ACTA INFOLOGICA 2023;7(1):107-123
dergipark.org.tr/acin

İSTANBUL UNIVERSITY PRESS

DOI: 10.26650/acin.1132510

RESEARCH ARTICLE

Analysis of City Demographics in Turkiye Using Data Mining Techniques

Türkiye’de Şehir Nüfus İstatistiklerinin Veri Madenciliği Yöntemleriyle İncelenmesi

Özge Doğuç[1], Kevser Şahinbaş[2], Gökhan Silahtaroğlu[1]

[1](Prof. Dr.) Medipol University, Faculty of Business and Management Sciences, Department of Management Information Systems, Istanbul, Turkiye
[2] (Asst. Prof.) Medipol University, Faculty of Business and Management Sciences, Department of Management Information Systems, Istanbul, Turkiye

ORCID: Ö.D. 0000-0002-5971-9218;
K.Ş. 0000-0002-8076-3678;
G.S. 0000-0001-8863-8348

Corresponding author:
Özge DOĞUÇ
Medipol University, Faculty of Business and Management Sciences, Department of Management Information Systems, Istanbul, Turkiye
E-mail address: odoguc@medipol.edu.tr

Submitted: 18.06.2022
Revision Requested: 05.09.2022
Last Revision Received: 14.03.2023
Accepted: 12.04.2023
Published Online: 13.06.2023

Citation: Doguc, O., Sahinbas, K. & Silahtaroglu, G.(2023). Analysis of city demographics in Turkiye using data mining techniques. *Acta Infologica, 7*(1), 107-123. https://doi.org/10.26650/acin.1132510

ABSTRACT

In this study, a data mining model has been developed and used to analyze how cities and regions in Turkey can be grouped, aiming to find similarities and differences between them. For this purpose, data is obtained from Turkish Statistical Institution (TUIK) and fuzzy c-means clustering algorithm was used to find categorizations. The data set contains 142 variables from 8 categories such as education, health, happiness, and development levels. The results showed that in all categories, the biggest 3 cities in Turkey, İstanbul, Ankara, and İzmir are different from the rest of the country. Also, cities located in the western and eastern regions of Turkey are mostly grouped among themselves, showing the clear distinction between those two regions. Finally, small cities with big neighbors are grouped with other big cities, showing the direct impact of big cities on their neighbors. This also implies that small cities with no big neighbors are often isolated, as their residents don’t have access to the services provided in the big cities.

Keywords: Data mining, clustering analysis, data management, demographic analysis

ÖZ

Bu çalışmada, Türkiye’deki şehirlerin ve bölgelerin nasıl gruplanabileceğini analiz etmek, aralarındaki benzerlik ve farklılıkları bulmak amacıyla bir veri madenciliği modeli geliştirilmiş ve kullanılmıştır. Bu amaçla Türkiye İstatistik Kurumu’ndan (TÜİK) veriler elde edilmiş ve kategorizasyonları bulmak için bulanık c-ortalamalar kümeleme algoritması kullanılmıştır. Veri setinde eğitim, sağlık, mutluluk ve gelişmişlik düzeyleri gibi 8 kategoriden 142 değişken yer almaktadır. Sonuçlar, Türkiye’nin en büyük 3 şehri olan İstanbul, Ankara ve İzmir’in tüm kategorilerde ülkenin geri kalanından farklı olduğunu gösterdi. Ayrıca, Türkiye’nin batı ve doğu bölgelerinde yer alan şehirler çoğunlukla kendi aralarında gruplandırılarak bu iki bölge arasındaki açık fark görülmektedir. Son olarak, büyük komşuları olan küçük şehirler, büyük şehirlerin komşuları üzerindeki doğrudan etkisini göstererek diğer büyük şehirlerle gruplandırılmıştır. Bu aynı zamanda, sakinlerinin büyük şehirlerde sağlanan hizmetlere erişimi olmadığı için, büyük komşuları olmayan küçük şehirlerin genellikle izole edildiği anlamına gelir.

Anahtar Kelimeler: Veri madenciliği, kümeleme analizi, veri yönetim, demografik analiz

1. INTRODUCTION

Turkey has 81 provinces and 7 geographic regions. Although most cities have unique characteristics, cities in each region may have similar cultural characteristics due to regional similarities. Therefore, a city in the Marmara Region and a city in the Eastern Anatolia region may show similarities in some of their demographic characteristics. Knowing in which features the cities are similar can help with the decisions that are made at the state level. For example, a proposed investment in a city can be planned for other cities that have been experiencing similar problems or have similar needs. Areas that cities need to develop or investments they need can be observed in more detail, and it will be easier for the investments to spread throughout the country if these similarities and differences are well-analyzed.

Data analytics is a science that analyzes existing historical data with statistical and machine learning methods and extracts useful information (Silahtaroğlu, 2008). As its name suggests, the quality of the data to be provided to machine learning is extremely important. For example, providing the age of the person as input can help the method to discover patterns and make predictions, but also providing the sign of the same person, along with the age category (such as child, young, old, very old) will help the machine to reveal different and hidden information about that individual. Similarly, correlating provinces or cities to people, such as the person's place of birth and city of residence helps the data analytic methods to extract information and patterns about populations. At this point, like the age example, in addition to the name of the province, using variables such as its population, literacy, and patients per physician will help the method produce a more detailed result.

This study uses 142 demographic variables about Turkish cities, spanning 5 years. This is the first study in the literature that uses such as wide range of demographic variables in Turkish cities to perform data analytics; as the studies in the literature had much limited scopes. This study provides a general clustering analysis of the cities that was carried out by collecting data from a wide range of categories. Therefore, unsupervised algorithms are used to find the groups of attributes as well as clusters of the cities that provide the best resemblance (similarities). This study aims to help the local governments and institutions by providing a detailed analysis of how and why some of the cities exhibit similarities and discuss how these similarities can be used in budget and investment planning.

2. LITERATURE SURVEY

There are several studies in the literature on cluster analysis using data provided by the Turkish Statistical Institution (TUIK). In one of these studies, Bulut (2019) aimed to cluster Turkish citizens based on the life satisfaction index values. In the study, a total of 11 index values and indicators showing the life satisfaction rates of the citizens were used as variables. The index values used in the study were housing, health, environment, access to infrastructure services, social life, income and wealth status, security, business life, civic participation, education, and life satisfaction. The value range of index values is between 0 and 1, and values closer to 1 represent higher living standards. In the study, Expectation Maximization (EM) and k-means algorithms are used. As a result of the study, the number of interpreted clusters was obtained as 2. Cities in the 2nd cluster generally consist of the cities located in the Southeastern and Eastern regions of Turkey. With this result, it has been seen that the satisfaction or dissatisfaction of the citizens in Turkey is caused by a regional problem. (Bulut, 2019).

In his study, Yilanci (2010) classified the cities in Turkey from a socioeconomic point of view using fuzzy clustering analysis. In addition to the fuzzy cluster analysis method, the k-means method was also used in the study for comparison. The study used 11 socioeconomic variables such as population density, unemployment rate, number of insured people, public investment expenditures, rate of higher education graduates, total agricultural production values, population per physician, infant mortality rate, the total number of people receiving a pension, gross domestic per capita. The study revealed two clusters of cities where the cities in the first cluster can be classified as developed when evaluated in terms of socioeconomics, and the cities in the second cluster can be classified as underdeveloped cities when evaluated in terms of socioeconomics. (Yılancı, 2010).

