Journal of Engineering and Natural Sciences
Mühendislik ve Fen Bilimleri Dergisi

Sigma 27, 83-104, 2009

Research Article / Araştırma Makalesi

USING VISUAL DATA MINING TECHNIQUES IN CLUSTERING ANALYSIS AND AN APPLICATION

Metin VATANSEVER, Ali Hakan BÜYÜKLÜ*

Yıldız Teknik Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, Esenler-İSTANBUL

Received/Geliş: 18.11.2008 Revised/Düzeltme: 13.03.2009 Accepted/Kabul: 18.03.2009

ABSTRACT

Cluster analyses play a very important role in data mining process. A critical and important issue is to decide outliers, potential cluster structures, optimal number of clusters, to choose suitable cluster algorithms and to effectively evaluate the cluster results in cluster analysis. Various quantitative methods can overcome the problems. However, using the quantitative methods, some details can be missed that the details would be important. In the article, dealing with visual data mining techniques, with human visual system' supports, is showed how to effectively decide outliers, potential cluster structures, optimal number of clusters, choose suitable cluster algorithms and evaluate the cluster results in cluster analysis. In this way, more efficient the result of clustering can found in the field of data mining.
Keywords: Data mining, visualization techniques, visual data mining, outlier detection, cluster analysis, cluster validity, visual cluster validity.

GÖRSEL VERİ MADENCİLİĞİ TEKNİKLERİNİN KÜMELEME ANALİZLERİNDE KULLANIMI VE UYGULANMASI

ÖZET

Veri madenciliği çalışmalarında kümeleme analizleri önemli bir yer teşkil etmektedir. Kümeleme analizlerinde sapan değerlerin, potansiyel küme yapılarının, uygun küme sayılarının keşfi, uygun kümeleme algoritmalarının seçimi ve küme sonuçlarının değerlendirilmesi kritik bir öneme sahiptir. Çeşitli, sayısal yöntemlerle bu tür sorunların üstesinden gelinebilir. Ancak sayısal yöntemlerle bazı önemli olabilecek ayrıntılar gözden kaçırılabilir. Bu çalışmada görsel veri madenciliği yöntemleri yardımıyla, insan algı sisteminin de devreye girmesiyle etkili bir şekilde, sapan değerlerin, potansiyel küme yapılarının, küme sayılarının keşfedilebileceği, uygun kümeleme algoritmalarının seçilebileceği ve küme sonuçlarının değerlendirilebileceği gösterilmiştir. Bu sayede veri madenciliği alanında daha etkin küme sonuçlarına ulaşılabilinecektir.
Anahtar Sözcükler: Veri madenciliği, görselleştirme teknikleri, görsel veri madenciliği, sapan değer tespiti, kümeleme analizi, küme doğrulama, görsel küme doğrulama.

1. GİRİŞ

Veri madenciliği ile ilgili çalışmalarda kümeleme analizleri önemli bir yer tutmaktadır. Kümeleme analiziyle, gruplanmamış veriler benzerliklerine göre sınıflandırılarak araştırmacıya

*Corresponding Author/Sorumlu Yazar: e-mail/e-ileti: hbuyuklu@yildiz.edu.tr, tel: (212) 383 41 01

uygun, kullanılabilir özetleyici bilgiler verilir. Bu özelliğiyle kümeleme analizi daha sonra yapılabilecek çalışmalar için bir ara adım olma özelliği de taşıyabilir. Bu nedenle kümeleme işleminin başarıyla tamamlanması çok önemlidir.

Kümeleme analizlerinde, kaliteli küme sonuçlarının elde edilmesinde sapan değerlerin, potansiyel küme yapılarının, küme sayılarının keşfi, uygun kümeleme algoritmalarının seçilmesi ve küme sonuçlarının değerlendirilmesi kritik bir öneme sahiptir. Çeşitli sayısal yöntemlerle sapan değerler, uygun küme yapıları, küme sayıları keşfedilebilir, uygun kümeleme algoritmaları seçilebilir ve kümeleme sonuçları değerlendirilebilir; ancak bunlar yeterli olmayabilir. Görsel yöntemler sayesinde bu tür sorunların, insan algı sisteminin de devreye girmesiyle, üstesinden gelinebilir. Görsel yöntemler sayesinde sayısal yöntemlerle keşfedilemeyen ilginç örüntüler keşfedilebilir.

