are 7 geographical areas in Turkey.

3.3.1.1 Year 2010

- The customers of the cities in the cluster 0 (16, 28, 46, 50, 63, 65) are remarkably satisfied.
- On the other hand, the customers in the cluster 6 (4, 9, 14, 29, 31) are remarkably dissatisfied along the year.
- Although almost all clusters are down in November and December, the cities in the cluster 0 (16, 28, 46, 50, 63, 65, 70, 74) are up in terms of satisfaction.
- These results are shown in Table 3.18 and Figure 3.1.

Table 3.18 The Centroid table for 2010's CSPS

MONTH	C_0	C_1	C_2	C_3	C_4	C_5	C_6
JANUARY	83.9	74.1	85.2	79.3	81.0	83.5	76.0
FEBRURARY	86.9	81.5	80.6	80.5	82.5	82.6	78.9
MARCH	86.7	81.7	81.8	79.8	83.0	81.8	78.3
APRIL	86.6	80.3	81.0	81.5	81.5	84.9	78.5
MAY	85.4	80.6	83.4	81.1	82.7	83.5	78.4
JUNE	82.8	78.3	79.5	81.5	82.5	80.9	74.4
JULY	84.4	80.7	78.2	79.6	84.8	81.9	76.2
AUGUST	82.6	79.1	79.0	80.2	82.4	81.7	77.2
SEPTEMBER	84.3	81.4	80.9	77.6	80.9	80.0	73.6
OCTOBER	83.0	78.5	77.7	77.9	84.2	80.9	74.1
NOVEMBER	85.5	80.0	77.4	77.9	82.8	81.6	76.7
DECEMBER	85.0	78.4	79.0	77.4	80.0	80.1	74.8

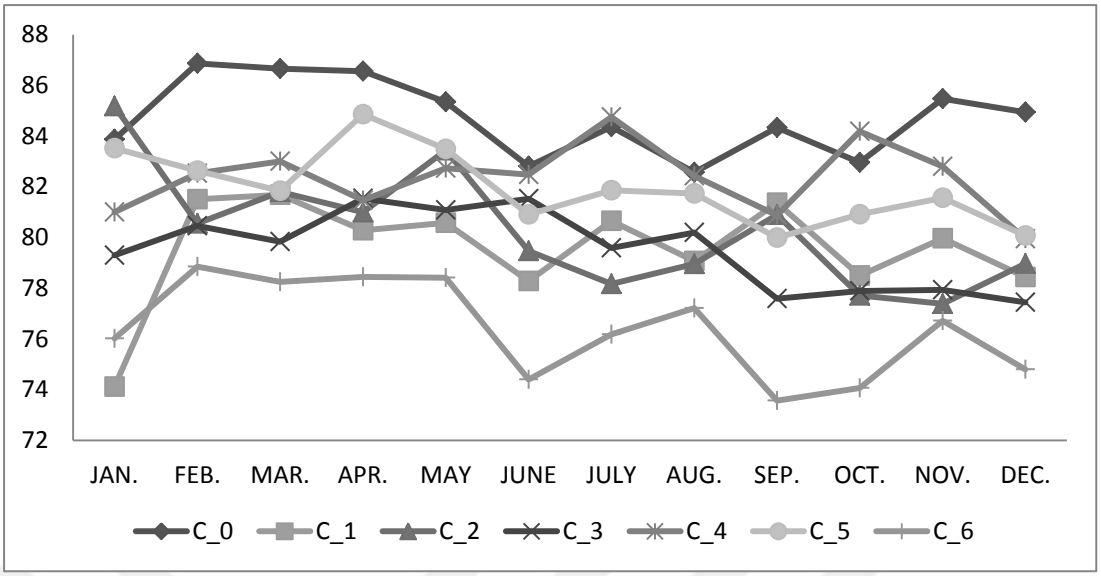


Figure 3.1 Cluster centroids for 2010

3.3.1.2 Year 2011

Table 3.19 The Centroid table for 2011's CSPS

MONTH	C_0	C_1	C_2	C_3	C_4	C_5	C_6
JANUARY	81.2	76.0	82.4	77.2	76.1	80.1	76.4
FEBRURARY	78.4	75.6	80.8	75.8	71.3	78.9	78.1
MARCH	81.0	72.8	80.1	76.9	74.6	79.4	77.4
APRIL	79.2	74.1	80.5	76.9	72.3	77.5	70.5
MAY	82.2	75.4	82.0	79.7	74.8	80.8	77.6
JUNE	80.2	76.5	81.7	79.0	72.9	80.3	78.8
JULY	81.9	75.3	82.6	78.1	74.2	78.8	79.2
AUGUST	80.8	73.9	81.6	76.6	72.3	79.4	78.4
SEPTEMBER	80.6	73.7	78.5	74.4	66.5	77.9	76.2
OCTOBER	80.6	71.7	79.0	74.4	67.7	76.6	78.3
NOVEMBER	82.9	70.0	79.2	76.0	69.6	76.4	75.7
DECEMBER	80.5	74.4	79.7	76.1	66.6	77.9	78.2

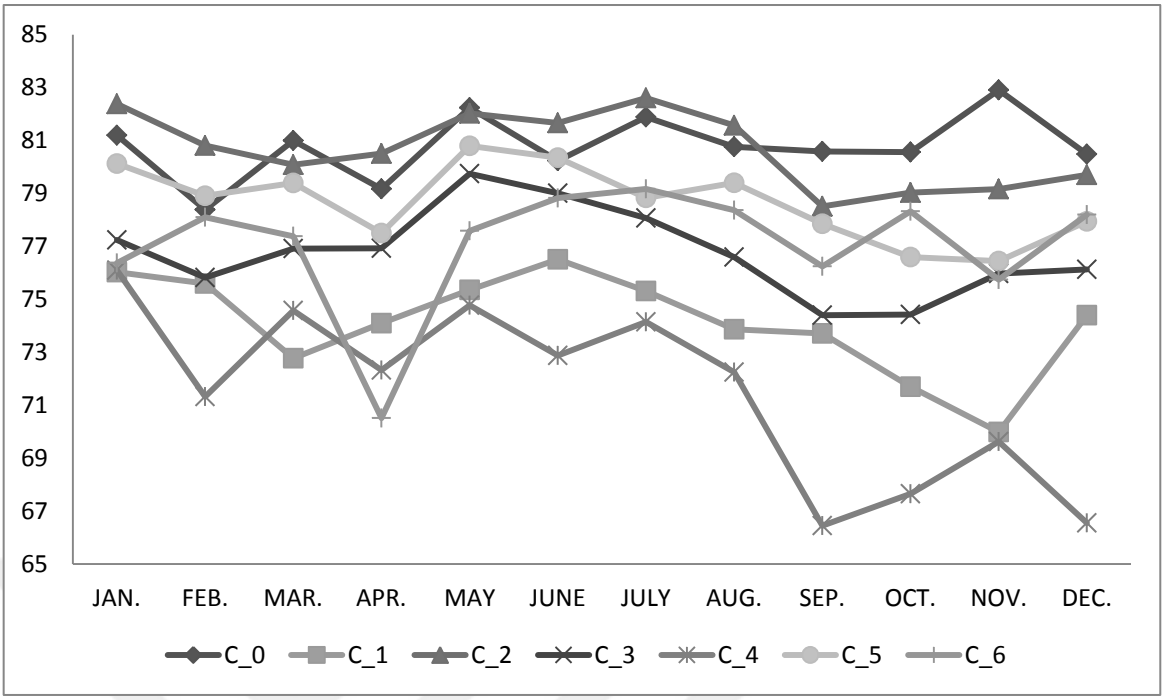


Figure 3.2 Cluster centroids for 2011

- The customers of the cities in the cluster 0 are remarkably satisfied.
- On the other hand, the customers in the cluster 4 (4, 31, 36, 60) are remarkably dissatisfied along the year.
- The customer satisfaction level of the cities in the cluster 6 has a negative peak in April.
- These results are shown in Table 3.19 and Figure 3.2.

3.3.1.3 Year 2012

- The satisfaction levels of the cities are remarkably changing along the year.
- The cities in cluster 1 (9, 38, 76) has positive peak in no accordance with the year.
- It can be said that the customer satisfaction levels of the clusters show instability along the year.
- These results are shown in Table 3.20 and Figure 3.3.

Table 3.20 The Centroid table for 2012's CSPS

MONTH	C_0	C_1	C_2	C_3	C_4	C_5	C_6
JANUARY	78.0	81.6	75.6	79.9	77.5	72.5	73.0
FEBRURARY	78.3	76.7	74.1	80.3	80.7	80.1	73.2
MARCH	76.9	91.3	77.7	79.9	78.6	74.7	73.7
APRIL	75.3	74.2	74.0	77.3	80.2	71.9	68.6
MAY	78.9	82.3	77.2	82.8	80.7	80.3	71.7
JUNE	81.2	83.1	73.6	83.4	78.6	74.6	77.9
JULY	79.1	78.5	77.9	81.9	76.6	76.7	74.1
AUGUST	79.7	77.4	74.2	78.9	75.9	85.5	80.5
SEPTEMBER	79.3	72.6	76.4	82.0	79.1	78.7	74.7
OCTOBER	78.0	77.0	78.6	81.2	74.7	71.2	71.6
NOVEMBER	76.4	82.4	77.6	82.3	76.5	85.1	71.7
DECEMBER	79.1	81.4	77.7	82.0	78.2	79.1	74.3

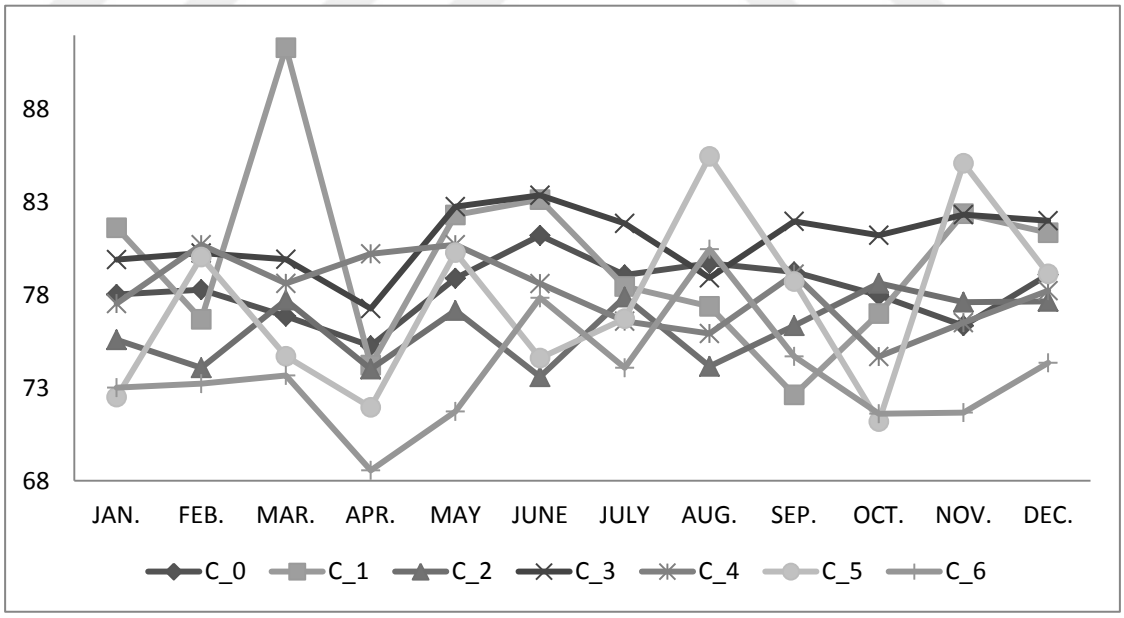


Figure 3.3 Cluster centroids for 2012

3.3.1.4 Average of 2010, 2011, 2012 Scores

Table 3.21 The centroid table for average CSPS

MONTH	C_0	C_1	C_2	C_3	C_4	C_5	C_6
JANUARY	81.4	75.0	77.8	78.8	70.9	80.3	79.7
FEBRURARY	82.0	75.5	78.0	78.4	71.9	80.0	80.3
MARCH	81.9	76.0	77.8	77.7	72.3	80.4	82.6
APRIL	80.8	74.4	75.8	79.0	75.1	79.3	72.1
MAY	83.2	76.9	78.9	80.4	69.4	81.3	80.4
JUNE	82.5	76.6	77.4	80.4	71.3	80.7	80.1
JULY	82.5	75.8	78.8	78.4	71.2	80.4	80.9
AUGUST	81.2	77.1	79.1	77.3	71.1	80.0	76.2
SEPTEMBER	81.8	73.6	77.9	77.4	72.4	78.8	74.9
OCTOBER	81.6	73.3	77.2	76.3	68.1	78.5	79.4
NOVEMBER	82.9	74.4	77.8	75.9	70.8	78.8	79.0
DECEMBER	81.2	73.8	77.3	77.2	73.9	79.9	78.8

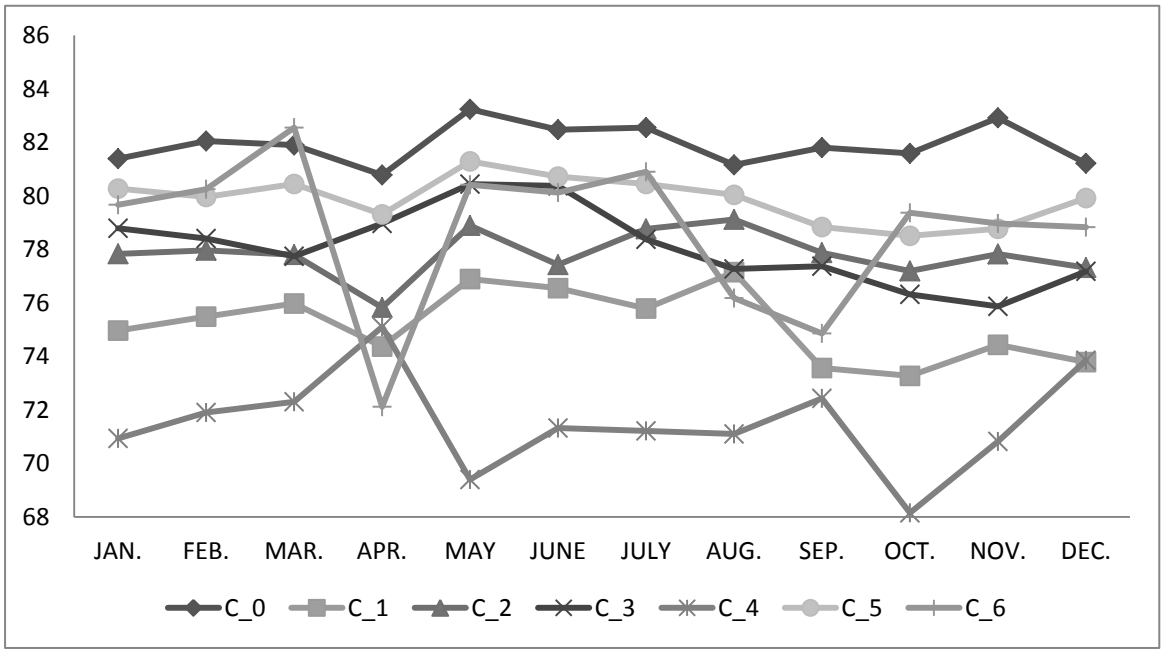


Figure 3.4 Cluster centroids for average values

When taking the average for the years 2010, 2011 and 2012;

- The city in the cluster 4 (4) are remarkably dissatisfied along the year.
- The customer satisfaction level of the cities in the cluster 0 and cluster 5 are high and stable.
- The cities in the cluster 6 (9, 76) have low satisfaction level in April.
- These results are shown in Table 3.21 and Figure 3.4.