In the study by İncekirik and Altin (2021), the transportation data of the cities in Turkey were evaluated and classified by cluster analysis. The data set used for the study includes the transportation statistics between the years 2004-2018 and contained 40 variables. As a result, 5 clusters of cities were discovered; where the clustering algorithm placed the cities that are developed and located on the coast with significant export and import activities in different clusters than the cities that did not have advanced transportation alternatives and thus were open to further development. (İncekırık & Altın, 2021).

In another study done by Tekin (2015), health data from 81 provinces of Turkey were examined to identify similar provincial groups. The study used data from 2013 with 16 health indicators; and aimed to provide a comparison of development levels of different cities using health and socioeconomic indicators. The study revealed significant differences in the quality of health services provided in the western and eastern regions of Turkey. (Tekin, 2015).

In Kandemir's (2018) study, the clustering of provinces in Turkey was made according to accommodation statistics. The study aims to determine the cities with priority in tourism and discuss planning to allocate resources to the cities that do not have priority. Using the data from 2016, the study showed that Ankara, Antalya, and Istanbul do not have a definite cluster membership with other cities and the distance between cities does not affect this situation. When it comes to domestic and foreign tourists, it is seen that Antalya is the most preferred city, as expected. And for foreign tourists, Çorum province is the least preferred and according to the number of domestic tourists, Osmaniye is the least preferred province (Kandemir, 2018).

This study analyzes the demographics of Turkish cities across 142 attributes by using an unsupervised learning method, clustering. Unlike the previous studies in the literature, this study analyzes the demographic data from a broad perspective and finds demographic attributes that show a resemblance between Turkish cities. The sections discuss the clustering methods that are used in this study and introduce the data set that is used for generating results.

2.1 K-MEANS CLUSTERING ALGORITHM

K-means is a popular clustering algorithm that is used to classify data. The algorithm divides the data to be classified into k classes or clusters based on their properties. Classification is done by the distribution of the selected data around the center points of the clusters, which show the closest features in terms of similarity to each other. The name of the algorithm comes from the requirement for the number of clusters to be known and constant. The number of clusters is expressed with the letter k, and it is also the representation of the number of clusters that will be created according to the similarity between the data. (Silahtaroğlu, 2004).

Figure 1 illustrates how the k-means algorithm works. In the visualized example, k is taken as 3. The white symbols (triangle, circle, square) seen in (a) on the left resemble the initial cluster centers, and they were chosen randomly. Afterward, the data points shown in the first iteration (b) are assigned to the nearest center and shown with the same symbols (triangle, circle, square). Next, the cluster centers are recalculated, considering the average of the elements in each cluster. Cluster centers (shown by looking at moving arrows) are redetermined. Steps b and c are performed again if there is a change in the cluster centers. In the example in the figure, the algorithm ends up dividing it into three clusters. (Böhm, 2001).

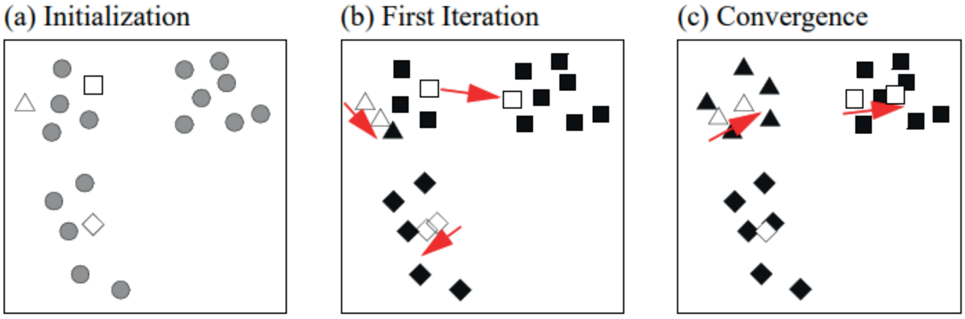

Figure 1. Representation of the K-Means algorithm (Böhm, 2001)

2.2 FUZZY C-MEANS ALGORITHM

The Fuzzy c-means algorithm was first introduced by Dunn in 1973 and implemented by Bezden in 1981. (Höppner et al., 1999) It is one of the most popular and most used fuzzy clustering algorithms in the literature. Different than regular clustering algorithms such as k-means, fuzzy c-means allows data elements to be assigned to multiple clusters. (Kruse et al., 1999) It uses 'goal functions' for cluster assignments and tries to minimize the goal score at every iteration. (Işık & Çamurcu, 2011). Like the k-means algorithm, it reevaluates the cluster centers and recalculates the goal function at every iteration. The algorithm iterates until the goal score is less than ε. (Silahtaroğlu, 2013). Like the k-means algorithm, the performance of fuzzy c-means is dependent on the initial cluster center assignments. (Hekim & Orhan, 2011). However, unlike the k-means algorithm, fuzzy c-means provide 'soft assignments' to the clusters, where each assignment contains a probability. Therefore, the fuzzy c-means algorithm can track an unlimited number of cluster assignments and pick the ones with the highest probabilities.

3. DATA ACQUISITION AND METHOD

The data set used in this study was collected from multiple data sources through TUIK. Data set contains data from 2015-2020 and includes 23 major categories as follows:

> *"Education", "Demography", "Population and Migration", "Health", "Environment", "Transportation", "Justice", "Foreign Trade", Energy", "Construction and Housing", "Culture", "Industry", "Agriculture and Livestock", "Tourism", "National Accounts", "Working Life", "Civil Participation", "Income and Wealth", "Security", "Access to Infrastructure Services", "Social Life", "Happiness" Level".*

Each category involves a number of variables; and thus, the data set contains 142 variables in total. Table 1 below shows the variables in each category.

Table 1
Demographic categories and variables used in this study

Category Name	Variables
General Information	Population Female Population Male Population Population Percentage Area (Km^2) Number Of Districts Number Of Municipalities Number Of Villages
Education	Literacy rate (%) Number of primary school students Number of secondary school students Number of high school students Number of primary school teachers Number of secondary school teachers Number of high school teachers Number of primary schools Number of secondary schools Number of high schools Primary and Secondary Schools / Number of Students Per Classroom High Schools / Number of Students Per Classroom Number of Students Per Teacher (Primary School) Number of Students Per Teacher (Secondary School) Number of Students Per Teacher (High School) Primary School Enrolment Rate (female) Primary School Enrolment Rate (overall) Primary + Secondary Schooling Rate (female) Primary + Secondary Schooling Rate (overall) High School Enrolment Rate (female) High School Enrolment Rate (overall) Net enrolment rate in pre-primary education (3-5 years) (%) The mean score for placing the TEOG system. YGS average score Percentage of faculty or college graduates (%) Satisfaction rate of public education services (%)