Geçmişte yapılan çalışmalara bakacak olursak, Andrews [1], Everitt ve Nicholls [2] çok değişkenli görsel yöntemlerle sapan değerlerin ve küme yapılarının nasıl keşfedilebileceğini incelemişlerdir. Chen ve Liu çok değişkenli, büyük veri setlerinde potansiyel küme sayılarını, küme biçim ve sınırlarını grafiksel yöntemlerle keşfetmeye çalışmışlardır [3]. Xu ve Wunsch [4], Han ve Kamber [5] ve Tan, Steinbach ve Kumar [6], uygulamanın amacına, veri tipine, verinin büyüklüğüne göre farklı kümeleme yöntemlerini incelemişlerdir. Halkidi, Batistakis ve Vazirgiannis küme algoritmalarının sonuçlarını değerlendirmek için çeşitli sayısal yöntemler sunmuşlardır [7-8]. Hathaway ve Bezdek kümeleme algoritmalarının sonuçlarını değerlendirebilmek için matris grafiklerinin görsel küme doğrulama (visual cluster validity) yöntemi olarak kullanılabileceğini göstermişlerdir [9].

Bu çalışmada görsel yöntemler yardımıyla etkili kümeleme analizleri gerçekleştirilmeye çalışılmıştır. Çalışmanın 2. bölümünde veri madenciliği, görselleştirme, görsel veri madenciliği kavramlarından bahsedilmiş, literatürde yer alan görsel veri madenciliği yöntemlerinin sınıflandırılmasına değinilerek, çalışmada kullanılan görsel yöntemler incelenmiştir. 3. bölümde çalışmanın odağını oluşturan kümeleme analizi ve küme doğrulama yöntemlerine değinilmiş, 4. bölümde 81 ildeki 918 ilçe, 20 sosyoekonomik özelliğe göre görsel veri madenciliği yöntemleri yardımıyla kümelenmiştir. Son bölüm olan 5. bölümde sonuç ve önerilerde bulunulmuştur.

2. VERİ MADENCİLİĞİ, GÖRSELLEŞTİRME VE GÖRSEL VERİ MADENCİLİĞİ

Veri madenciliği, veri ambarlarında veya diğer bilgi depolarında tutulmakta olan büyük miktardaki verinin işlenerek içindeki değerli olabilecek bilginin ortaya çıkarılması sürecidir. Veri görselleştirme, algılanabilirliği arttırmak için verinin etkileşimli ve bilgisayar desteği ile görsel olarak temsil edilmesidir [10]. Görsel veri madenciliği ise görselleştirme ile veri madenciliğini sentezleyerek veri madenciliği döngüsünü daha efektif hale getirmektir. Görsel veri madenciliği, veri tabanı bilgi keşfi sürecinin bir aşaması olarak bilgisayarla kullanıcı arasında iletişim aracı olarak görselliği kullanan bir adımdır. Görsel veri madenciliği sayesinde veriden yeni, yorumlanabilir örüntüler elde edilebilir [11].

2.1. Görsel Veri Madenciliği Yöntemlerinin Sınıflandırılması

Veri madenciliği çalışmalarında çok fazla sayıda kayıt ve çok fazla sayıda boyuttan oluşan veri yığınlarıyla uğraşılır. İnsanların algılama sistemleri de yalnızca 3 boyutla sınırlı olduğu için daha fazla boyut içeren veriler insan algı sisteminin dışına çıkmaktadır. Bundan dolayı, veri görselleştirme yöntemleri çok boyutlu veriyi 2 veya 3 boyuta indirgeyerek görselleştirmeli, diğer taraftan da veriler arasındaki ilişkiyi muhafaza edebilmelidir. Kutu, çizgi, histogram gibi bilinen çeşitli görselleştirme yöntemleri verileri görselleştirmek için kullanılabilir; ancak bu teknikler az sayıda boyutu gösterebilmektedirler. Çok boyutlu verileri görselleştirmek içinse çeşitli görselleştirme yöntemleri geliştirilmiştir[12].

Görselleştirme yöntemlerinin, bu alanda yapılan çalışmalar incelendiğinde, 5 temel sınıfa ayrılabileceği görülmektedir [6][12-13]. Bu beş sınıf:

- standart 2 ve 3 boyutlu gösterimler (Kutu grafikleri, Histogram,…),
- geometrik olarak dönüştürülmüş gösterimler, (Andrews eğrileri, PolyViz grafikleri,…),
- simgesel gösterimler, (Chernoff yüzleri, Star,….),
- yoğun piksel gösterimler, (Matris grafikleri,…..),
- istiflenmiş gösterimlerdir (Dünyalar içinde dünyalar (worlds within worlds), Treemap, …).