3.3.2 The Results of X-Means Clustering Algorithm

In this section, the results of x-means algorithm are given for the years 2010, 2011, 2012 and the average. The number of cluster k is chosen as seven again as there are seven geographical areas in Turkey.

3.3.2.1 Year 2010

Table 3.22 The centroid table for 2010's CSPS

MONTH	C_0	C_1	C_2	C_3	C_4	C_5	C_6
JANUARY	86.4	78.4	77.6	75.5	83.2	79.6	84.9
FEBRURARY	80.0	80.1	80.4	80.7	82.4	83.0	87.7
MARCH	81.7	79.8	79.4	80.2	81.9	84.1	87.1
APRIL	81.0	81.3	79.8	78.9	84.1	83.0	87.5
MAY	84.7	80.4	81.5	75.8	82.7	83.8	85.7
JUNE	79.5	80.2	80.8	71.9	81.2	81.5	83.2
JULY	77.6	78.9	80.1	75.0	81.7	84.2	85.5
AUGUST	79.6	79.3	79.6	76.6	81.9	81.8	82.5
SEPTEMBER	81.5	74.3	78.8	77.8	79.6	82.9	84.5
OCTOBER	77.4	73.7	79.4	76.3	80.7	84.1	82.6
NOVEMBER	77.4	79.3	77.8	75.4	81.0	83.6	85.8
DECEMBER	77.7	75.9	78.1	75.6	79.8	81.4	84.8

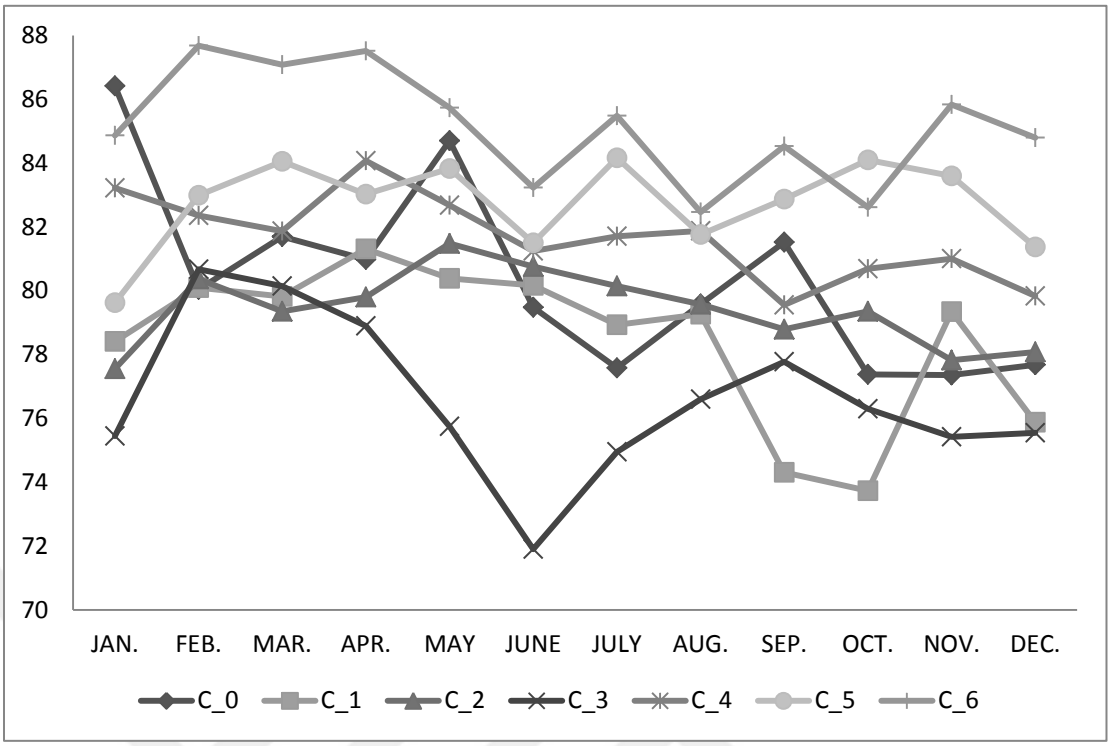


Figure 3.5 Cluster centroids for 2010

- The cities in the cluster 3 are remarkably dissatisfied (29, 42, 69) along the year.
- The customer satisfaction level of the cities in the cluster 6 and cluster 5 are high and stable.
- In September and October, the customer satisfaction level of the cities in the cluster 1 is remarkably declining.
- In June, the satisfaction level of the cluster3 reached the lowest level.
- These results are shown in Table 3.22 and Figure 3.5.

3.3.2.2 Year 2011

- The cities in the cluster 5 are remarkably dissatisfied (4, 31, 26, 42, 45, and 60) along the year.
- The customer satisfaction level of the cities in the cluster 0 and cluster 4 are high and stable.
- The satisfaction level of the cities in the cluster 3 has a negative peak in September.

- The satisfaction level has a positive peak in June for the cluster 0.
- These results are shown in Table 3.23 and Figure 3.6.

Table 3.23 The centroid table for 2011's CSPS

MONTH	C_0	C_1	C_2	C_3	C_4	C_5	C_6
JANUARY	86.8	76.5	80.1	76.7	81.0	75.3	77.6
FEBRURARY	82.0	74.6	79.0	77.3	79.8	73.0	76.0
MARCH	81.6	73.4	79.3	77.2	80.5	73.4	77.1
APRIL	78.6	75.6	77.7	71.5	80.5	71.9	76.8
MAY	82.9	77.3	80.6	77.2	82.4	73.8	80.1
JUNE	86.1	77.1	80.5	78.2	80.6	74.0	78.7
JULY	82.1	77.8	79.3	79.1	82.6	73.3	77.5
AUGUST	81.6	75.0	79.1	78.3	81.6	71.9	76.9
SEPTEMBER	81.1	74.2	77.6	76.2	79.3	68.7	74.3
OCTOBER	81.4	71.6	76.7	78.7	79.5	69.4	74.2
NOVEMBER	81.9	70.8	76.6	77.2	80.6	69.8	75.6
DECEMBER	82.6	74.6	78.0	79.7	79.3	68.5	75.9

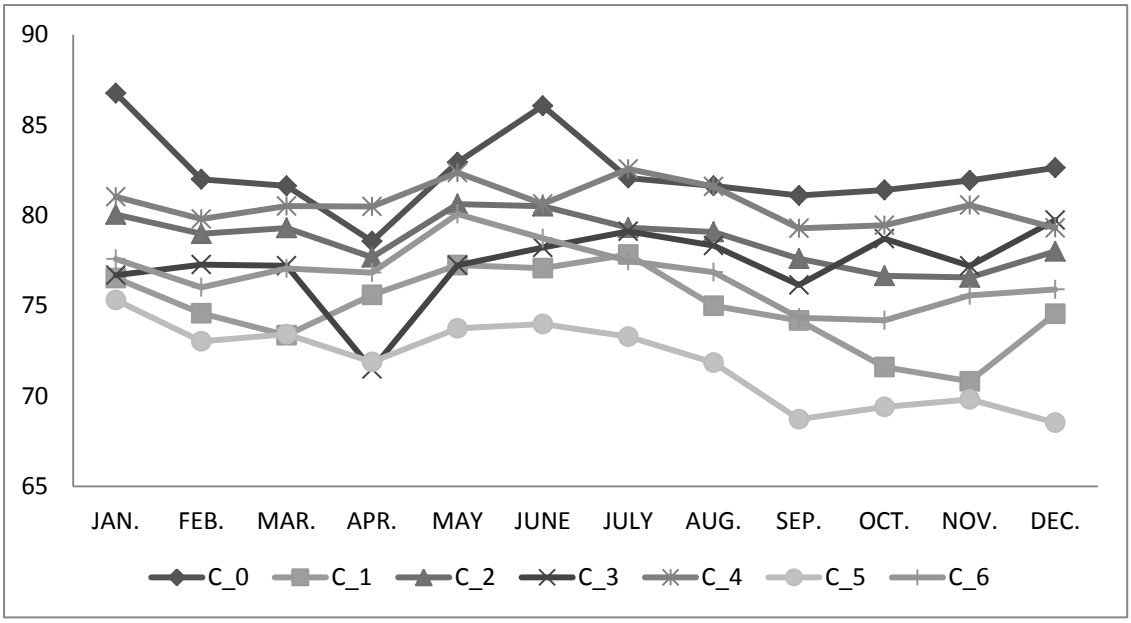


Figure 3.6 Cluster centroids for 2011

3.3.2.3 Year 2012

Table 3.24 The centroid table for 2012's CSPS

MONTH	C_0	C_1	C_2	C_3	C_4	C_5	C_6
JANUARY	77.6	77.3	76.0	81.7	78.5	74.6	79.4
FEBRURARY	79.5	78.5	74.8	82.0	79.1	73.8	80.3
MARCH	77.2	83.7	78.4	85.9	77.4	73.7	80.6
APRIL	79.8	75.9	74.1	71.7	74.8	71.1	79.5
MAY	79.9	80.6	78.0	82.9	79.8	73.6	83.5
JUNE	80.8	78.4	73.8	86.2	80.4	77.8	83.3
JULY	78.1	75.4	77.4	82.3	80.0	74.5	82.0
AUGUST	74.8	82.3	72.3	76.5	81.3	79.4	78.0
SEPTEMBER	79.2	72.5	79.2	82.0	80.0	74.4	82.1
OCTOBER	76.3	73.5	78.7	81.7	78.9	73.2	80.6
NOVEMBER	75.1	81.8	77.4	86.4	78.9	72.3	81.9
DECEMBER	76.4	80.6	78.4	81.3	80.2	76.0	82.2

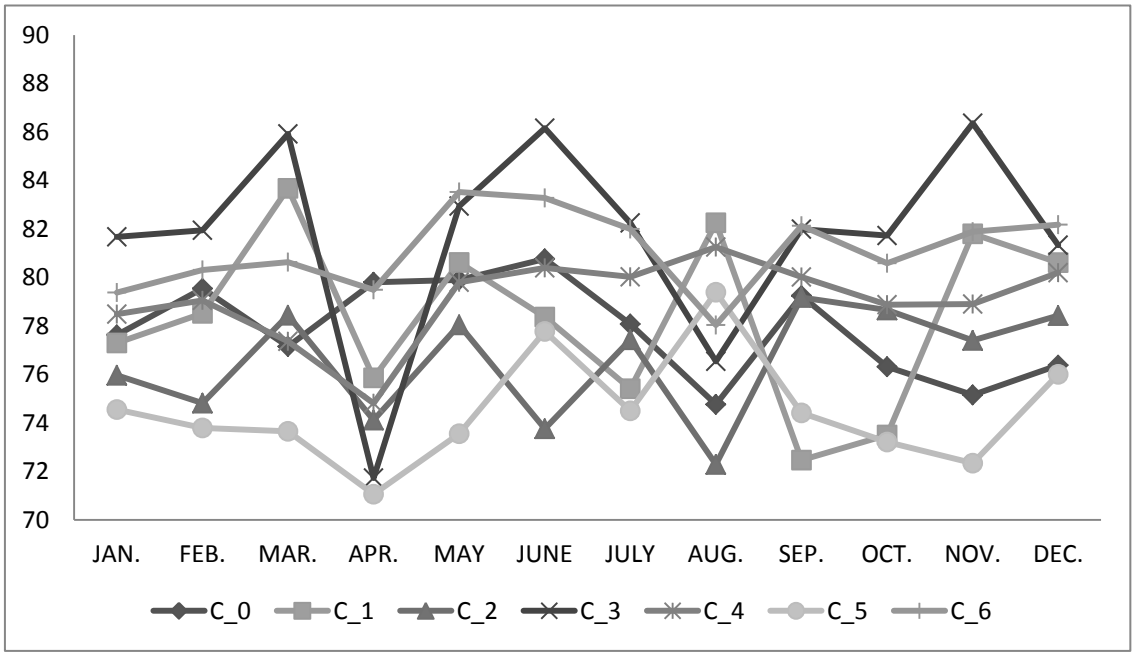


Figure 3.7 Cluster Centroids for 2012

- The satisfaction levels of the cities except the cluster 6 are remarkably changing along the year.
- Almost each cluster has a negative peak in April.
- These results are shown in Table 3.24 and Figure 3.7.

3.3.2.4 Average of 2010, 2011, 2012 Scores

- The satisfaction level of the cities in the cluster 6 is remarkably high for all the months.
- The cities in the cluster 3 (9, 76) has a negative peak value in April and shows no consistency along the year.
- The customer satisfaction level of the cluster 4 is remarkably low.
- These results are shown in Table 3.25 and Figure 3.8.