Demography	Infant Mortality Rate (per thousand)
	Number of Divorces
	Number of Births
	Expected Lifespan at Birth (years)
	Number of Marriages
	Divorce Rate (per thousand)
	Birth Rate (per thousand)
	Marriage Rate (per thousand)
	Death Rate (per thousand)
	Number of Deaths
	Total Fertility Rate (number of children)
	Under-5 Mortality Rate (per thousand)
Population And Migration	Child Dependency Rate (%)
	Net Migration Rate (per thousand)
	Population Density (number of people per square kilometer)
	Average Household Size
	Total Number of Households
	Total Age Dependency Rate (%)
	Immigration from Turkey to Abroad
	Elderly Dependency Rate (%)
	Annual Population Growth Rate (per thousand)
	Migration from Abroad to Turkey
Health	Total Number of Physicians Per Thousand People
	Number of Hospitals
	Number of Hospital Beds
	Total Number of Hospital Beds per Hundred Thousand People
	Number of Applications Per Physician
	Health satisfaction rate (%)
	Satisfaction rate of public health services (%)
Environment	Wastewater Collection Rate (%)
	Wastewater Recycle Rate (%)
	Access to Potable Water Rate (%)
	Potable Water Recycle Rate (%)
	(Air Pollution) Average of PM10 Station Values (μg/m^3)
	Forest Area per km^2 (%)
	Rate of Street Noise Complaints (%)
	Cleaning Services Satisfaction Rate (%)
Transport	Number of Cars per Thousand People
	Number of Land Vehicles
	Number of Cars
	Number of Traffic Accidents
Justice	Number of Convicts Entering the Penitentiary Institution According to the Committed Crime
Foreign Trade	Total Exports (thousand $)
	Total Imports (thousand $)
Energy	Total Electricity Consumption per Person (kWh)
Construction And Housing	Number of Housing sales (first sale)
	Number of Housing Sales (total)
	Number of Buildings by Occupancy Permit
	Number of Flats by Occupancy Permit
	Area According to the Occupancy Permit (square meters)
	Number of Buildings by Building Permit
	Number of Flats by Building Permit
	Area According to Building Permit (square meters)
	Number of Rooms Per Person
	Availability of toilets in the residences (%)
	Housing services quality satisfaction rate (%)

Culture	Number of Public Library Users Number of Books in Public Libraries Number of Public Libraries Number of Museum Artifacts Affiliated with the Ministry of Culture and Tourism Number of Museums Affiliated to the Ministry of Culture and Tourism Number of Museum Visitors Number of Movie Theaters Number of Cinema Audiences Number of Theater Halls Number of Theater Audiences
Industry	Total Number of Initiatives
Agriculture And Livestock	Crop Production Value (thousand TL) Number of Cattle (head) Live Animals Value (thousand TL) Animal Products Value (thousand TL) Number of Ovine (head) Greenhouse Vegetable and Fruit Production (tons) Cereals and Other Herbal Products Production (tons) Total Agricultural Area (hectares)
Tourism	Total number of overnight stays Total Number of Arrivals (person) Number of Foreign Stays Number of Foreign Arrivals (person)
National Accounts	GDP (thousand TL) GDP per Capita ($) GDP per Capita (TL)
Work Life	Employment rate (%) Unemployment rate (%) Average daily earnings (TL) Job satisfaction rate (%)
Civil Participation	Local Administrations Election Rate (%) Political Parties Membership Rate (%) Rate of Union/Association Membership (%)
Income And Wealth	Savings Deposits Per Capita (TL) Percentage of Households in the Middle- and Upper-Income Group (%) Percentage of Households That Do Not Meet Basic Needs (%)
Security	Murder Rate (per one million people) Number of Fatal and Injured Traffic Accidents (per thousand) Rate of Feeling Safe When Outside at Night (%) Satisfaction Rate of Public Security Services (%)
Access To Infrastructure Services	Number of Internet Subscribers (per hundred) Access to Sewerage and Potable Water (%) Airport Access Rate (%) Public Transportation Services Satisfaction Rate (%)
Social Life	Shopping Center Area Per Thousand People (m^2) Social Relations Satisfaction Rate (%) Social Life Satisfaction Rate (%)
Happiness Level	Life Happiness Level (%)

This study uses the Fuzzy C-means clustering algorithm to analyze TUIK data from different perspectives. After an initial cluster analysis is done with all 142 variables, data is further analyzed from 7 different perspectives. The next section provides details about the cluster analysis. Figure 2 summarizes the approach taken in this study.

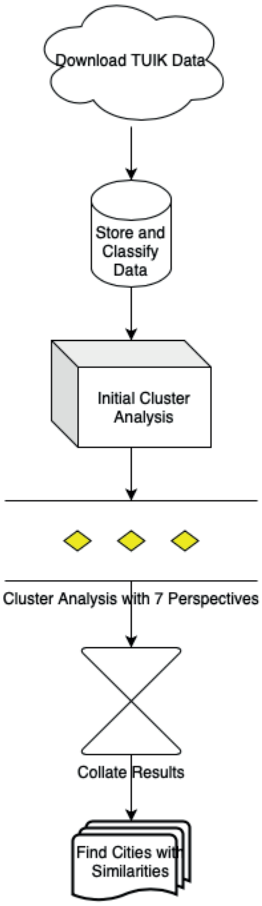

Figure 2. Workflow of the approach used in this study.

4. CLUSTER ANALYSIS

In the analysis section, the major categories in the data were examined in order to find similar ones among the cities. For data analysis, this study uses an enhanced version of the Fuzzy C-means clustering with a variable number of clusters. The study uses clustering indices such as Xie Beni and partition coefficient along with the Fuzzy C-means algorithm's scoring function, to find the most optimal number of clusters. Initially, all 142 variables were used in a general cluster analysis to identify overall similarities between the cities. Next, data is analyzed from 7 perspectives such as "Education", "Population and Settlement", "Livelihood, Income, and Purchasing Power", "Health", "Satisfaction", "Security", "Cultural Opportunities and Utilization" and at the same time, a clustering analysis was performed with the variables. Thus, by working on the data from 8 aspects, similarities between the cities across different categories were considered. Cities are colored based on the clusters they belong to and presented on the maps. The analysis results obtained when all variables are included in the analysis are shown in the table below. While determining the cluster numbers, Xie Beni, Partition Entropy, and Partition Coefficient values were used, different cluster numbers were tried and the number of clusters that gave the best coefficient was presented as the final cluster number in the study.

In the general clustering analysis performed with 142 different variables, 9 clusters were obtained. The provinces and the clusters they belong to are shown on the map in Figure 3. As can be seen, Istanbul, Ankara, Konya, and Izmir are separated from all provinces, and each is represented as a cluster.

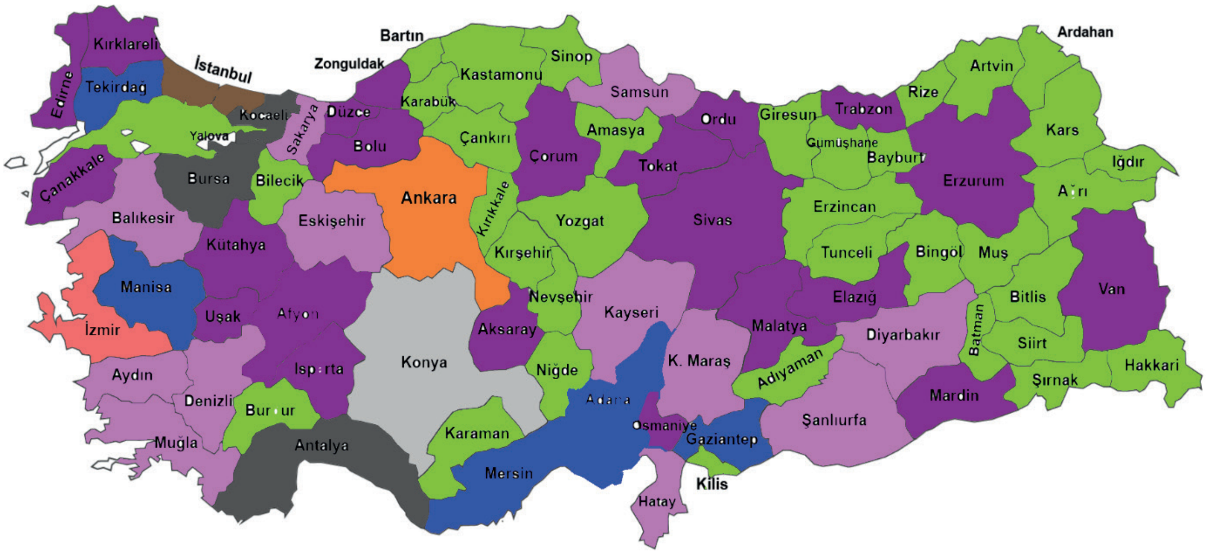

Figure 3. Representation of the clustering study based on all 142 variables on the map.