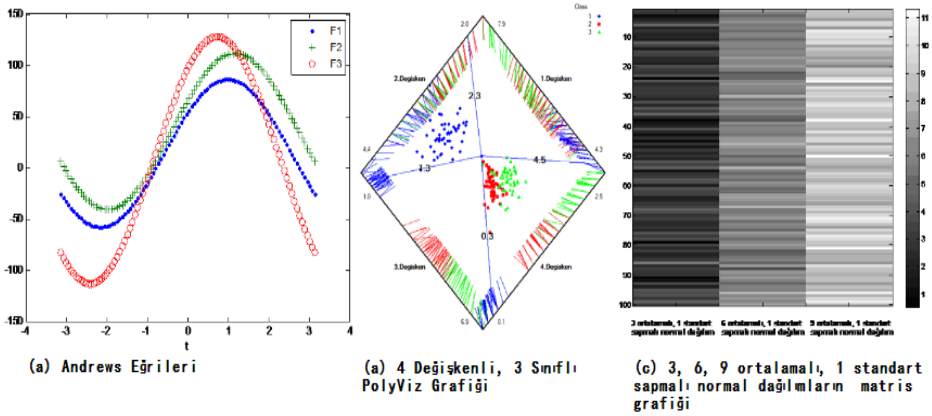
Çalışmamızda küme yapılarını ve sapan değerleri göstermelerine olan duyarlılıklarından dolayı Andrews Eğrileri, PoliViz ve Matris Grafikleri kullanılmıştır.

2.1.1. Andrews Eğrileri (Andrews Curves)

Andrews eğrileri çok boyutlu verileri görselleştirmek için bir yöntem olarak geliştirilmiştir. Andrews eğrilerinde, gözlem değerleri eşitlik (1) deki fonksiyon kalıbı kullanılarak dönüştürülürler. Dönüşen bu değerlerin daha sonra çizgi grafikleri çizilerek Andrews eğrilerine ulaşılır [14].

$$f_x(t) = x_1/\sqrt{2} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \ldots, \qquad -\pi < t < +\pi \tag{1}$$

Burada x1, x2, … verilerimizin değişkenleridir. Şekil 1 (a)' da $n_1 = (20,60,40)$, $n_2 = (50,70,30)$ ve $n_3 = (10,80,90)$ veri noktaları için çizilen Andrews eğrileri bulunmaktadır.



Şekil 1. Andrews eğrileri, PoliViz ve matris grafikleri

Andrews eğrileri orijinal veri setinin uzaklıklarını içersinde barındırırlar. Andrews eğrilerinin kullandığı fonksiyon kalıplarından elde edilen eğrilerin birbirine yakın olması gözlem değerlerinin birbirine yakın olduğunu, birbirine uzak olması gözlem değerlerinin de birbirine uzak olduğunu gösterir. Buradan hareketle Andrews eğrileri verilerin küme yapılarının anlaşılmasında da, sapan değerlerin tespitinde de kullanılabilirler [1][14].

Andrews eğrilerine getirilen en büyük eleştirilerden bir tanesi de çizilen şeklin biçiminin değişkenlerin sıralarına olan bağımlılığıdır. Yani değişken sıraları değiştikçe Andrews eğrilerinin şekilleri değişecektir. Andrews eğrilerinde dönüştürme işlemi için kullanılan eşitlik (1)' deki seride ilk sıraya yerleşen değişken grafiğimizin üzerinde en büyük ağırlığa sahip olan değişkendir. Yani değişkenlerin sıraları değiştikçe, ilk sırada olan değişkenin, grafiğin ağırlığına

olan etkisi fazla olacak şekilde grafiğimizin biçimi değişecektir. Sonuç olarak değişken sayısının permutasyonu kadar farklı sayıda grafik çizilecektir [14]. Temel bileşenler analiziyle boyut indirgemesi yapılarak Andrews eğrilerinin bu dezavantajının üstesinden gelinebilir. Bu sayede en büyük varyans açıklayıcılık oranı sahip değişken şeklimizde en büyük etkiye sahip olacaktır.

2.1.2. PolyViz (Geliştirilmiş RadViz)

PolyViz çok değişkenli bir veri görselleştirme yöntemidir. Burada ki amaç serpilme grafiklerindeki gibi eksenlerde yer alan değişkenler arasındaki serpilmeyi çizmektir. Ancak serpilme grafiğinden farklı olarak PolyViz grafiğinde değişkenleri temsil eden eksenler bir nokta etrafında çokgen oluştururlar. Veri değerlerimiz, değişken eksenleri boyunca kısa çizgilerle temsil edilerek, değişkenlerde kesişen değerler grafikte nokta olarak konumlandırılır. PolyViz yönteminde, grafikteki bütün değişkenlere eşit önem verilebilmesi adına, veri setimizde yer alan değişkenlerin standartlaştırılması gerekmektedir [13].