Table 3.25 The centroid table for average CSPS

MONTH	C_0	C_1	C_2	C_3	C_4	C_5	C_6
JANUARY	78.8	79.7	77.5	79.7	81.4	74.5	81.7
FEBRURARY	78.4	79.6	77.6	80.3	81.2	75.1	82.8
MARCH	77.7	80.5	77.5	82.6	80.9	75.6	82.5
APRIL	79.0	78.6	76.1	72.1	80.7	74.4	80.9
MAY	80.4	80.9	78.9	80.4	82.2	76.1	84.0
JUNE	80.4	80.5	77.5	80.1	81.4	76.0	82.9
JULY	78.4	80.1	78.7	80.9	81.2	75.3	83.7
AUGUST	77.3	80.1	78.6	76.2	80.9	76.5	81.1
SEPTEMBER	77.4	78.7	77.9	74.9	80.0	73.4	82.7
OCTOBER	76.3	77.7	77.4	79.4	80.6	72.7	82.2
NOVEMBER	75.9	78.7	77.6	79.0	81.3	74.0	82.7
DECEMBER	77.2	79.8	77.0	78.8	80.0	73.8	82.6

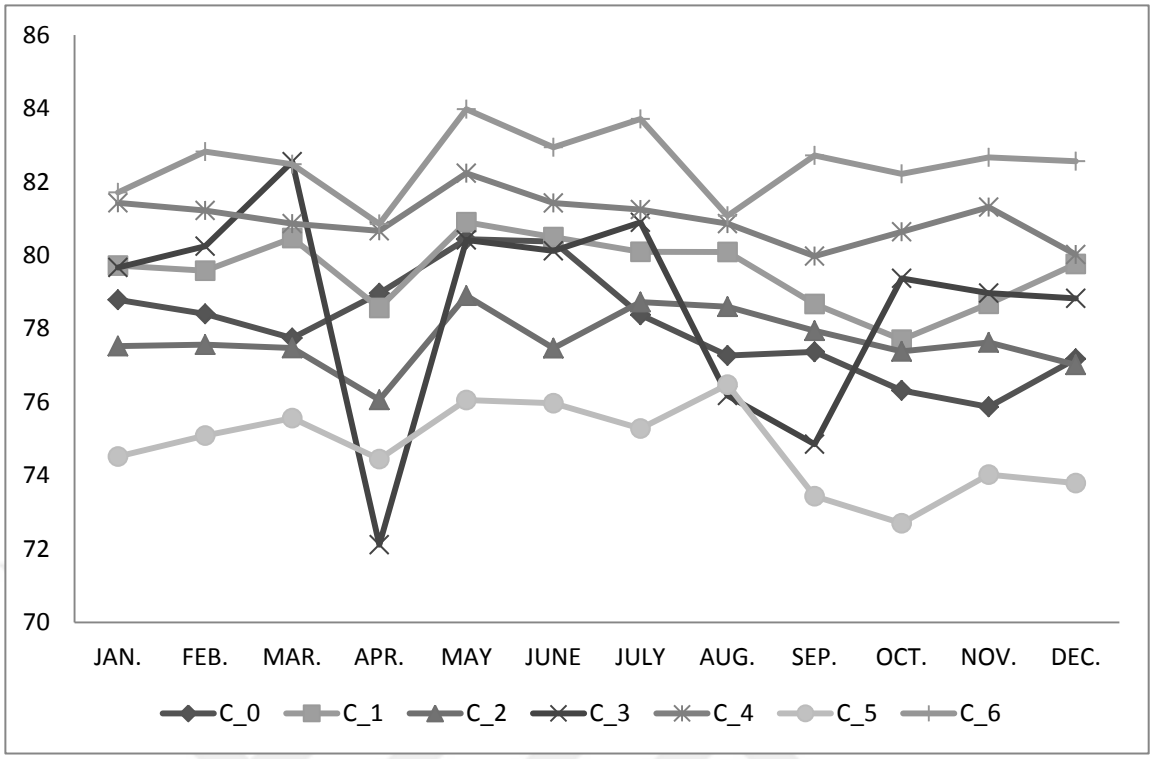


Figure 3.8 Cluster centroids for average values

3.3.3 The Results of K-medoids Clustering Algorithm

The k-medoids algorithm results are given in this section. After getting results for the years 2010, 2011, 2012 and the average, monthly changes of the customer satisfaction of the cities are interpreted.

3.3.3.1 Year 2010

- The satisfaction level of the cities in the cluster 6 is remarkably high for all the months.
- The cities in the cluster 3 (22 cities) have a negative peak value in April and shows no consistency along the year.
- The customer satisfaction level of the cluster 1 is remarkably low.
- The cities in cluster 2 (9 cities) have a peak in October.
- These results are shown in Table 3.26 and Figure 3.9.

Table 3.26 The centroid table for 2010's CSPS

MONTH	C_0	C_1	C_2	C_3	C_4	C_5	C_6
JANUARY	80.2	79	83.3	78.6	72.8	86.6	83.5
FEBRURARY	81.2	81	83.2	83.8	77.5	80.6	89.4
MARCH	79.4	76	79.8	80.8	83.8	85.2	90.9
APRIL	82.1	79.8	79.9	82.1	81.7	77.1	88.9
MAY	82	82.3	80.3	85.8	80.7	82.3	86.2
JUNE	81	81.3	82.6	84	83.8	77.7	83.6
JULY	79.3	76.9	83.4	83.8	84.3	76.4	88.5
AUGUST	84.9	76.9	78.8	83.1	79.4	78.2	82.7
SEPTEMBER	81.7	76.5	78.3	83.3	76.4	84.6	84.9
OCTOBER	78.5	80.3	88.6	81.1	78.7	78.1	85
NOVEMBER	75.3	74.8	80.9	84.4	82	77.6	83.9
DECEMBER	77.3	73.8	77.7	79.2	79.4	80.4	86.2

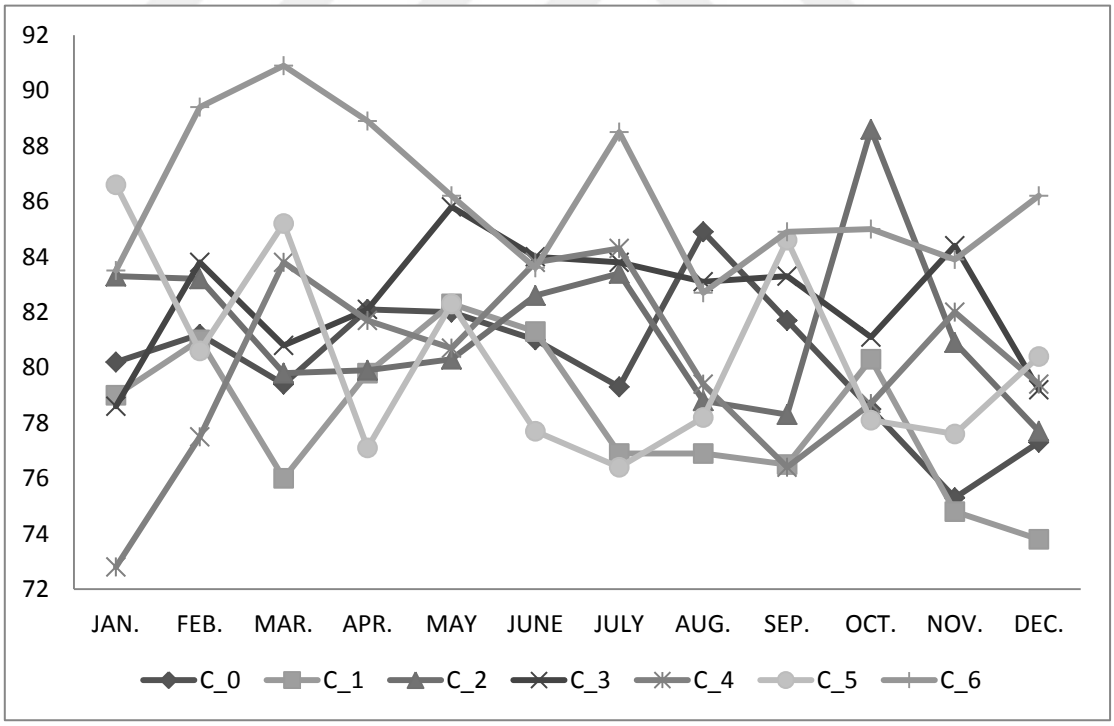


Figure 3.9 Cluster centroids for 2010

3.3.3.2 Year 2011

Table 3.27 The Centroid table for 2011's CSPS

MONTH	C_0	C_1	C_2	C_3	C_4	C_5	C_6
JANUARY	88.4	74.2	78.6	72.6	76.5	80.5	81.6
FEBRURARY	83.1	73.3	76.7	78.7	80.7	78.6	79.3
MARCH	80.6	76	73	80.1	76.9	78.4	81.4
APRIL	78.9	76.1	77.7	76.3	70.6	76.9	81.8
MAY	80.7	78.4	81.3	80.2	78.5	80.8	83.9
JUNE	86.8	81.7	80	76.3	81.7	81.9	80.1
JULY	78.9	80	73.3	73	81	78.8	81.8
AUGUST	80.3	77.6	76.5	77.6	78.9	79.7	81.1
SEPTEMBER	79.1	73.6	73.7	76.1	76.7	78.5	80.5
OCTOBER	76	76.3	75.6	74.9	76.7	76.2	78.7
NOVEMBER	82.2	80.8	74.8	72.7	77	78.8	79.3
DECEMBER	83.2	82.8	74.3	78.7	73.6	75.9	79.4

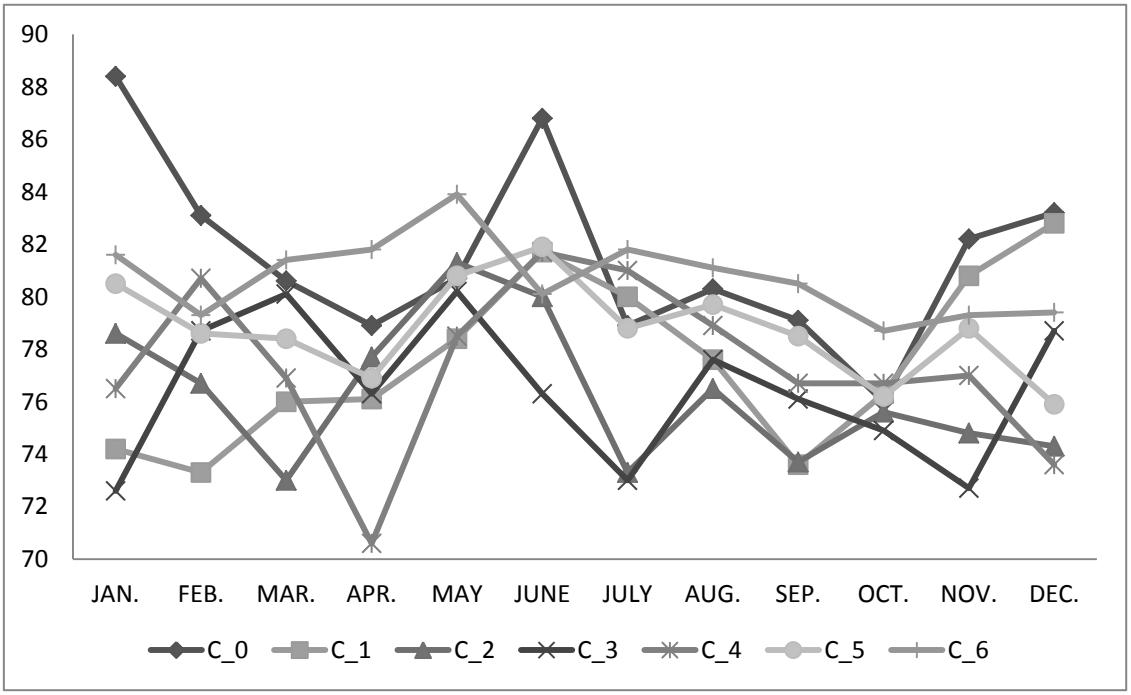


Figure 3.10 Cluster centroids for 2011

- The satisfaction level of the cities in the cluster 0 show an unstable line despite its level is high along the year.
- The cities in the cluster 4 (9, 76) has a negative peak value in April and shows no consistency along the year.
- These results are shown in Table 3.27 and Figure 3.10.

3.3.3.3 Year 2012

- The satisfaction level of all the clusters shows an unstable view along the year.
- The cities in the cluster 5 (6 cities) has a negative peak value in April and shows no consistency along the year.
- The customer satisfaction level of the cluster 5 is remarkably low.
- These results are shown in Table 3.28 and Figure 3.11.

Table 3.28 The centroid table for 2012's CSPS

MONTH	C_0	C_1	C_2	C_3	C_4	C_5	C_6
JANUARY	75.2	80.6	77.4	78.5	78.4	72.4	77.7
FEBRURARY	71.7	77.6	74.8	70.2	81.0	76.4	85.8
MARCH	71.6	78.7	74.1	79.5	82.2	75.3	80.3
APRIL	70.1	76.8	87.5	82.1	76.5	64.2	81.3
MAY	77.4	79.8	86.1	85.8	85.1	67.2	81.2
JUNE	75.3	82.4	81.6	83.2	82.5	72.0	78.1
JULY	80.6	79.4	74.9	82.3	85.2	73.2	77.9
AUGUST	71.9	74.4	78.6	74.3	79.0	80.9	76.7
SEPTEMBER	79.0	76.3	77.3	80.7	85.7	75.2	78.0
OCTOBER	72.3	78.4	72.7	79.2	81.2	72.6	77.9
NOVEMBER	82.5	75.3	76.3	89.1	84.8	69.2	82.4
DECEMBER	78.1	75.6	72.1	87.7	85.1	73.5	80.5

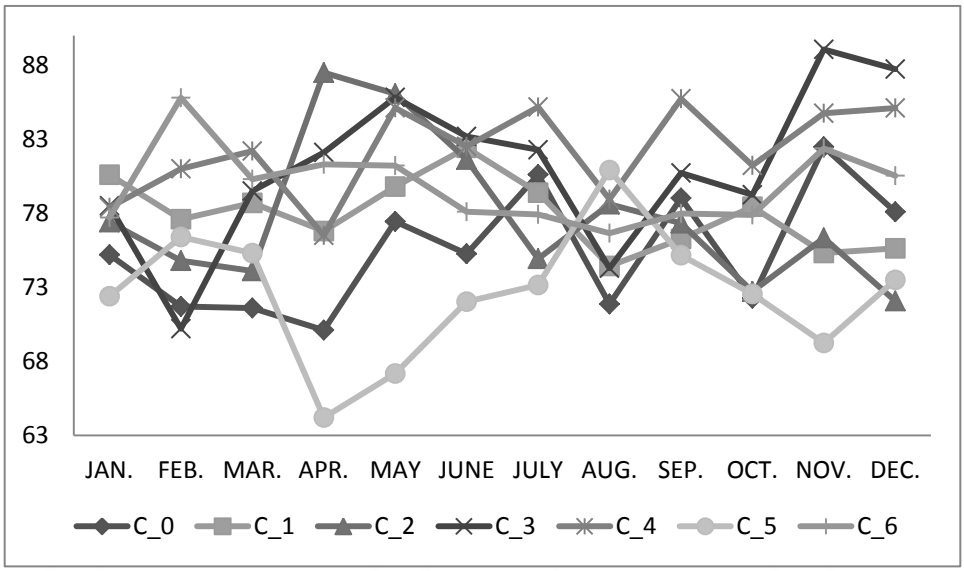


Figure 3.11 Cluster centroids for 2012

3.3.3.4 Average of 2010, 2011 and 2012 Scores

- The satisfaction level of the cities in the cluster 0 is remarkably high for all the months.
- The cities in the cluster 5 (9 cities) has a negative peak value in April and shows no consistency along the year.
- The customer satisfaction level of the cluster 5 is remarkably low.
- These results are shown in Table 3.29 and Figure 3.12.