Cluster 0: Aydın, Balıkesir, Denizli, Diyarbakır, Eskişehir, Hatay, Kahramanmaraş, Kayseri, Muğla, Sakarya, Samsun, Şanlıurfa

Cluster 1: Antalya, Bursa, Kocaeli

Cluster 2: İstanbul

Cluster 3: Adıyaman, Amasya, Ardahan, Artvin, Ağrı, Bartın, Batman, Bayburt, Burdur, Bilecik, Bingöl, Bitlis, Erzincan, Gümüşhane, Giresun, Hakkâri, Iğdır, Karabük, Karaman, Kars, Kastamonu, Kırıkkale, Kırşehir, Kilis, Muş, Nevşehir, Niğde, Rize, Sinop, Siirt, Tunceli, Yalova, Yozgat, Çankırı, Şırnak

Cluster 4: Adana, Gaziantep, Manisa, Mersin, Tekirdağ

Cluster 5: Konya

Cluster 6: Ankara

Cluster 7: İzmir

Cluster 8: Afyonkarahisar, Aksaray, Bolu, Düzce, Edirne, Elâzığ, Erzurum, Isparta, Kırklareli, Kütahya, Malatya, Mardin, Ordu, Osmaniye, Sivas, Tokat, Trabzon, Uşak, Van, Zonguldak, Çanakkale, Çorum

It is observed from the cluster results that the biggest 3 cities in Turkey, Istanbul, Ankara, and İzmir are located in separate clusters. Cluster 0 contains the cities that are developing and have more than 1M residents, although they are geographically separate. Similarly, Cluster 3 contains underdeveloped cities. The initial clustering analysis with all variables revealed that the dataset can be used to distinguish developed and underdeveloped cities. So, as a next step data is analyzed focusing on different perspectives.

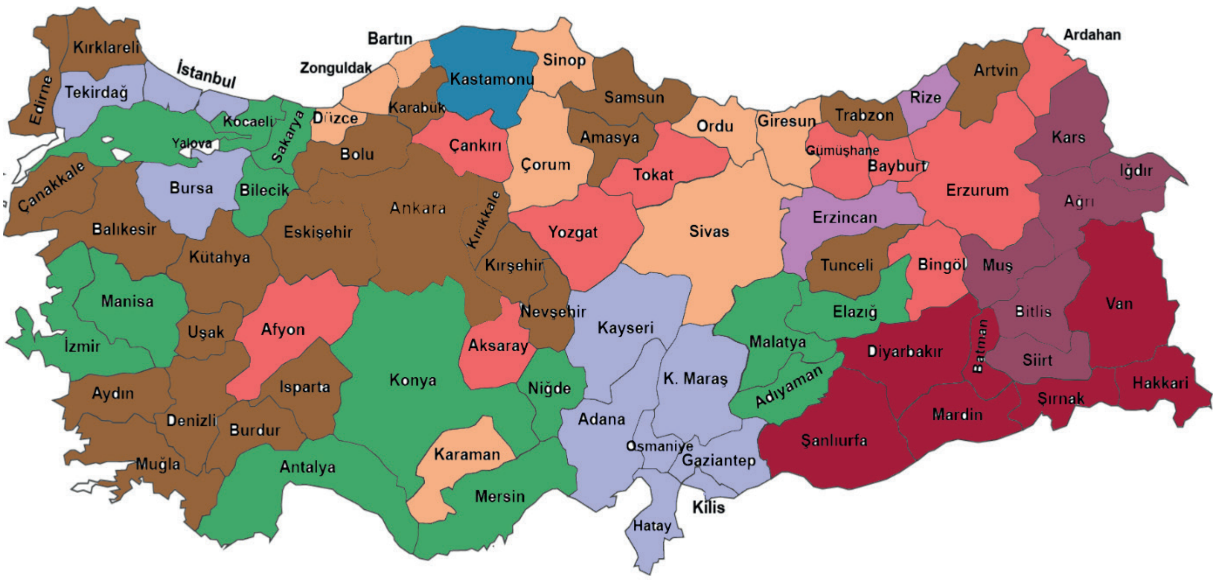

Figure 4. Representation of the cities based on clustering analysis for *education*.

Cluster 0: Amasya, Ankara, Artvin, Aydın, Balıkesir, Bolu, Burdur, Denizli, Edirne, Eskişehir, Isparta, Karabük, Kırıkkale, Kırklareli, Kırşehir, Kütahya, Muğla, Nevşehir, Samsun, Trabzon, Tunceli, Uşak, Çanakkale

Cluster 1: Erzincan, Rize

Cluster 2: Adıyaman, Antalya, Bilecik, Elâzığ, Kocaeli, Konya, Malatya, Manisa, Mersin, Niğde, Sakarya, Yalova, İzmir

Cluster 3: Kastamonu

Cluster 4: Ağrı, Bitlis, Iğdır, Kars, Muş, Siirt

Cluster 5: Bartın, Düzce, Giresun, Karaman, Ordu, Sinop, Sivas, Zonguldak, Çorum

Cluster 6: Batman, Diyarbakır, Hakkâri, Mardin, Van, Şanlıurfa, Şırnak

Cluster 7: Adana, Bursa, Gaziantep, Hatay, Kahramanmaraş, Kayseri, Kilis, Osmaniye, Tekirdağ, İstanbul

Cluster 8: Afyonkarahisar, Aksaray, Ardahan, Bayburt, Bingöl, Erzurum, Gümüşhane, Tokat, Yozgat, Çankırı

There is a resemblance between the clusters in Figure 4 with the clusters created with all 142 variables. Some of the developing cities such as Istanbul, Adana, and Bursa are in Cluster 7 along with their neighbor cities, although they are not developed (e.g., Osmaniye and Kilis). This suggests that the education level of a developed city can impact its neighbors. This can be explained by students traveling or relocating to big neighboring cities for advanced education.

On the contrary, clusters 1 – 5 contain mostly underdeveloped cities with no neighboring big cities. Residents of these small cities do not have easy access to higher education institutions that are generally located in big cities.