PolyViz grafiğinde Andrews eğrilerinde olduğu gibi küme yapıları, sapan değerler gözlenebilmektedir. Ayrıca PolyViz grafiğinde veri setinin her bir değişkeni için veri dağılımları hakkında da bilgiler elde edilebilmektedir. Veri dağılımları değişken eksenleri boyunca çizilen kısa çizgilerin dağılımlarından anlaşılmaktadır. Çünkü bu kısa çizgilerin her biri veri setindeki veri birimlerine karşılık gelmektedir [13]. Şekil 1 (b)' de 3 sınıf, 4 boyuttan oluşan bir PolyViz grafik örneği bulunmaktadır.

PolyViz grafik yönteminde, daire içersinde dağılan verilerin şekilleri, daire etrafında dizilen değişkenlerin sıralarına göre farklılık göstermektedir. Mesela m değişken için dairenin etrafında $(m-1)!/2$ tane farklı şekilde değişken sıralanabilir. Her farklı sıralamaya göre PolyViz grafiğinde verilerimizin dağılımları farklılık göstermektedir.

2.1.3. Matris Grafikleri (Matrix Plots)

Matris grafiği değişkenler arasındaki ikili ilişkileri kullanıcıya göstermeye yarayan bir çeşit saçılma çizgisidir. Matris grafiğinin ana fikri, veri matrisinde bulunan verilerin büyüklüğünü matris grafiğinde renkli karelerle temsil etmektir. n satır p sütundan oluşan veri matrisi için matris grafiği nxp şeklinde renkli karelerden oluşur. Veri matrisinde bulunan değerler matris grafiğinde büyüklüklerine göre renklendirilirler. Bu teknik çok boyutlu büyük veri setlerinin görselleştirilmesi için elverişlidir. Matris grafiğinde, grafikteki bütün değişkenlere eşit önem verilebilmesi adına, veri setimizde yer alan değişkenlerin standartlaştırılması gerekmektedir. [6][15]. Şekil 1 (c)' de 100 birimden oluşan 3, 6 ve 9 ortalamalı, 1 standart sapmalı normal dağılımların matris grafiği bulunmaktadır.

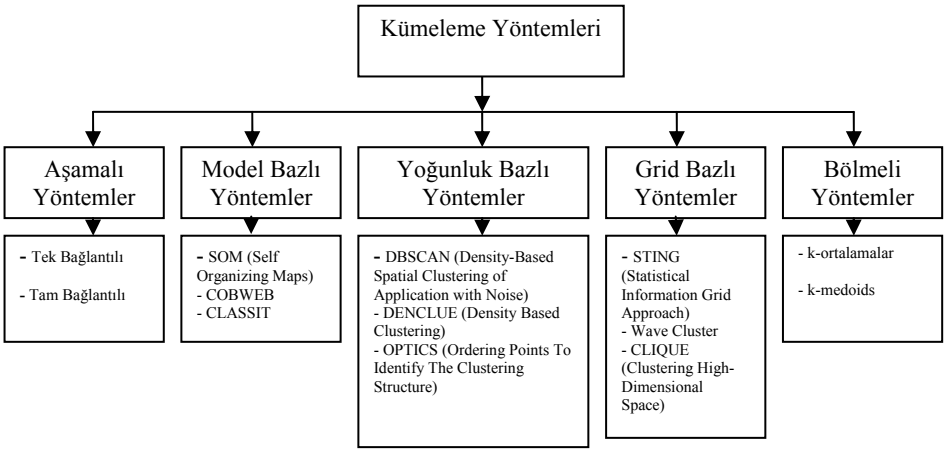
3. KÜMELEME ANALİZİ

"Kümeleme analizi X veri matrisinde yer alan ve doğal grupları kesin olarak bilinmeyen birimleri, değişkenleri ya da birim ve değişkenleri birbirleri ile benzer olan alt kümelere ayırmaya yardımcı olan yöntemler topluluğudur [16]." Kümeleme analizi, birimleri değişkenler arası benzerlik (similarity) ya da uzaklıklara (dissimilarity) dayalı olarak hesaplanan bazı ölçülerden yararlanarak homojen gruplar oluşturmaya çalışır [4][14][16]. Kümeleme analizi sonucunda kümeleri oluşturan elemanlar birbirine benzerlik, başka kümelerin elemanlarından farklılık gösterirler. Kümeleme işlemi başarılı olursa, bir geometrik çizim yapıldığında birimler küme içerisinde birbirilerine çok yakın, kümeler ise birbirilerinden uzak olacaktır.