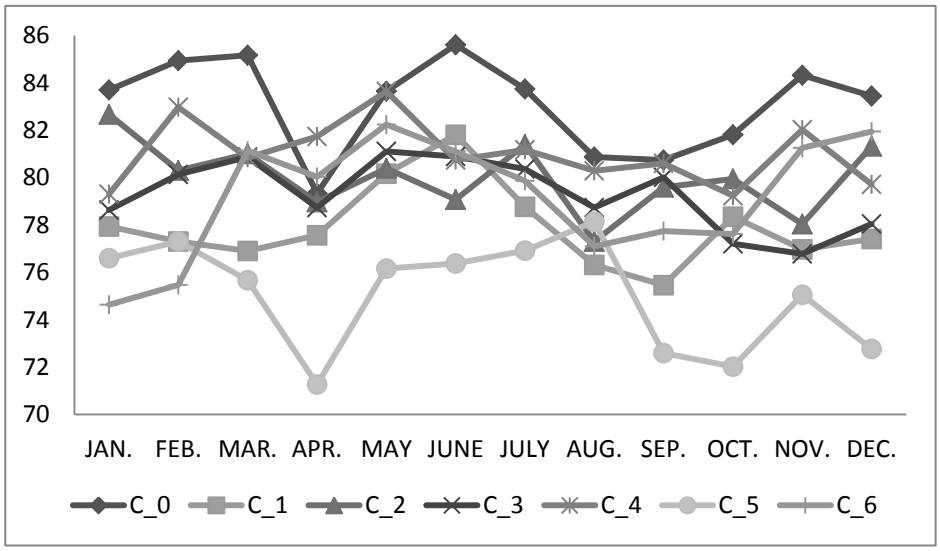


Figure 3.12 Cluster centroids for average scores

Table 3.29 The centroid table for average CSPS

MONTH	C_0	C_1	C_2	C_3	C_4	C_5	C_6
JANUARY	83.7	77.9	82.7	78.6	79.3	76.6	74.6
FEBRURARY	84.9	77.3	80.3	80.1	83.0	77.3	75.5
MARCH	85.2	76.9	81.0	80.9	80.8	75.7	81.1
APRIL	79.3	77.6	79.0	78.7	81.7	71.3	80.0
MAY	83.6	80.2	80.4	81.1	83.6	76.2	82.2
JUNE	85.6	81.8	79.1	80.9	80.7	76.4	81.1
JULY	83.7	78.8	81.4	80.4	81.2	76.9	79.9
AUGUST	80.9	76.3	77.3	78.7	80.3	78.1	77.1
SEPTEMBER	80.7	75.5	79.6	80.0	80.6	72.6	77.7
OCTOBER	81.8	78.3	79.9	77.2	79.2	72.0	77.6
NOVEMBER	84.3	77.0	78.0	76.8	82.0	75.0	81.3
DECEMBER	83.4	77.4	81.3	78.0	79.7	72.8	81.9

3.3.4 Comparison of K-means, X-Means and K-medoids Results

In Table 3.30 distance measurements of the clusters for all algorithms of the year 2010 CSPS and in the Table 3.31 density distributions are given. As known, evaluation of unsupervised methods is hard. Best way is to see if the results prove to be useful in practice. However, interpreting the values in the tables could give an idea together with the actual data and dispersion of the cities in the clusters. For example, average distance in cluster 6 for k-medoids algorithm is zero. This means that there is a just one city in that cluster. In another way, it could be said that, average within centroid distance in cluster 5 for the algorithm k-medoids is 92.875 meaning that the similarity in that cluster is lowest among other clusters.

Table 3.30 The Centroid distance table for 2010's CSPS

	K-Means	X-Means	K-Medoids
Avg. within centroid distance	-32.199	-31.579	-46.077
Avg. within centroid distance_cluster_0	-36.427	-28.229	-33.932
Avg. within centroid distance_cluster_1	-31.45	-30.334	-45.77
Avg. within centroid distance_cluster_2	-23.708	-23.846	-35.472
Avg. within centroid distance_cluster_3	-29.001	-14.774	-40.38
Avg. within centroid distance_cluster_4	-69.527	-22.794	-39.287
Avg. within centroid distance_cluster_5	-34.733	-64.626	-92.875
Avg. within centroid distance_cluster_6	-18.006	-33.365	0
Davies Bouldin	-1.838	-1.596	-1.359

The cluster efficiency measurements algorithms try to find out the similarity in the cluster while determining the optimum k value (Davies & Bouldin, 1979). In the study, to find out the relation between customer satisfaction scores and the geographical dispersion of the cities, k value was determined as 7, since there are 7 geographical districts in Turkey.

Table 3.31 The density table (k-means, x-means and k-medoids respectively)

Cluster	2010	2011	2012	Average
0	8, 5, 20	8, 3, 2	21, 11, 6	17, 16, 4
1	7, 9, 15	7, 6, 3	3, 5, 34	8, 20, 19
2	7, 17, 9	14, 24, 17	10, 8, 1	10, 9, 7
3	20, 4, 22	20, 8, 7	25, 4, 1	16, 2, 23
4	11, 28, 6	4, 16, 7	12, 23, 19	1, 15, 17
5	21, 11, 3	20, 6, 29	2, 12, 6	26, 9, 9
6	6, 6, 5	7, 17, 18	7, 17, 13	2, 9, 1

So, to evaluate the results in this study efficiently, the centroid tables, the distance measurement table and the density distribution table have to be taken into consideration together with the actual data itself. The future work would be to find out the optimal k value and then compare the efficiencies of the clustering algorithms. Another future work would be to add a special attribute called label to the CSPS's data such as "is the customer satisfied?" to be able to apply supervised methods. The cluster density table is given in Table 3.31.

3.3.5 Results for Decision Support System

How the community structure such as economy, education, security, development would affect the customer expectations. Are there any relation between the social issues and customer satisfaction while the service quality is remaining the same? Are there any relations among the cities dispersed around Turkey in terms of customer satisfaction scores? It has been shown the following results by using the Spearman's coefficient correlation rank using SPSS with the p-value<0.05:

- The more the habitableness order is, the less the customer satisfaction score (CSS) is.
- The more the economic development order is, the less the CSS is.
- The better the educational status of the cities is, the less CSS is.
- There is no meaningful relation between the health developments of the cities and the CSS.
- There is meaningful positive relation between the cultural-art developments of the cities and the CSS.
- There is no meaningful relation between the security developments of the cities and the CSS.
- There is no meaningful relation between the city life and the CSS

As can be seen from the results, examination of habitableness order, economic development order and educational status are consistent while there are no consistent results among others.

CHAPTER FOUR
CLUSTERING THE CITIES USING DEMOGRAPHIC ISSUES IN TURKEY

In this chapter, by combining the CNBC-E magazine research (Mavi, 2011) and the customer satisfaction performance scores (CSPS) for the 81 cities of Turkey, the Agglomerative hierarchical clustering algorithm is applied to investigate the cities in Turkey in terms of some demographic issues and customer satisfaction. In the CNBC-E's habitableness research, 34 topics of different demographic issues such as air utilization rate, unemployment rate, number of pre-school student per classroom etc. are taken into consideration in order to sort the cities in terms of these topics. In addition to the CNBC-E research results, the CSPS of a fixed line telecom operator in Turkey is added to ranking titles. All the data are conducted in the year of 2010. By using the CSPS apart from the CNBC-E data, how the customer satisfaction ranks affect the clustering is investigated as well.

By clustering the cities depending on the CNBC-E research and the CSPS, a landscape of Turkey will be acquired. To cluster cities, the Agglomerative hierarchical clustering algorithm and the k-medoids 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 data to the map. As known, the agglomerative clustering methods have three modes such as single link, complete link and average link. After applying these modes, the efficiencies of the modes are calculated by using cophenetic correlation coefficient. Due to its efficiency, the average link mode is used to map the data and interpret the cluster results. As the measure types, the mixed measures and as the mixed measure parameter, the mixed Euclidian has been chosen.

After applying the algorithms, the accuracy and item distribution performance are calculated as well as optimizing the cluster number of the k-medoids. It can be seen from the results that, there are explicit relations among the cities spread in the same geographical areas with some exemptions. By using the maps for several levels, the status for the cities can be investigated by bottom-up manner. In this respect, the study can be considered a mixture of a geographic information system and a data mining system to present a decision support system. First, the data was clustered and after clustering process, the results were entered to a geographical information system (i.e. MapInfo) to visualize the results.

4.1 Applying the Agglomerative and K-medoids Algorithm to Data

The annual average of the 2010 monthly customer satisfaction scores (CSPS) of the fixed line operator for 81 cities in Turkey and the habitable cities evaluation rank in Turkey research of the CNBC-E Business Magazine in September 2011 for 2010 (Mavi, 2011) are used for the inputs of data mining algorithms.

The 35 attributes for the cities in terms of order of cities are used in the clustering algorithms (Some are listed in Table 4.1). 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 number is 48 for Unemployment Rate.

For better understanding and simplicity, the city names are numbered as showed in Table A.1 in appendix A. Using the numbers rather than names provides a clear view on the maps. Consequently, all of the ranking titles will be given in Appendix B as well.

Table 4.1 Sample view of the data used in the study

Demographic Info/City Name	Ankara (1)	Eskisehir(2)	Isparta (3)	Trabzon (4)	Istanbul (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 Universıty Graduates	1	2	5	7	4
Literacy Rate	4	3	15	40	20

4.2 The Main Process Design

In this section, the process design of Rapidminer will be given. The steps of process will be given in the following (Figure 4.1);

- Multiply the data to get the inputs of the single link.
- Agglomerative algorithm, complete link Agglomerative algorithm, average link Agglomerative algorithm and 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 Cophenetic Correlation Coefficient to calculate the performances of the three Agglomerative clustering algorithms.
- The output of the data to similarity operator is also multiplied to calculate each of the Agglomerative clustering algorithm (Single Link, Complete Link and Average Link).

- 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 be able to compare with the results of the Agglomerative algorithm.

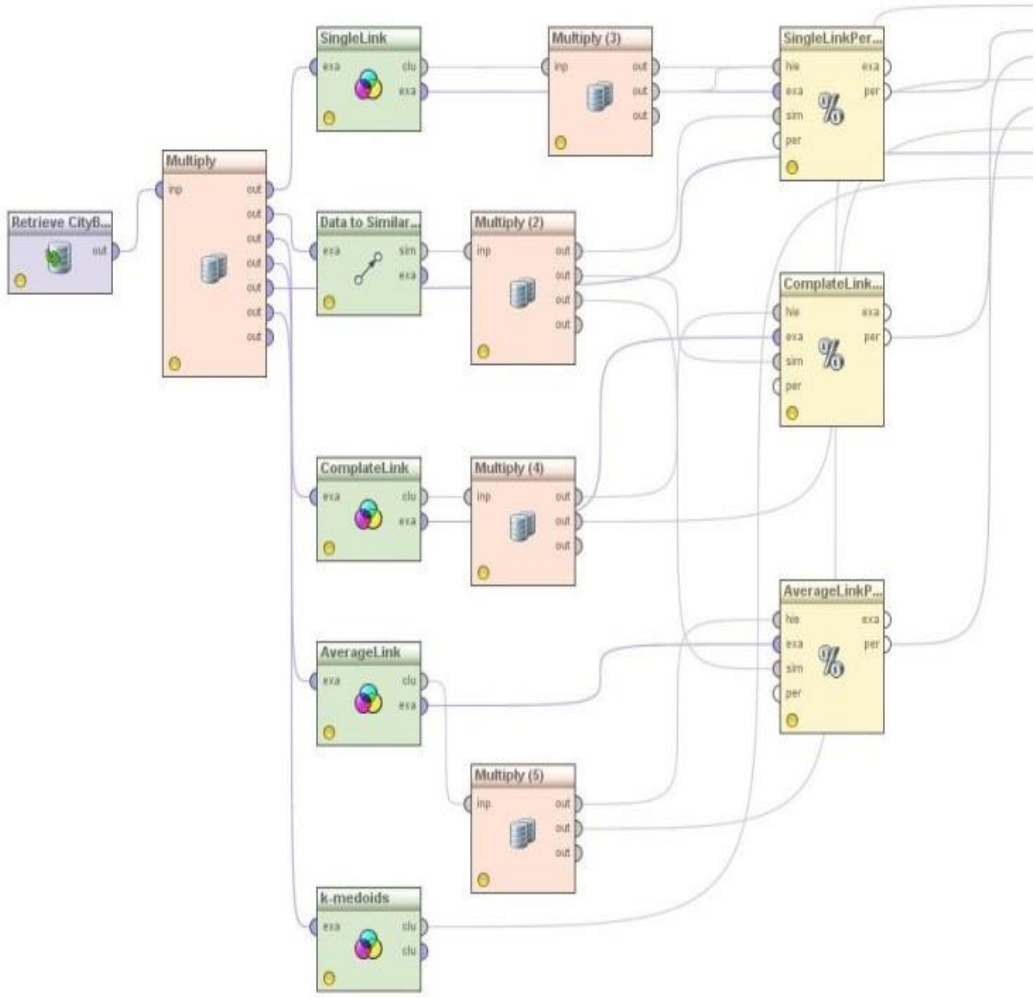


Figure 4.1 Main process diagram.

The operators used in this diagram are explained briefly as follows;

- Retrieve CityBasedList: This operator reads the data from the source.
- Multiply: This operator copies its inputs object to all connected output ports without modifying the input object.
- Single Link: This operator performs the single link Agglomerative algorithm

to data.

- Complete Link: This operator performs the complete link Agglomerative algorithm to data.
- Average Link: This operator performs the average link agglomerative algorithm to data.
- K-medoids: This operator performs the k-medoids algorithm to data.
- Data to Similarity: This operator measures the similarity of each example of the given Example set with every other example of the same Example set.
- Cophenetic Correlation Coefficient Operator (CPCC): In this example, this operator computer the CPCC for single link, complete link and average link outputs of data to similarity operator.