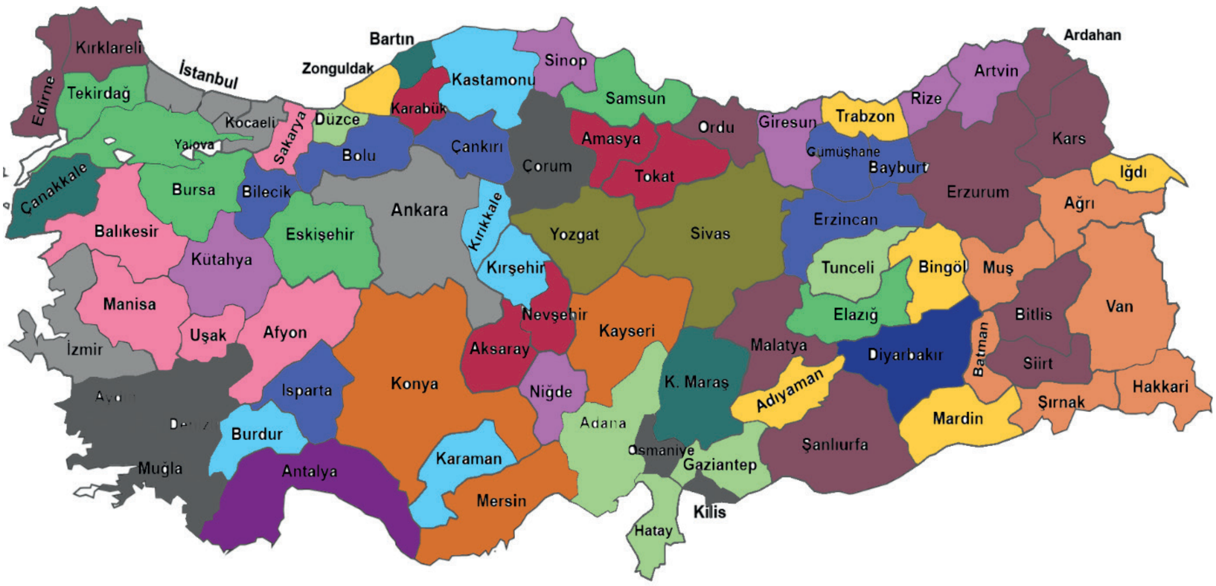

Figure 5. Representation of the cities based on clustering analysis for *security*.

Cluster 0: Ankara, Kocaeli, İstanbul, İzmir

Cluster 1: Ağrı, Batman, Hakkâri, Muş, Van, Şırnak

Cluster 2: Burdur, Karaman, Kastamonu, Kırıkkale, Kırşehir

Cluster 3: Afyonkarahisar, Balıkesir, Manisa, Sakarya, Uşak

Cluster 4: Artvin, Giresun, Kütahya, Niğde, Rize, Sinop

Cluster 5: Bursa, Elâzığ, Eskişehir, Samsun, Tekirdağ, Yalova

Cluster 6: Aksaray, Amasya, Karabük, Nevşehir, Tokat

Cluster 7: Aydın, Denizli, Kilis, Muğla, Osmaniye, Çorum

Cluster 8: Kayseri, Konya, Mersin

Cluster 9: Bayburt, Bolu, Bilecik, Erzincan, Gümüşhane, Isparta, Çankırı

Cluster 10: Adıyaman, Bingöl, Iğdır, Mardin, Trabzon, Zonguldak

Cluster 11: Diyarbakır

Cluster 12: Adana, Düzce, Gaziantep, Hatay, Tunceli

Cluster 13: Ardahan, Bitlis, Edirne, Erzurum, Kars, Kırklareli, Malatya, Ordu, Siirt, Şanlıurfa

Cluster 14: Bartın, Kahramanmaraş, Çanakkale

Cluster 15: Sivas, Yozgat

Cluster 16: Antalya

Although there are more clusters than the previous analysis results, the biggest 3 cities (Istanbul, Ankara, and Izmir) are placed in the same cluster. Also, the developed and underdeveloped cities show different security characteristics and are therefore placed in different clusters. Cities that are in the western and eastern regions are put in separate clusters. Clusters 13 and 14 are the only exceptions to this where Ardahan-Edirne and Kahramanmaraş-Çanakkale are placed in the same cluster. It can be observed that political instability in the eastern regions over the years and the refugee migrations due to the Syrian internal conflict affected Security in the cities in eastern and Southeastern regions of Turkey.

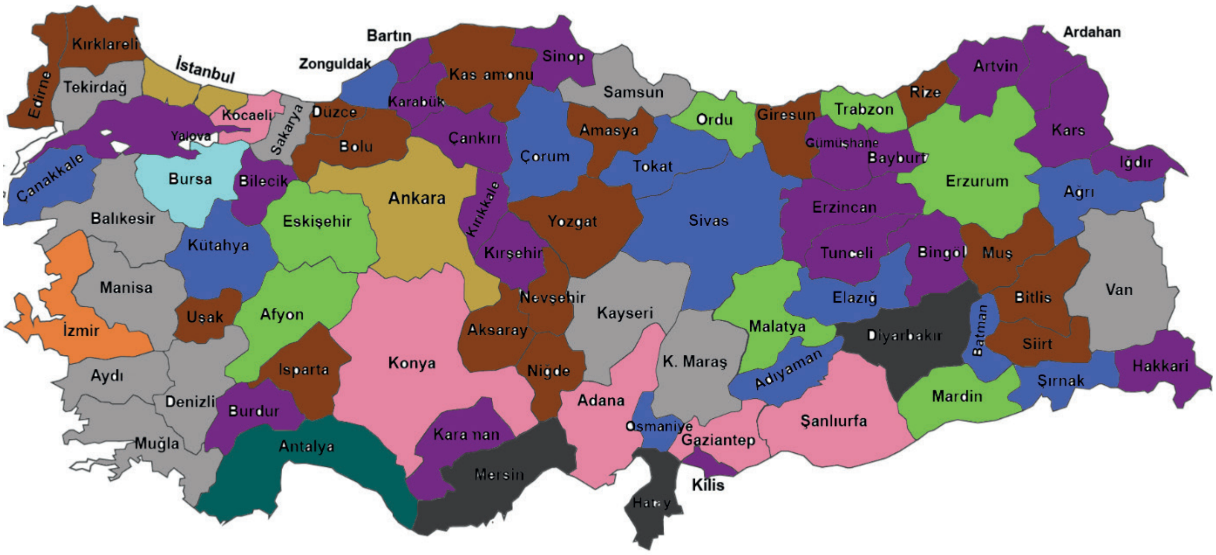

Figure 6. Representation of the cities based on clustering analysis for *population*.

Cluster 0: Aydın, Balıkesir, Denizli, Kahramanmaraş, Kayseri, Manisa, Muğla, Sakarya, Samsun, Tekirdağ, Van

Cluster 1: Adıyaman, Ağrı, Batman, Elâzığ, Kütahya, Osmaniye, Sivas, Tokat, Zonguldak, Çanakkale, Çorum, Şırnak

Cluster 2: İzmir

Cluster 3: Adana, Gaziantep, Kocaeli, Konya, Şanlıurfa

Cluster 4: Afyonkarahisar, Erzurum, Eskişehir, Malatya, Mardin, Ordu, Trabzon

Cluster 5: Ardahan, Artvin, Bartın, Bayburt, Burdur, Bilecik, Bingöl, Erzincan, Gümüşhane, Hakkâri, Iğdır, Karabük, Karaman, Kars, Kırıkkale, Kırşehir, Kilis, Sinop, Tunceli, Yalova, Çankırı

Cluster 6: Bursa

Cluster 7: Ankara, İstanbul

Cluster 8: Aksaray, Amasya, Bolu, Bitlis, Düzce, Edirne, Giresun, Isparta, Kastamonu, Kırklareli, Muş, Nevşehir, Niğde, Rize, Siirt, Uşak, Yozgat

Cluster 9: Antalya

Cluster 10: Diyarbakır, Hatay, Mersin

Not surprisingly, the top 5 biggest cities such as İstanbul, Ankara, İzmir, Bursa, and Antalya are separated from the rest of the cities. Clusters 3 and 10 show the developing cities with more than 1 million residents, and the rest of the cities were placed in clusters based on their geographic regions. These results are on par with the previous clustering results with other perspectives where large and small cities are placed in separate clusters.