4.3 The Results of Agglomerative and K-medoids Algorithms

In this section, the results of the single, complete and average link of Agglomerative algorithm and the results of k-medoids algorithm will be 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 will be used. In Figure 4.2, the tree of the average link mode agglomerative algorithm is depicted.

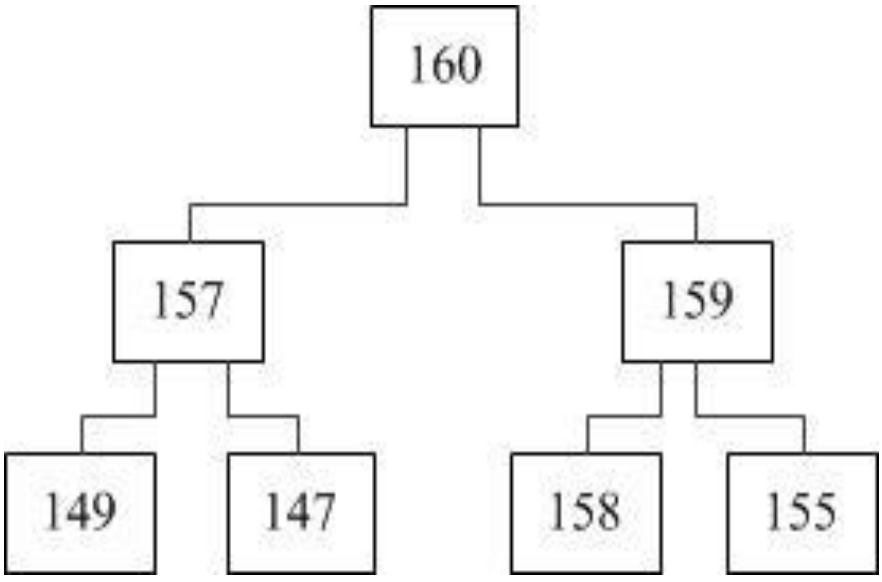


Figure 4.2 Tree view of average link agglomerative clustering algorithm.

In the k-medoids algorithm, because of parameter optimization the cluster number k is chosen as 4. The detailed explanation for this process will be given in the section 4.5.

In Figure 4.2, 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 is depicted in Figure 4.3 as well.

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, city 22, 30, 15 and 54 are participated to the same cluster (Cluster number 104).

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

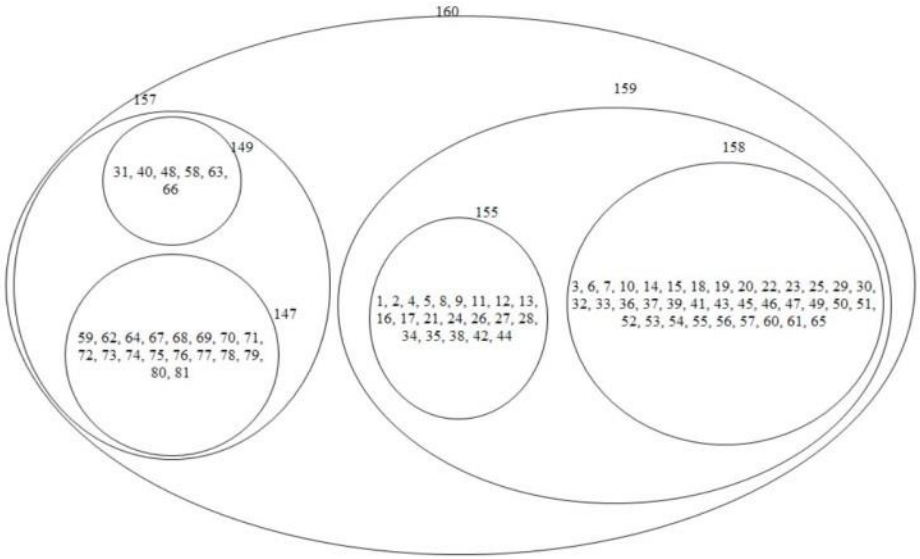


Figure 4.3 Nested clusters for k=4.

In Figure 4.2, the clusters for k=4 are depicted for the Agglomerative clustering. The cluster branches are also given in Figures 4.4, 4.5, 4.6 and 4.7 in detail respectively. The tree views of the clusters provide an opportunity of viewing data distribution as well.

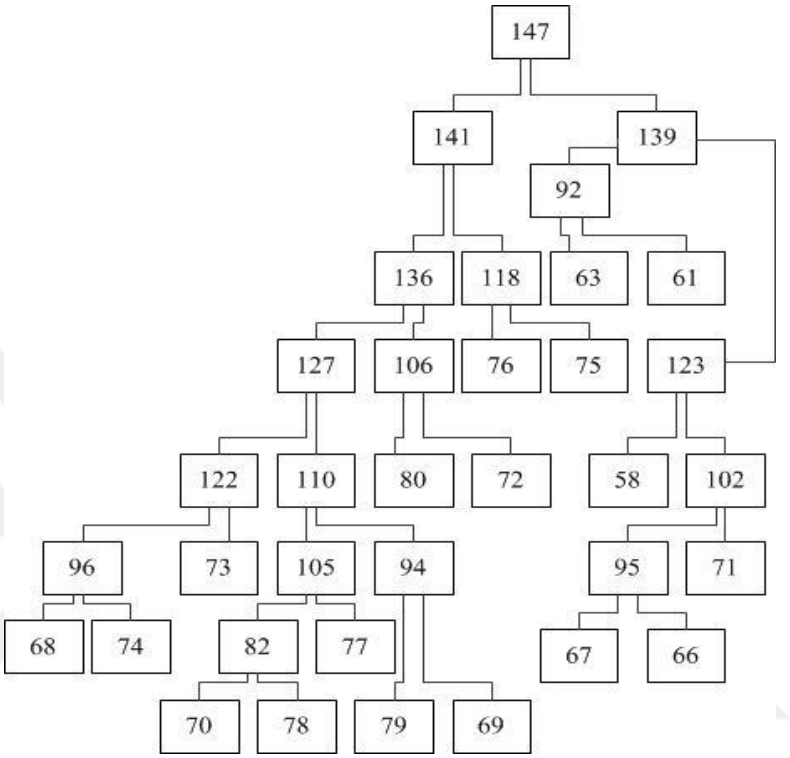


Figure 4.4 Cluster 147 (Tree depth is 2).

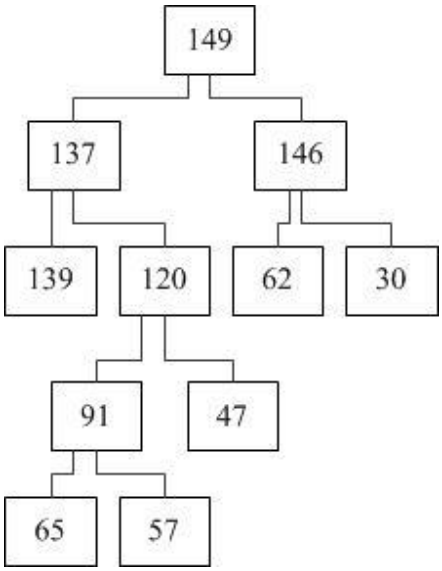


Figure 4.5 Cluster 149 (Tree depth is 2).

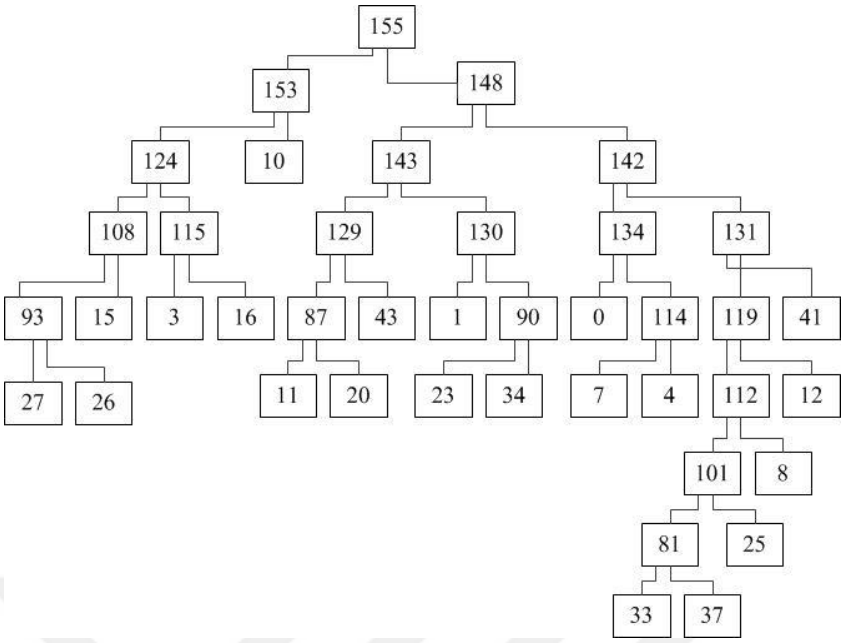


Figure 4.6 Cluster 155 (Tree depth is 2).

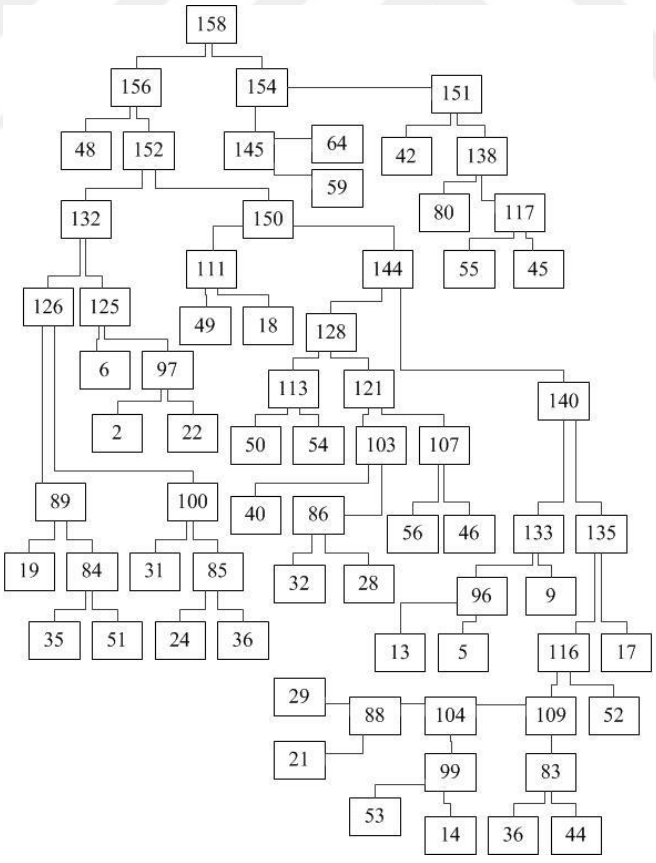


Figure 4.7 Cluster 158 (Tree depth is 2).

In Figure 4.8, the k-medoids algorithm clusters are given to compare with the agglomerative clustering results. These results will also be compared in Section 4.4 by using the map views.

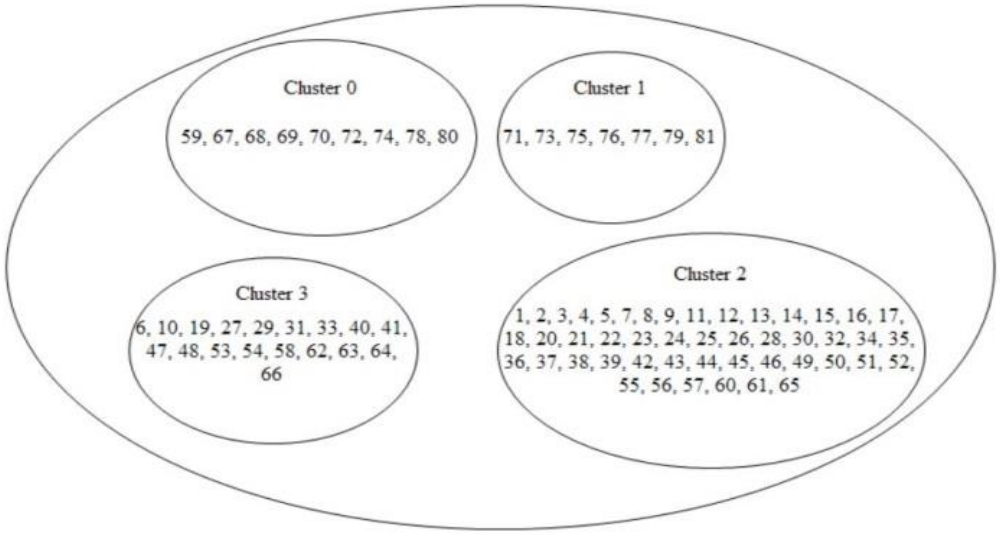


Figure 4.8 K-Medoids Clusters for k=4

4.4 The Map Views of the Clustering Results

In this section, the results of the Agglomerative clustering and k-medoids clustering results are mapped to visualize the data. The maps are used to investigate the data in terms of geographical issues. The maps of Agglomerative clustering for k 8, 16, 27, and 40 are given in Figures C.2, C.3, C.4 and C.5 respectively in Appendix C. The cluster number is selected as 4 for mapping.

As can be seen the results depicted in the Figure 4.2, when we would take the depth of the result tree as 2, there would be four clusters. The distribution of the cities is depicted in Figure 4.9. When cluster number k is chosen as 4, the cities in the same geographical areas are clustering into same cluster with some exemptions. In other words, the people living in the same areas are behaving in the same manner in terms of our 35 parameters.

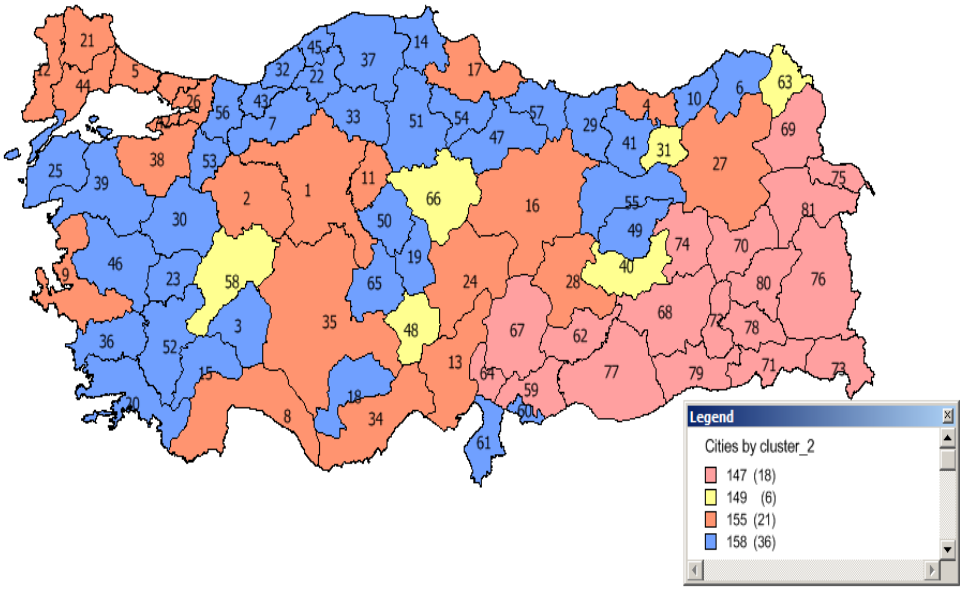


Figure 4.9 City Distribution for k=4 (Agglomerative Clustering).