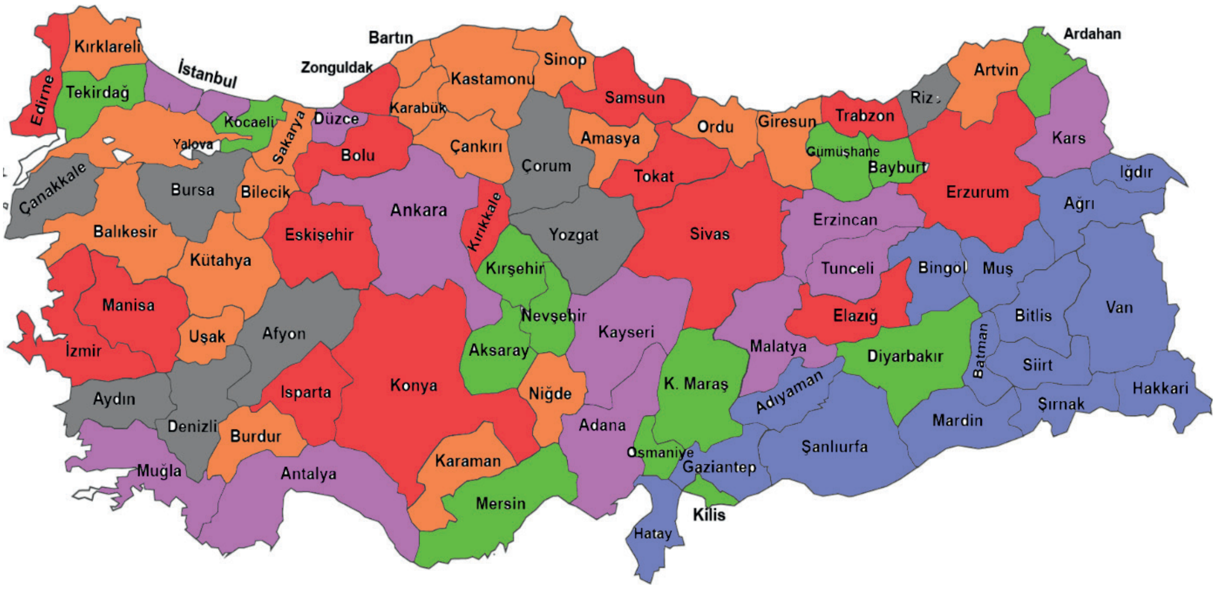

Figure 7. Representation of the cities based on clustering analysis for *health services*.

Cluster 0: Amasya, Artvin, Balıkesir, Bartın, Burdur, Bilecik, Giresun, Karabük, Karaman, Kastamonu, Kırklareli, Kütahya, Niğde, Ordu, Sakarya, Sinop, Uşak, Yalova, Çankırı

Cluster 1: Aksaray, Ardahan, Bayburt, Diyarbakır, Gümüşhane, Kahramanmaraş, Kırşehir, Kocaeli, Kilis, Mersin, Nevşehir, Osmaniye, Tekirdağ

Cluster 2: Bolu, Edirne, Elâzığ, Erzurum, Eskişehir, Isparta, Kırıkkale, Konya, Manisa, Samsun, Sivas, Tokat, Trabzon, Zonguldak, İzmir

Cluster 3: Afyonkarahisar, Aydın, Bursa, Denizli, Rize, Yozgat, Çanakkale, Çorum

Cluster 4: Adana, Ankara, Antalya, Düzce, Erzincan, Kars, Kayseri, Malatya, Muğla, Tunceli, İstanbul

Cluster 5: Adıyaman, Ağrı, Batman, Bingöl, Bitlis, Gaziantep, Hakkâri, Hatay, Iğdır, Mardin, Muş, Siirt, Van, Şanlıurfa, Şırnak

These results clearly show that geography has a direct impact on access to health services. Cluster 5 contains cities in the eastern regions only. Diyarbakır and Erzurum are the only 2 exceptions, as these cities are bigger than the others in the eastern regions. Smaller cities in the western regions were expected to be placed in the same cluster with their larger neighbors, as residents of the small cities can easily travel to a large city nearby to receive health services. However, this is not the case: Tekirdağ and Kocaeli are in a different cluster than İstanbul; Aydın and İzmir, Kırşehir and Ankara are also good examples of this case. These results suggest that the health standards in Western cities are different from the ones in Eastern cities. Also, in the Western regions small cities have similar health standards to the large cities. Large cities in the Western regions have better and bigger hospitals and large numbers of health professionals, so residents living in their neighboring cities can also benefit from these services. One can even argue that in the Western regions the 'number of doctors per capita' is higher in small cities.

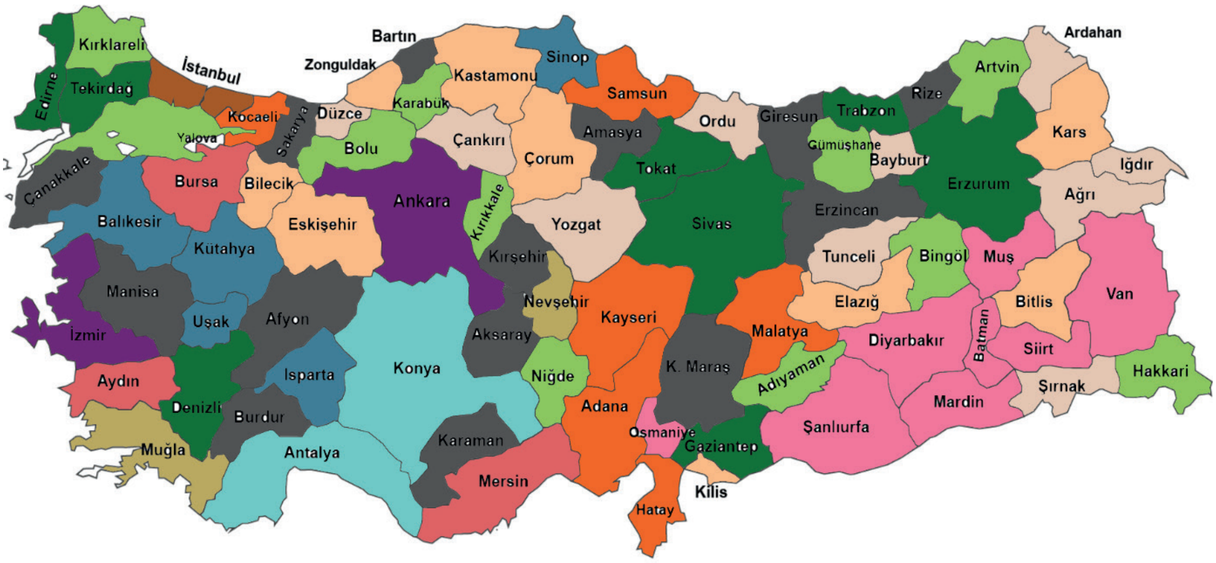

Figure 8. Representation of the cities based on clustering analysis for *cultural and social activities*.