Again, cities 58 (Afonkarahisar), 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. In other words, the cities in cluster 149 exhibit a different behavior compared to others. This situation can be seen from Figure 4.10 (in k-medoids) as well.

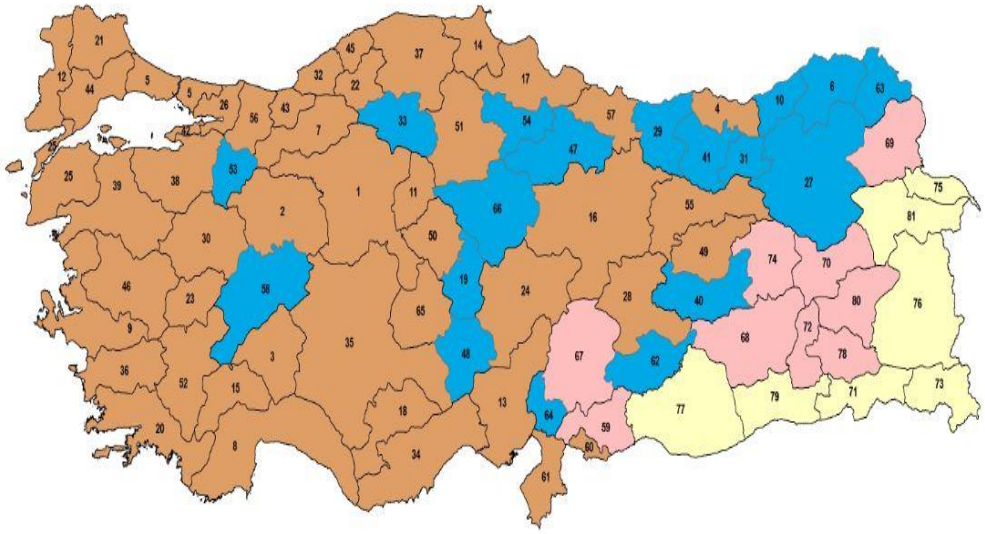


Figure 4.10 K-medoids results for the data (i.e., CNBC-E Data and CSPS).

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 a data organized as sorting of the city scores. To get the better performance, in this case, the actual points of cities are used rather than sorting of scores. The comparisons of the results depicted in Figures 4.11 and 4.12 could be done.

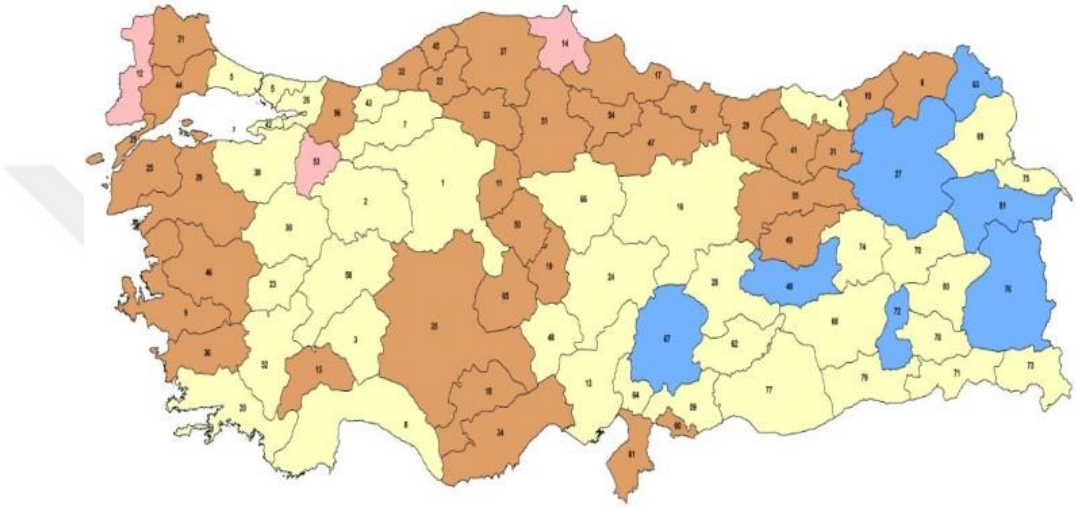


Figure 4.11 Agglomerative Algorithm Results for CSPS.

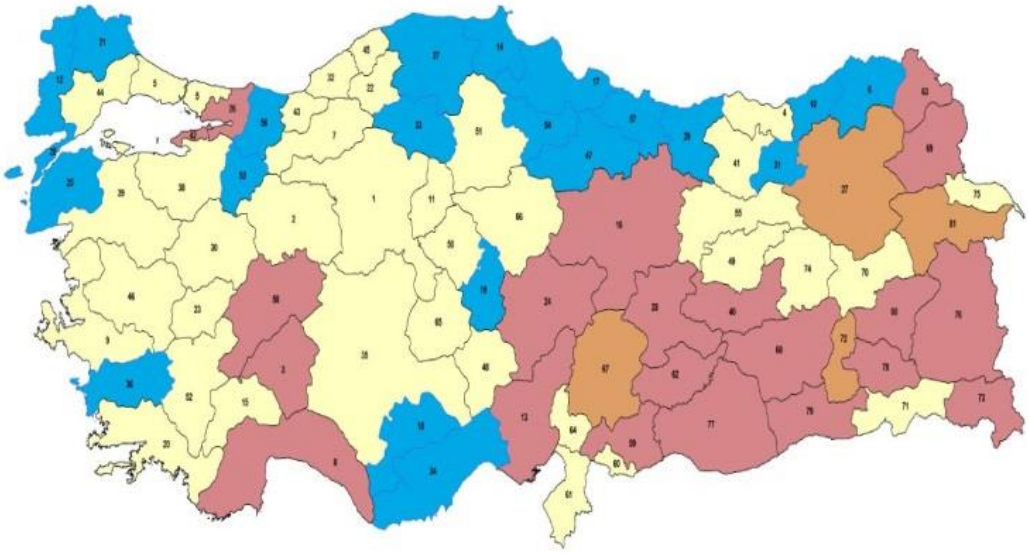


Figure 4.12 K-Medoids results for CSPS.

By comparing agglomerative clustering results depicted in Figure 7.11 and k-medoids results depicted in Figure 7.12, it can be said that there is no explicit effect of customer satisfaction to clusters.

4.5 Cluster Performance and Accuracy Analysis

In the cluster performance and accuracy analysis, the main process depicted in Figure 4.13 is used. With the set role operator, the cluster names (i.e. cluster 0, cluster 1, etc.) are assigned to the class label. In this way, which city that has certain social parameters belongs to which cluster is investigated.

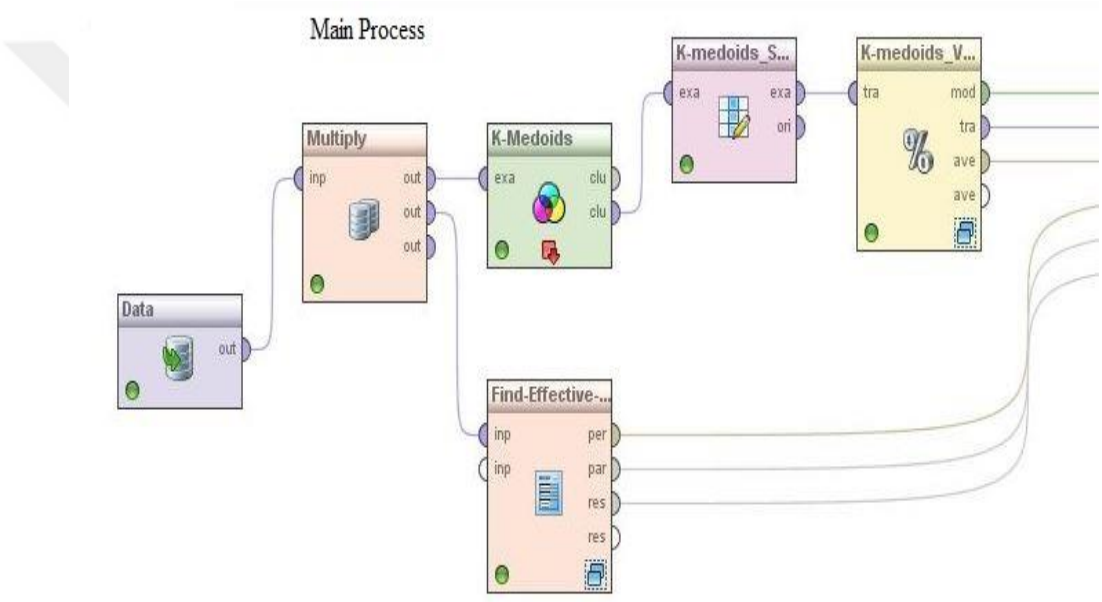


Figure 4.13 Main accuracy and performance measurement process.

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 correlation performance parameters are selected in Rapidminer. The sub-process schema of the X-validation process is given in Figure 4.14. The decision tree is used since there are too many parameters in the data table such as 81 cities and 35 social items.

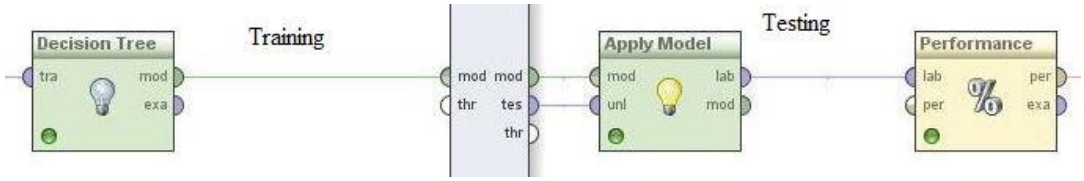


Figure 4.14 Validation sub processes.

In Figure 4.15, the optimization sub-process is depicted. This process is used to find the most effective k value. The results are given in Table 4.2. As can be seen from results in Table 4.2, the item distribution performances for all k values are higher than 90%.

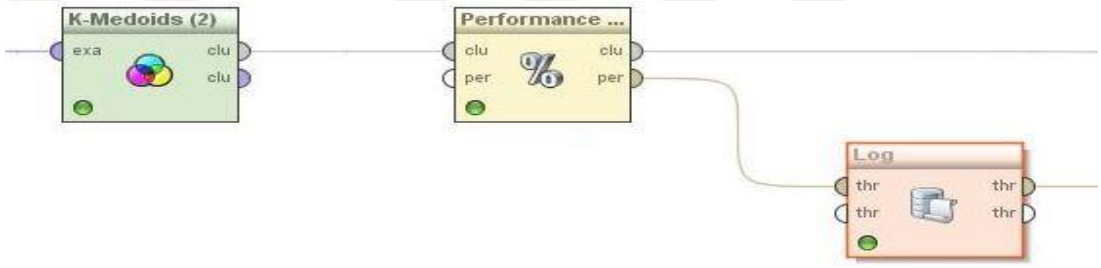


Figure 4.15 Optimization sub process.

Table 4.2 Item distribution performance for k value optimization

Number of Clusters	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%. As a result 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 86.55%. To make difference among cities, the optimal k value is selected as 4 by checking the accuracy levels and class recall values (Table 4.3). Additionally, in the agglomerative clustering section, the optimal tree level was selected as 2 for tree depth. This also means that there were four clusters in investigating the cities in terms of geographic distribution.

Table 4.3 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 were found as household per capita consumption of electricity, forest area ratio, rate of primary school students per teacher and rate of traffic accent per vehicle.

As can be seen from the distinctive parameters depicted in the Figure 4.16, the first distinctive item for clustering is the household per capita consumption of electricity value. The city having households per capita consumption of electricity higher than 78.50 is assigned to cluster 0. On the other hand, for other 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 belongs to cluster 0 again, otherwise the rate of primary school students per teacher value will be considered for the next distinctive parameter.

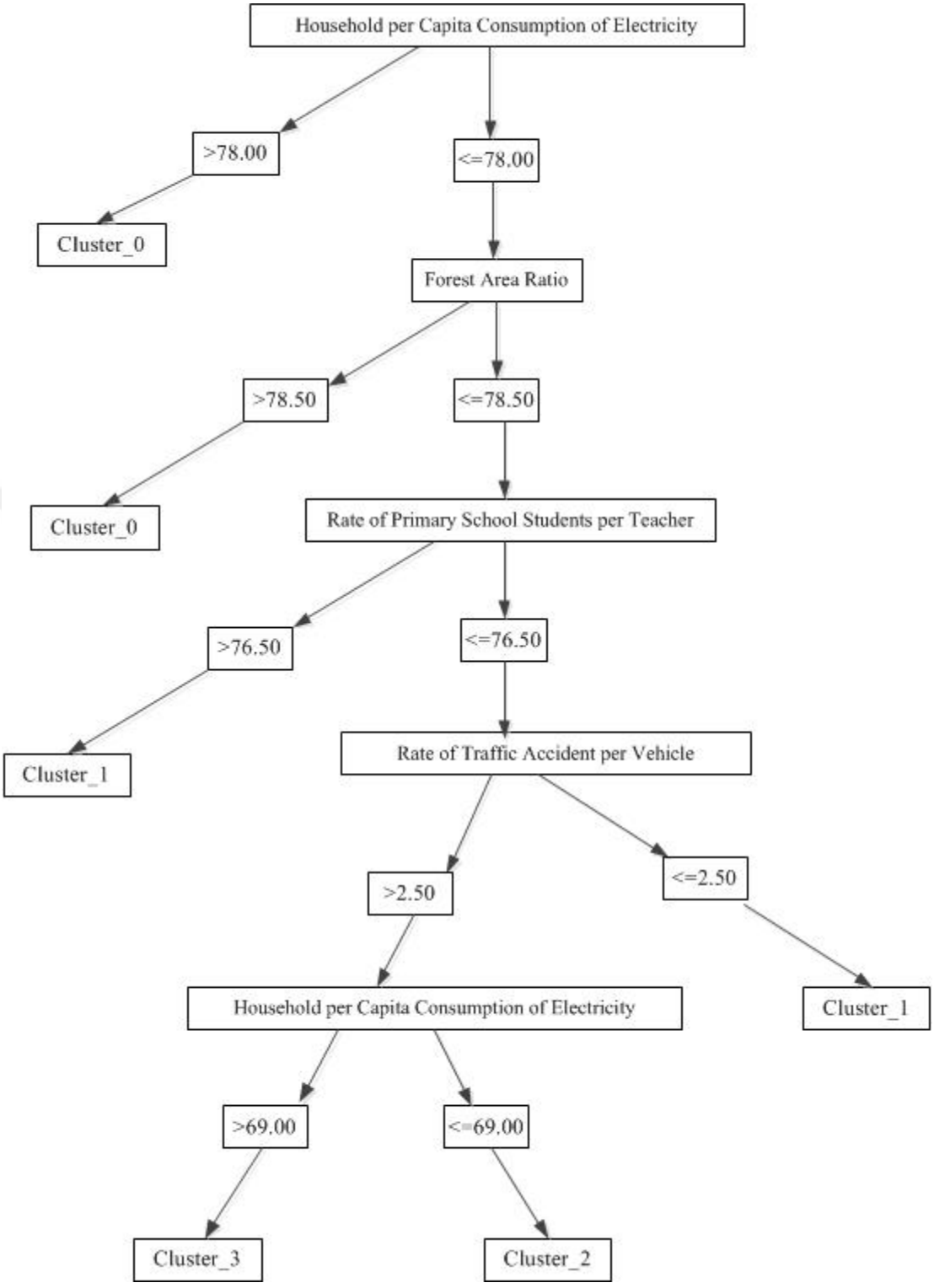


Figure 4.16 Distinctive parameters for clustering (for k=4).

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 traffic accident per vehicle is considered. As can be seen from Figure 4.8, 9 cities are assigned to cluster 0 in this way. By using Figure 4.8, other items of the clusters can be find in the same way.

If the rate of traffic accident per vehicle value is higher than 2.50, the household per capita consumption of electricity value is considered again. If this value is higher than 69.00, the city is assigned to the cluster 3 and if the value is less than or equal to 69.00 then the city is assigned to the 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, since the effect of the customer satisfaction to the clustering, it can be seen that the CSPS value is not a distinctive parameter.

Since, the optimum k value is computed as 4, there are four distinctive parameters. It is obvious that, the distinctive parameters will change when changing the cluster number k.

CHAPTER FIVE
CONCLUSION

5.1 Conclusion

In this study, the customer satisfaction has been handled from a different point of view as a first part of the thesis. How the community structure such as economy, education, security, development would affect the customer expectations? Are there any relation between the social issues and customer satisfaction while the service quality is remaining the same?

It has been shown the following results by using the Spearman's coefficient correlation rank using SPSS for the significance level at 0.05:

- The more the habitableness order is, the less the customer satisfaction score is.
- The more the economic development order is, the less the customer satisfaction score is.
- The better the educational status of the city is, the less the customer satisfaction score is.
- There is no meaningful relation between the health developments of the city and the customer satisfaction scores.
- There is meaningful positive relation between the cultural-art developments of the cities and the customer satisfaction scores.
- There is no meaningful relation between the security developments of the cities and the customer satisfaction scores.
- There is no meaningful relation between the city life and the customer satisfaction scores.

As can be seen from the results, examination of habitableness order, economic development order and educational status are consistent while there are no consistent results among others. By examining these results, a customer experience program can be conducted for every cities separated from each other in terms of habitableness levels. For example, a new questionnaire can be designed for the cities that have high

habitableness ranks. On the other hand, while examining customer experience programs, cities can be grouped by checking the ranks and every city is evaluated by the group to which it belongs.

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. To understand the customer behavior is very important to deliver the services needed by the customers. In this article, a decision support consisting of GIS and data mining techniques is proposed as a case study for Turkey to depict a picture showing the customer behavior in terms of the demographic issue of the cities in Turkey. It has also been showed that the people living the same geographic areas are behaving in similar manners. When designing a customer experience program, the results of the article can be used as a decision support for managers.

As seen in the topic 4.5, household per capita consumption of electricity, forest area ratio, rate of primary school students per teacher and rate of traffic accent per vehicle parameters are revealed as distinctive parameters in structuring clusters. Depending on the clustering using these parameters, the result will be given in the following;

- 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. When cluster number k= 8, it can be seen that there are just these two cities fit in cluster 146.
- 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. As can be seen from the data, the household consumption of electricity ranks for these cities are 41, 61, 64,

56, 62 and 66 respectively and all these ranks are smaller than 69. By using, the distinctive parameters tree, it can be seen that these cities are located in the cluster 2.

- The cophenetic correlation coefficient for single link Agglomerative clustering is 0.353, for complete link agglomerative clustering is 0.589 and for average link clustering algorithm is 0.644. So, the average link agglomerative clustering algorithm was used to examine the results.
- The optimal cluster number k for the k-medoids algorithm is calculated as four.

These results can beconsidered a data driven decision support system as explained in the section 2. After clustering the cities, by examining the map results, a new sales campaign can be conducted for every cluster. Consequently, a customer experience programs can be designed in different way for every cluster.

5.2 Future Works

A future work would be to find a correction index for the customer satisfaction scores. As shown in this study, while the services and products are the same, the customer satisfaction levels can be varying in according to community structure. Therefore, it could be found a correction index to get the actual results of the customer satisfaction scores and find the most proper order of the cities. This leads the better actions to be taken in the right offices to service better service to customers.

In the second part of the thesis, how the cities are clustered in terms of the habitableness sub-topics has examined. As a future work of this part, after founding distinctive parameters, by using these parameters merely, the agglomerative algorithm could be applied again to the same data and the new values of accuracy and efficiency parameters could be calculated and compared to the previous ones to observe how the class efficiency parameters being affected.

Finally, the ranked data is used in mining operations in the study. This is why the Agglomerative clustering algorithmis chosen. If the real data is used, the other clustering algorithms can be used to find distinctive parameters and to get different efficiency levels. By comparing the results of these algorithms, more efficient decision support could be designed.

REFERENCES

Anselin, L., & Getis, A. (2010). Spatial Statistical analysis and geographic information systems in Anselin L.,and Rey S. J. (Ed). *Perspectives on Spatial Data Analysis,* Berlin: Springer

Averweg, U., & Richard, F. (2012). *Decision-making support systems: Theory & Practice*.

Azad, N., & Hashemi S. (2013).A Study of important factors influencing customer relationship management: A Case study of mobile service provider. *Management Science Letters, 3*(4), 1161-1166.

Bayraktar, E., Tatoglu, E., Turkyılmaz, A., Delen D., & Zaim S. (2012). Measuring the efficiency of customer satisfaction and loyalty for mobile phone brands with DEA. *Expert Systems with Applications, 3*(39), 99-106.

Bhargava, H. K., Power, D. J., & Sun, D. (2007). Progress in Web-based decision support technologies. *Decision Support Systems, 43*(4), 1083-1095.

Chiang, W. Y. (2011). To mine association rules of customer values via a data mining procedure with improved model: An empirical case study. *Expert Systems with Applications, 38* (3), 1716-1722.

Cranshaw, J., Schwartz R., Hong J. I., & Sadeh N. M. (2012). The live hoods project: Utilizing social media to understand the dynamics of a city. *Proceeding 6th International AAAI Conference on Weblogs and Social Media*, 58-65.

Codd, E. F., Codd, S. B., & Salley, C. T. (1993). Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate. *Codd and Date*, *32*.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure.Pattern analysis and machine intelligence. *IEEE Transactions*, (2), 224-227.

Deng, Z., Lu, Y., Wei, K.K., & Zhang J. (2010). Understanding customer satisfaction and loyalty: An empirical study of mobile instant messages in Chine. *International Journal of Information Management, 30* (4), 289-300.

Dvoroznak, M. (2104). *Clustering performance plug-in,* Retrieved May 20, 2014, from http://korek.name/web/moje-tvorba/rapidminer-clustering_performance_plugin-average_silhouette-cophenetic_coefficient.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*, London: Sage

Frank, B., & Enkawa, T. (2008). How economic growth affects customer satisfaction. *Asia Pacific Management Review, 12* (2), 531-544.

Gachet, A. (2000). A framework for developing distributed cooperative decision support systems: inception phase (working paper). *Department of Informatics-University of Fribourg.* Retrieved July 20, 2015, from https://www.researchgate.net/publication/243604133_A_Framework_for_Developing_Distributed_Cooperative_Decision_Support_Systems_-_Inception_Phase

Garla, S., Chakraborty, G., Gaeth U. G., & Iowa U. S. (2011). Comparison of k-means, mormal mixtures and probabilistic-d clustering for B2B segmentation using customers' perceptions. *Conference Proceedings of the SAS Global Forum*, 1-8.

Gautheir, T. D. (2001). Detecting trends using spearman's rank correlation coefficient. *Environmental Forensics, 2*(4), 359-362.

Gorsevski, P. V., Donevska, K. R., Mitrovski, C. D., & Frizado, J. P. (2012). Integrating multi-criteria evaluation techniques with geographic information systems for landfill site selection: a case study using ordered weighted average. *Waste Management, 32*(2), 287-296.

Guha, S., Rastogi, R., & Shim, K. (1999). ROCK: A robust clustering algorithm for categorical attributes in data engineering. *Proceedings 15th International Conference on IEEE*, 512-521.

Guha, S., Rastogi, R., & Shim, K. (1998). CURE: An efficient clustering algorithm for large databases. *ACM SIGMOD Record, 27*(2), 73-84

Gupta, A., Jha, C.K., & Shukla A.K. (2012). Factors influencing M-commerce transaction satisfaction and impact of trust towards M-commerce service providers. *Journal of Computer Engineering, 3*(5), 01-07.

Hamad, M. M., & Qader, B. A. (2014). Knowledge-driven decision support system based on knowledge warehouse and data mining for market management. *Global Journal of Management and Business Research*, *13*(10).

Hamilton, H. (2012). Retrieved May 20, 2014, from *http://www2.cs.uregina.ca/~dbd/ cs831/notes/clustering/clustering.html*

Han, J., & Kamber, M. (2006). *Data mining: concepts and techniques*, San Francisco: Morgan Kaufmann.

Hossain M., & Leo, S. (2009). 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, 2*(4), 338-350.

Kantardzic, M. (2011). *Data Mining: Concepts, models, methods and algorithms*, New Jersey: John Wiley & Sons, Inc.

Karypis, G., Han, E. H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer, 32*(8), *68-75.*

Kaur, J., & Gupta, G. (2014). Optimized clustering algorithm with hybrid k-means and hierarchical algorithms. *International Journal for Multidisciplinary Engineeringand Business Management, 2*(1).

Kazi, A.K. (2013). An empirical study of factors influencing adoption of Internet banking among students of higher education. *International Journal of Finance & Banking Studies*, *2*(2), 87-99.

Kendall, M.G. (1962). *Rank correlation methods*, New York: Griffin.

Khan, K., Baharudin, B., & Khan, A. (2014). Identifying product features from customer reviews using hybrid dependency patterns. *The International Arab Journal of Information Technology*, *11*(3), 281-286.

Khokhar, S.Z., Hussain, F., Qureshi, T.M., Anjum, I., Samran, A., & Arshad R., (2011). Only customer satisfaction and customer loyalty is not enough: a study of Pakistan's telecom sector. *African Journal of Business Management, 5*(24), 10176-10181.

Kumar, P., & Wasan, S. K. (2010). Comparative analysis of k-mean based algorithms. *International Journal of Computer Sciences and Network Security*, *10*(4), 314-318.

Kuo, Y. F., Wu, C. M., & Deng, W. J. (2009). The relationships among service quality, perceived value, customer satisfaction, and post-purchase intention in mobile value-added services. *Computers in Human Behavior*, *25*(4), 887-896.

Kuronen, M., & Takala J. (2013). Measuring and developing customer satisfaction by balanced critical factor index. *Management and Production Engineering Review*, *4*(3), 34-39.

Land, S., & Fischer S. (2012). *RipedMiner 5*. Retrieved May 2014 from http://docs.rapidminer.com/downloads/RapidMiner_RapidMinerInAcademicUse_en.pdf

Manani, T.O., Nyaoga, R.B., Bosire, R.M., Ombati, T.O., & Kongere, T.O. (2013). Service quality and customer satisfaction at Kenya airways ltd. *European Journal of Business Management*, *5*(22), 170-179.

Mavi, B. (2011). Research on livable cities in Turkey. *CNBC-E Business Magazine*, 9

Mavi, B. (2012). Research on livable cities in Turkey. *CNBC-E Business Magazine*, 11

Muijs, D. (2004). *Doing quantitative research in education with SPSS*. London: Sage

Musso, J. A. (2001). The political economy of city formation in California: Limits to tiebout sorting. *Social Science Quarterly, 82*(1), 139-153.

Pamuk, A. (2004). Geography of immigrant clusters in global cities: a case study of San Francisco. *International Journal of Urban and Regional Research, 28*(2), 287-307.

Pelleg, D., & Moore, A. W. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. *In ICML*, 727-734.

Power, D. J. (2001). Supporting decision-makers: An expanded framework. *In e-Proceedings Informing Science Conference,* Krakow, Poland, 431-436

Power, D. J. (2004). Specifying an expanded framework for classifying and describingdecision support systems. *The Communications of the Association for Information Systems, 13*(1), 52.

Power, D. J., & Sharda, R. (2007). Model-driven decision support systems: Concepts and research directions. *Decision Support Systems, 43*(3), 1044-1061.

Power, D. (2007). What are the features of a communications-driven DSS? *DSS News, 8* (2), 15-20.

Power, D. J. (2008). Understanding data-driven decision support systems. *Information Systems Management, 25*(2), 149-154.

Power, D. J. (2008). Decision support systems: a historical overview. *In Handbook on Decision Support Systems 1*. Berlin: Springer

Rahman, S., Haque A., & Ahmad, M.I.S. (2010). Exploring influencing factors for the selection of mobile phone service providers: a structural equational modeling (sem) appraoch on Malaysian consumers. *African Journal of Business Management, 4*(13), 2885-2898.

Ramasubbu, N., Mithas, S., & Krishnan, M. S. (2008). High tech, high touch: The effect of employee skills and customer heterogeneity on customer satisfaction with enterprise system support services. *Decision Support Systems*, *44*(2), 509-523.