Cluster 0: Afyonkarahisar, Aksaray, Amasya, Bartın, Burdur, Erzincan, Giresun, Kahramanmaraş, Karaman, Kırşehir, Manisa, Rize, Sakarya, Çanakkale

Cluster 1: Antalya, Konya

Cluster 2: İstanbul

Cluster 3: Batman, Diyarbakır, Mardin, Muş, Osmaniye, Siirt, Van, Şanlıurfa

Cluster 4: Adıyaman, Artvin, Bolu, Bingöl, Gümüşhane, Hakkâri, Karabük, Kırıkkale, Kırklareli, Niğde, Yalova

Cluster 5: Balıkesir, Isparta, Kütahya, Sinop, Uşak

Cluster 6: Bilecik, Bitlis, Elâzığ, Eskişehir, Kars, Kastamonu, Kilis, Zonguldak, Çorum

Cluster 7: Aydın, Bursa, Mersin

Cluster 8: Muğla, Nevşehir

Cluster 9: Adana, Hatay, Kayseri, Kocaeli, Malatya, Samsun

Cluster 10: Ankara, İzmir

Cluster 11: Ardahan, Ağrı, Bayburt, Düzce, Iğdır, Ordu, Tunceli, Yozgat, Çankırı, Şırnak

Cluster 12: Denizli, Edirne, Erzurum, Gaziantep, Sivas, Tekirdağ, Tokat, Trabzon

The biggest 3 cities (İstanbul, Ankara, and İzmir) are isolated from the rest of the cities. This suggests that these cities attract the most cultural events and activities.

The rest of the cities are not very different from each other. Regardless of their sizes and locations, cities in Turkey (except for the top 3) have similar (and probably low) access to cultural activities.

The only exceptions to this are Antalya and Konya: Antalya is the tourism and conferencing center of Turkey, attracting millions of tourists every year. There are shows and cultural activities held throughout the year in the city, mostly for tourists. Similarly, Konya is the hometown of the world-famous philosopher and thinker Rumi. Millions of domestic and international tourists visit Konya to attend Rumi attractions throughout the year.

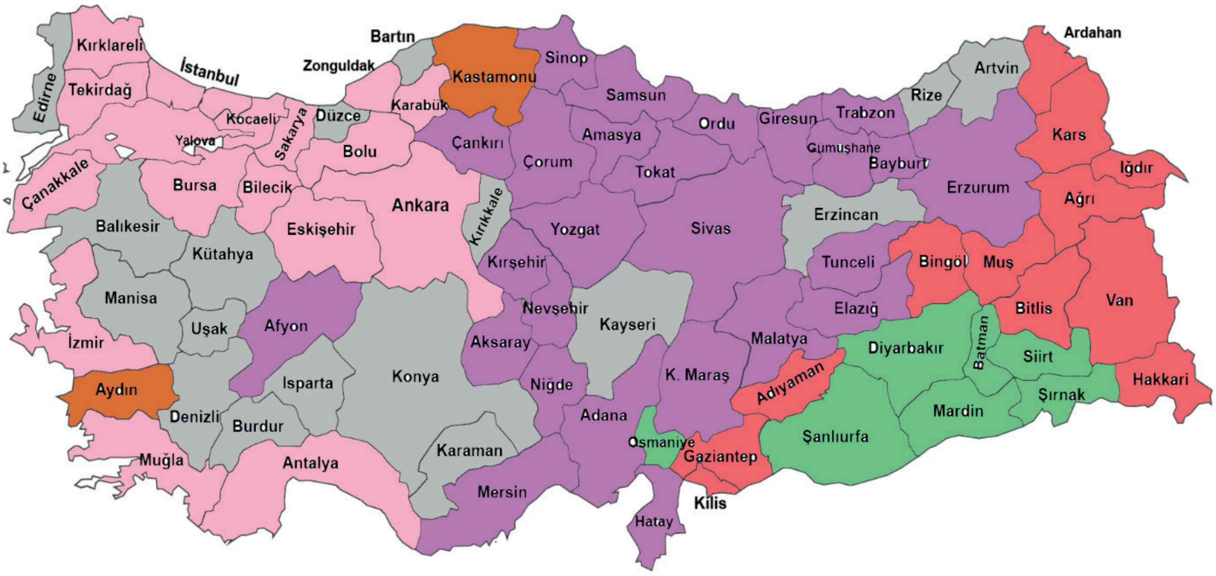

Figure 9. Representation of the cities based on clustering analysis for *purchasing power*.

Cluster 0: Adıyaman, Ardahan, Ağrı, Bingöl, Bitlis, Gaziantep, Hakkâri, Iğdır, Kars, Kilis, Muş, Van

Cluster 1: Artvin, Balıkesir, Bartın, Burdur, Denizli, Düzce, Edirne, Erzincan, Isparta, Karaman, Kayseri, Kırıkkale, Konya, Kütahya, Manisa, Rize, Uşak

Cluster 2: Adana, Afyonkarahisar, Aksaray, Amasya, Bayburt, Elâzığ, Erzurum, Gümüşhane, Giresun, Hatay, Kahramanmaraş, Kırşehir, Malatya, Mersin, Nevşehir, Niğde, Ordu, Samsun, Sinop, Sivas, Tokat, Trabzon, Tunceli, Yozgat, Çankırı, Çorum

Cluster 3: Aydın, Kastamonu

Cluster 4: Batman, Diyarbakır, Mardin, Osmaniye, Siirt, Şanlıurfa, Şırnak

Cluster 5: Ankara, Antalya, Bolu, Bursa, Bilecik, Eskişehir, Karabük, Kırklareli, Kocaeli, Muğla, Sakarya, Tekirdağ, Yalova, Zonguldak, Çanakkale, İstanbul, İzmir

Not surprisingly, the top 3 cities are located within the same cluster, along with several cities in the western regions. (Cluster 5) On the other hand, all eastern cities are placed in the same clusters (Cluster 0, 2, and 4) with no exceptions. This clustering analysis clearly shows how much the western and eastern regions of Turkey are separated from each other. From the economic perspective, the western and eastern regions are so divided that there are no exceptions in the clustering analysis results.

4.1 SIMILARITY ANALYSIS

As a next step in this study, a dual similarity analysis of the cities was performed. In the clustering analysis results detailed in the previous section, it can be observed that some of the cities are placed in the same cluster frequently by the algorithm. For example, Batman and Şırnak are two neighboring cities with similar population sizes, and they were placed in the same cluster in 7 out of 8 perspectives. While this result can be expected, Bartın and Karaman, two geographically distant cities were placed in the same cluster 7 times as well. These cities have similar sizes, and both are in the western part of Turkey. These results coincide with the previous ones from the clustering analysis, as the population sizes of the cities and their location in the east-west direction play important role in showing similarities between them.

Table 2 shows the cities that are placed in the same cluster for 6 or more times. Except for the Balıkesir-Uşak couple, all other city groups have similar sizes and are also in the same East-West region of Turkey.

Table 2
List of cities that are placed in the same cluster 6 or more times.

City 1	City 2	Size Comparison	Geographic Location (East / West)
Bartın	Karaman	Small / Small	West / West
Batman	Şırnak	Small / Small	East / East
Aydın	Denizli	Small / Small	West / West
Amasya	Giresun	Small / Small	East / East
Ağrı	Muş	Small / Small	East / East
Ağrı	Şırnak	Small / Small	East / East
Bayburt	Gümüşhane	Small / Small	East / East
Bayburt	Çankırı	Small / Small	East / East
Burdur	Karaman	Small / Small	West / West
Bitlis	Kars	Small / Small	East / East
Karabük	Yalova	Small / Small	West / West
Isparta	Uşak	Small / Small	West / West
Kütahya	Uşak	Small / Small	West / West
Balıkesir	Uşak	Big / Small	West / West

Turkey is divided into 7 geographic regions, where each region consists of roughly 10-15 cities. The central government provides funding and investment to the cities based on their geographic regions (Arslan, 2014) However, clustering analysis performed in this study showed that most of the cities that are in the same geographic regions are not necessarily similar and thus placed in different clusters.

Using the results presented in this study, further analysis may reveal groups of cities that are actually very similar to each other. In future work, analysis should be done for finding city groups of 5 and more and define the similarities in each group.