Raza, M.A., Siddique, A.N., Awan, H.M., & Bukhari, K. (2012). Relationship between service quality, perceived value, satisfaction and revisit intension in hotel

industry. *Interdisciplinary Journal of Contemporary Research in Business, 4*(8) 788-806.

Schwarz, N. (2010). Urban form revisited—Selecting indicators for characterizing European cities. *Landscape and Urban Planning, 96*(1*)*, 29-47.

Siddiqi, K.O. (2011). Interrelations between service quality attributes, customer satisfaction and customer loyalty in the retail banking in Bangladesh. *International Journal of Business and Management, 6*(3), 12-36.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72–101.

Suki, N.M. (2011). A structural model of customer satisfaction and trust in vendors involved in mobile commerce. *International Journal of Business Science and Applied Management, 6(*2), 17-30.

Suresh, S. S., Naidu, M. M., & Asha Kiran, S. (2012). An XML based knowledge-driven decision support system for design patterns election. *International Journal of Research in Engineering and Technology, 11*(3), 116-124.

Tabaei, Z., Fathian M., & Gholamian M.Z. (2011). Effective factors on electronic customer satisfaction. *3rd International Conference on Information and Financial Engineering*, 12, Singapore: IACSIT Press

Teknomo, K. (2012). *Hierarchical clustering tutorial*. Retrieved July 20, 2014, from http://people.revoledu.com/kardi/tutorial/Clustering/index.html

Wu, R. S., & Chou, P. H. (2011). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach, *Electronic Commerce Research and Applications, 10*(3), 331-341.

Williams, P., & Neumann, E. (2011). Customer satisfaction and business performance: a firm-level analysis. *Journal of Services Marketing, 25*(1*)*, 20-32.

Wyld, D. C., Wozniak, M., Chaki, N., Meghanathan, N., & Nagamalai, D. (2011). Advances in computing and information technology: First international conference. *ACITY 2011*, 98, India: Springer Science & Business Media.

Yi, Y., & Gong, T. (2008). The effects of customer justice perception and effect on customer citizenship behavior and customer dysfunctional behavior. *Industrial Marketing Management, 37*(7), 767-783.

Zar, J. H. (1972). Significance testing of the Spearman rank correlation coefficient. *Journal of the American Statistical Association, 67*(339), 578-580.

Zhang, T., Raghu, R., & Miron, L. (1996). BIRCH: An efficient data clustering method for very large databases. *ACM SIGMOD Record,* 25(2).

Zhao, Y. (2013). *R and data mining: Examples and studies,* Amsterdam: Elsevier

APPENDICES

A. City Codes

Table A.1 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	Kmaras	67
Istanbul	5	Kocaeli	26	Tokat	47	Diyarbakir	68
Artvin	6	Erzurum	27	Nigde	48	Kars	69
Bolu	7	Malatya	28	Tunceli	49	Mus	70
Antalya	8	Giresun	29	Kırsehir	50	Sırnak	71
Izmir	9	Kutahya	30	Corum	51	Batman	72
Rize	10	Bayburt	31	Denizli	52	Hakkari	73
Kırıkkale	11	Zonguldak	32	Bilecik	53	Bingol	74
Edirne	12	Cankırı	33	Amasya	54	Igdır	75
Adana	13	Mersin	34	Erzincan	55	Van	76
Sinop	14	Konya	35	Sakarya	56	Sanliurfa	77
Burdur	15	Aydın	36	Ordu	57	Siirt	78
Sivas	16	Kastamonu	37	Afyon	58	Mardin	79
Samsun	17	Bursa	38	Gaziantep	59	Bitlis	80
Karaman	18	Balıkesir	39	Kilis	60	Agri	81
Nevsehir	19	Elazıg	40	Hatay	61		
Mugla	20	Gumushane	41	Adiyaman	62		
Kirklareli	21	Yalova	42	Ardahan	63		

B. The CNBC-E Data

The total numbers of 34 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 Schools 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

C. Further Clusters

In the appendix, since the clusters of the agglomerative clustering algorithm are in the structure of a tree, the different levels of the trees are given here.

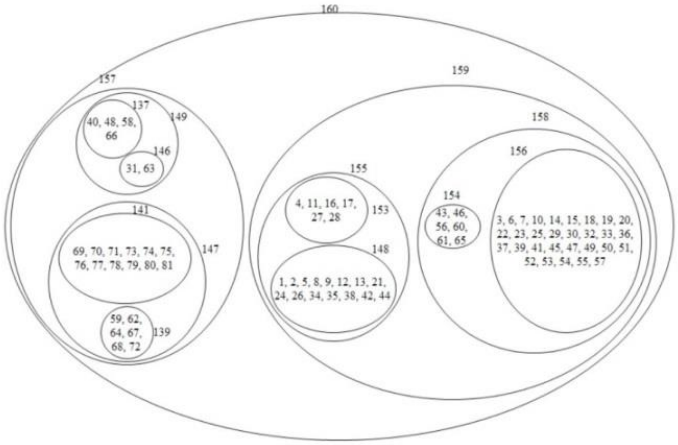


Figure C.1 Nested clusters for k=8

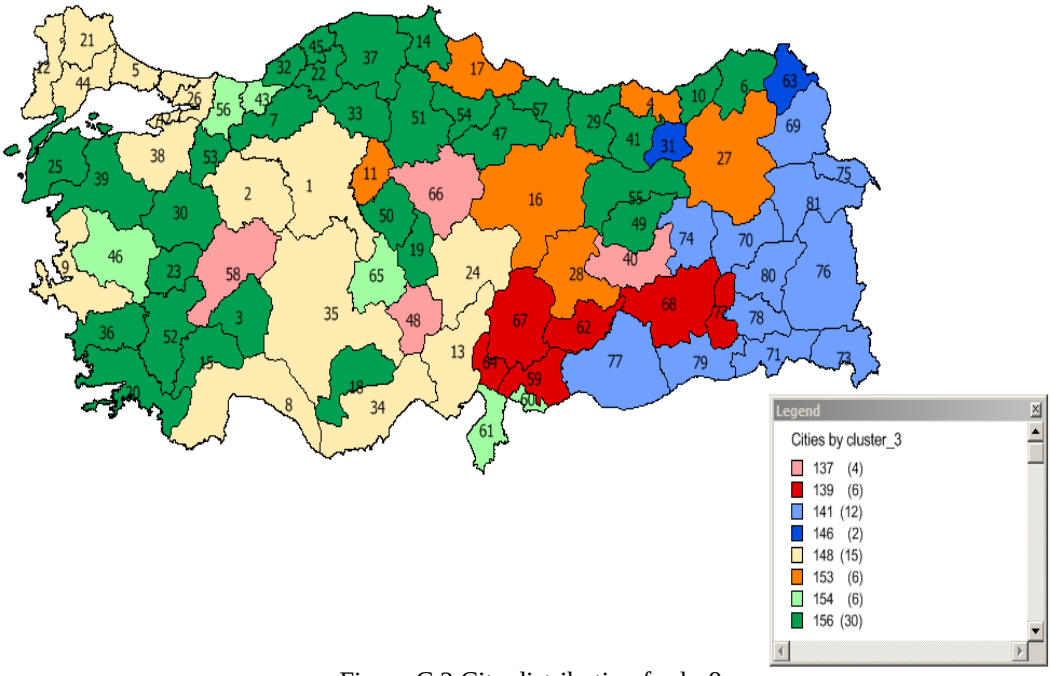


Figure C.2 City distribution for k=8

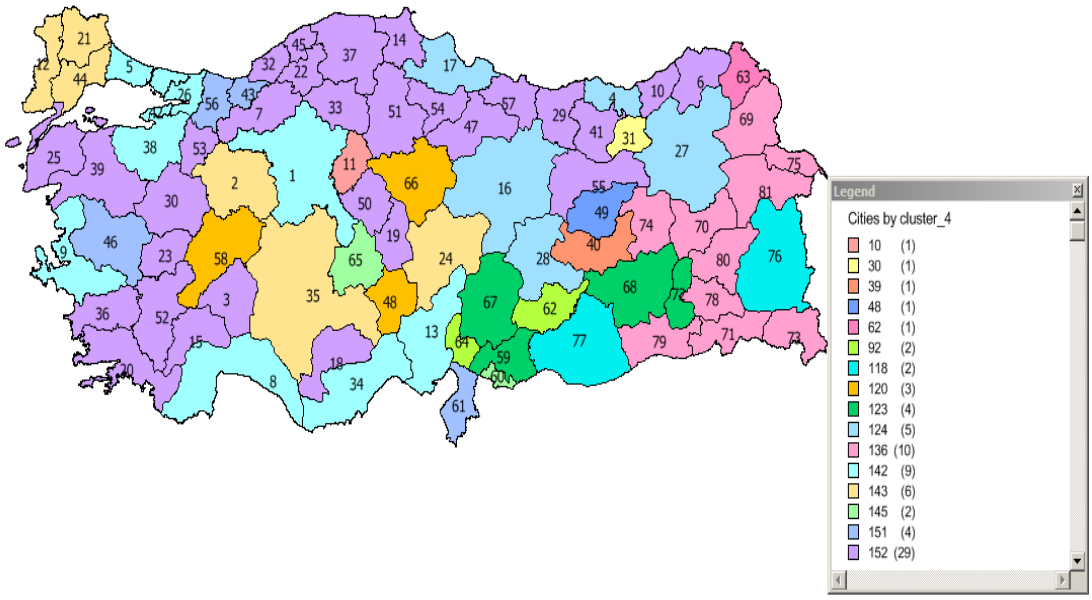


Figure C.3 City distribution for k=16

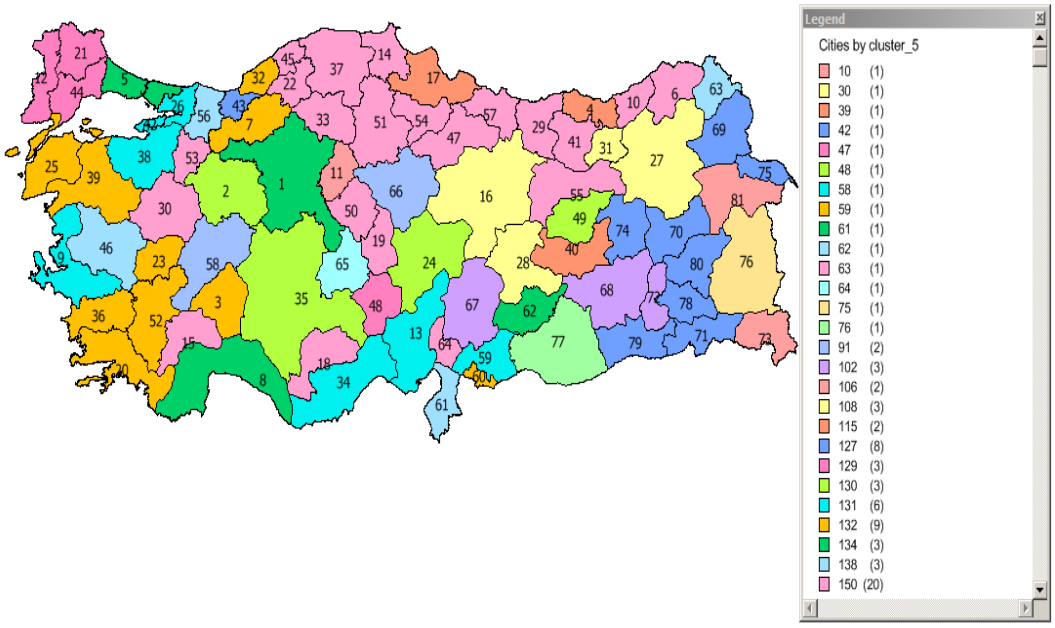


Figure C.4 City distribution for k=27

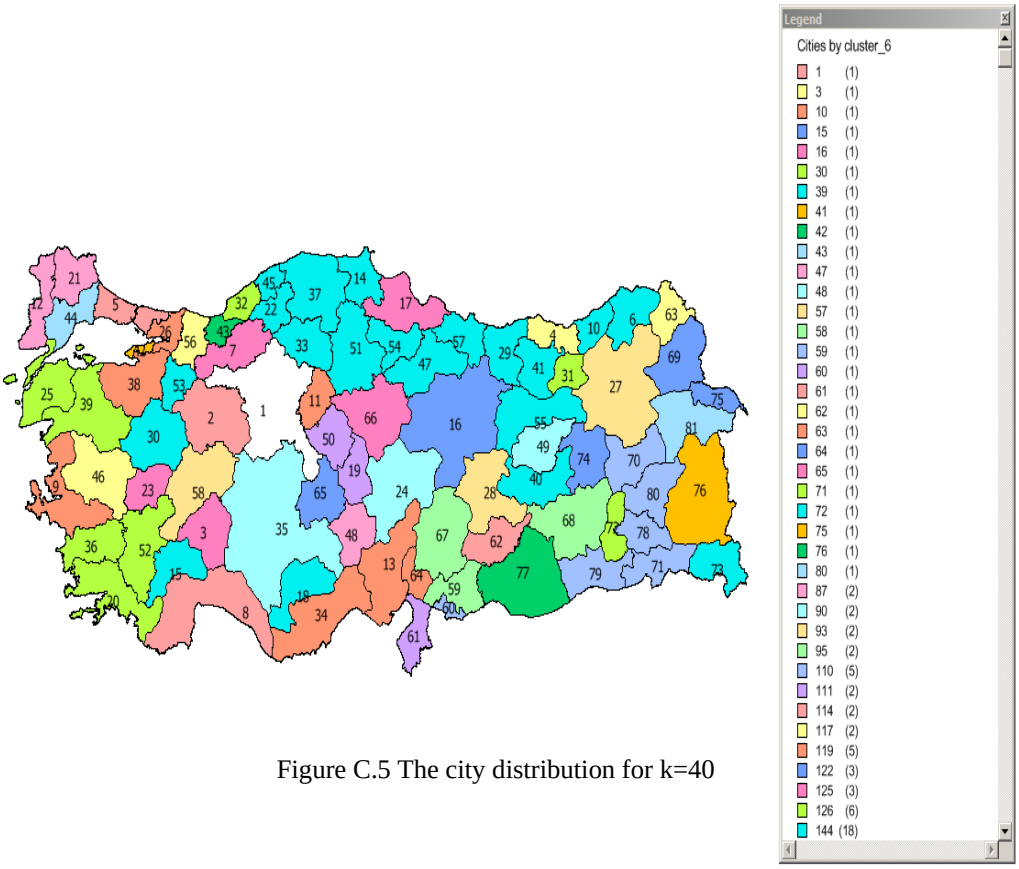


Figure C.5 The city distribution for k=40