5. DISCUSSION AND CON

In this study, demographic characteristics of the 81 cities in Turkey have been analyzed for similarities using variables from various categories. These categories are happiness, income and purchasing power, sporting and cultural activities, health and social protection, population and housing, security, and education. Demographic data for cities are obtained from TUIK, and data spans years between 2015 and 2020. Initially, the study performs a clustering analysis done all variables across all categories to illustrate the general demographic profile of the cities in Turkey. Next, cities were analyzed from seven different perspectives, as listed above. Finally, results from all 8-clustering analyses are compared to find city couples that are often placed in the same cluster. This study aims to help the local governments' budget and investment planning, by providing insights into city demographics and highlighting similarities between the cities that are geographically separate.

At each step, cities were segmented into different numbers of clusters, to find the most optimal cluster setup. The fuzzy c-means clustering algorithm, which evaluates cluster probabilities and decides the cluster assignments to maximize probability values is used for this purpose. The largest 3 cities in Turkey; Istanbul, Ankara, and İzmir are often placed in the same cluster and separated from the other cities. These cities not only have higher populations than others but also, they are located in the western regions of the country. Clustering results also show that cities in the western and eastern regions are often placed in separate clusters. While neighboring cities are expected to be placed in the same cluster, this study showed that cities in Turkey are categorized based on their east-west positions: Cities in the west are categorized together although they are not necessarily neighbors. The same outcome can be observed for the cities in the eastern region. These results clearly indicate that urban planning for cities can be separated at the top level; eastern and western regions have different levels of urbanization and therefore different needs. This study highlights 14 city pairs that are very frequently placed in the same cluster under each category. As expected, these city-pairs are from the same east/west regions of Turkey, although they are not geographically neighbors.

Also, big cities in each region often have a positive impact on their smaller neighbors. If a small city has a large neighbor, it is often placed inside the same cluster as the big cities, instead of the other small ones. While large cities attract more investment from the central government than small ones, they also help develop the smaller cities around them. The reverse is also true – if a small city doesn't have a large neighbor, it will require direct assistance from the central government to improve. This study also highlights the cities and regions that need more attention from the central government for investment planning.

Peer-review: Externally peer-reviewed.
Author Contributions: Conception/Design of Study- Ö.D., K.Ş., G.S.; Data Acquisition- Ö.D., K.Ş., G.S.; Data Analysis/Interpretation- Ö.D., K.Ş., G.S.; Drafting Manuscript- Ö.D., K.Ş., G.S.; Critical Revision of Manuscript- Ö.D., K.Ş., G.S.; Final Approval and Accountability- Ö.D., K.Ş., G.S.; Material and Technical Support- Ö.D., K.Ş., G.S.; Supervision- Ö.D., K.Ş., G.S.
Conflict of Interest: The authors have no conflict of interest to declare.
Grant Support: The authors declared that this study has received no financial support.

REFERENCES

Bulut, H. (2019). Türkiye'deki illerin yaşam endekslerine göre kümelenmesi. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 23(1), 74-82.

Yılancı, A. G. V. (2010). Bulanık Kümeleme Analizi İle Türkiye'deki İllerin Sosyoekonomik Açıdan Sınıflandırılması. Süleyman Demirel Üniversitesi *İktisadi ve İdari Bilimler Fakültesi Dergisi*, 15(3), 453-470.

İncekırık, A., & Altın, E. (2021). Türkiye'deki İllerin Ulaştırma Göstergelerine Göre Clusterleme Analizi Yöntemleriyle Sınıflandırılması. *Manisa Celal Bayar Üniversitesi Sosyal Bilimler Dergisi*, 19(3), 186-206.

Tekin, B. (2015). Temel sağlık göstergeleri açısından Türkiye'deki illerin gruplandırılması: bir Clusterleme analizi uygulaması. *Çankırı Karatekin Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 5(2), 389-416.

Atalay, M. (2019). Clusterleme analizi ile Türkiye'deki illerin turizm verileri açısından incelenmesi. *Ekonomi Maliye İşletme Dergisi*, 2(2), 103-115.

Dinçer, Ş. E. (2006). *Veri madenciliğinde K-means algoritması ve tıp alanında uygulanması* (Master's thesis, Fen Bilimleri Enstitüsü).

Silahtaroğlu, G. (2004). *Veri Madenciliğinde Clusterleme Analizi ve Öğretim Başarısının Değerlendirilmesine İlişkin Bir Uygulama* (Doctoral dissertation, Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İşletme Anabilim Dalı, İstanbul).

Mining, H. P. D. Powerful Database Support for High Performance Data Mining.

Höppner, F., Klawonn, F., Kruse, R., & Runkler, T. (1999). *Fuzzy Cluster analysis: methods for classification, data analysis and image recognition*. John Wiley & Sons.

Moertini, V. (2002). Introduction to Five DataClustering Algorithms Clustering Algorithm. *Integral*, 7(2).

Salem, S. A., & Nandi, A. K. (2005, September). New assessment criteria for Clustering algorithms. In 2005 IEEE Workshop on Machine Learning for Signal Processing (pp. 285-290). IEEE.

Kruse, R., Borgelt, C., & Nauck, D. (1999, August). Fuzzy data analysis: challenges and perspectives. In FUZZ-IEEE'99. 1999 IEEE International Fuzzy Systems. Conference Proceedings (Cat. No. 99CH36315) (Vol. 3, pp. 1211-1216). IEEE.

Meltem, I. Ş. I. K., & Çamurcu, A. Y. (2011). K-Means Ve Aşırı Küresel C-Means Algorıtmaları İle Belge Madenciliği. *Marmara Fen Bilimleri Dergisi*, 22(1), 1-18.

Ross, T. J. (2005). *Fuzzy logic with engineering applications*. John Wiley & Sons.

Hekim, M., & Orhan, U. (2011). Bulanık C-Means Clusterleme Yöntemine Çıkarımlı Yaklaşım. *İtüdergisi*/D, 10(1).

Atalay, A., & Tortum, A. (2010). Türkiye'deki illerin 1997-2006 yılları arası trafik kazalarına göre Clusterleme analizi. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 16(3), 1997-2006.

Kandemir, A. Ş. (2018). Bulanık Clusterleme Analizi İle Türkiye'deki İllerin Konaklama İstatistiklerine Göre Sınıflandırılması. *Seyahat ve Otel İşletmeciliği Dergisi*, 15(3), 657-668.

Atalay, M., & Öztürk, Ş. (2016, September). Türkiye'deki İllerin göç ve işsizlik istatistiklerine göre Clusterlenmesi. In 2nd International Congress on Applied Sciences:"Migration, Poverty and Employment", 23-25 Eylül, Konya/Turkey, Bildiriler Kitabı.

Servi, T., & Erişoğlu, Ü. (2020). Türkiye'deki Şehirlerin Sosyo-Ekonomik Gelişmişlik Düzeylerinin İstatistiksel Analizi. *Al Farabi Uluslararası Sosyal Bilimler Dergisi*, 5(2), 174-186.

Silahtaroğlu, G. (2008). *Data mining concepts and algorithms*. İstanbul: Papatya Publishing.

Silahtaroğlu, G. (2013). Veri madenciliği: Kavram ve algoritmaları. İstanbul: Papatya Publishing.

Arslan, İ. (2014). Türkiye'de Bölgesel Alanda Uygulanan İktisadi Politikalar (Yatırım Teşvikleri-İstihdam Analizi 1980-2006). *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi* , 4 (8) .