Veri tabanlarında toplanan veri miktarının artmasıyla orantılı olarak, kümeleme analizi son zamanlarda, özellikle veri madenciliği araştırmalarında genişçe yer bulur hale gelmiştir. Kümeleme analizi ayrıca istatistik, biyoloji, psikoloji, tıp, arkeoloji, sosyoloji, makine öğrenim ve örüntü tanıma gibi daha pek çok alanda kullanım olanağı bulmaktadır [16, 17].

3.1. Kümeleme Yöntemlerinin Sınıflandırılması

Veri madenciliğinde, uygulamanın amacına, veri tipine, verinin büyüklüğüne göre farklı kümeleme yöntemleri bulunmaktadır. Değişik kaynaklarda farklı kümeleme yöntemleri farklı şekillerde sınıflandırılmaktadır. Veri madenciliği ile ilgili kaynaklarda kümeleme yöntemleri aşağıdaki gibi sınıflandırılabilmektedir [5].



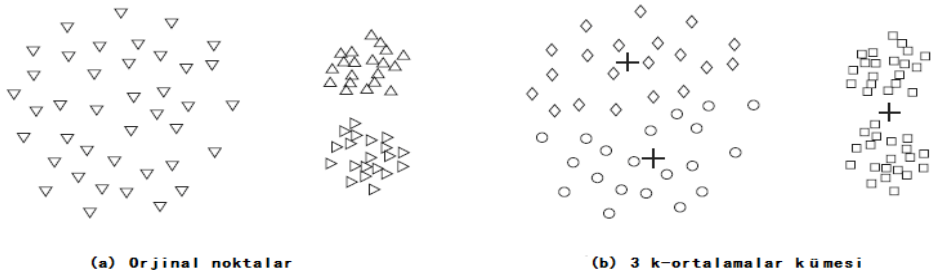
Şekil 3. Kümeleme yöntemleri

Bu çalışmada, kümeleme analizlerinde tek bağlantılı (single link), tam bağlantılı (complete link), kendinden düzenlenen haritalar (SOM-Self Organizing Maps) ve k-ortalamalar (k-means) yöntemleri kullanılacaktır. Tek ve tam bağlantılı hiyerarşik, k-ortalamalar kümeleme yöntemleri istatistiksel yöntemler olup, SOM kümeleme yöntemi ise yapay sinir ağı (neural network) yöntemidir [5].

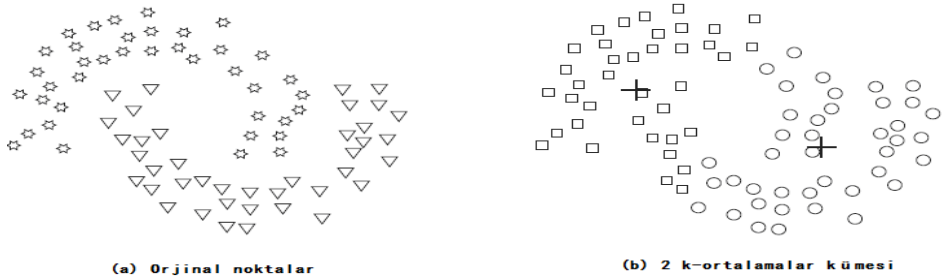
3.2. Kümelemede Farklı Tipte İdeal Olmayan Yapılar

K-ortalamalar ve SOM kümeleme yöntemleri farklı küme yapıları için birtakım kısıtlara sahiptir. Özellikle doğal küme yapılarının küresel biçimde olmaması, oldukça farklı küme hacimlerine ve yoğunluklarına sahip olması k-ortalamalar ve SOM yöntemlerinin başarısız sonuçlar vermesine neden olabilmektedir. Bu durumu örneklendirmek için Şekil 4, 5, 6'ya bakabiliriz. Şekil 4'deki kümelerden bir tanesinin hacmi diğer iki kümeye göre oldukça büyüktür. Küme hacminin büyük olmasından dolayı k-ortalamalar kümeleme yöntemi doğal küme yapılarını bulmada başarısız olmuştur. Şekil 4'de doğal küme yapısı küçük olan bir küme, k-ortalamalar yöntemi sonucunda büyük bir küme olarak bulunmuştur. Şekil 5'deki kümelerden iki tanesi diğer büyük hacimli kümeye göre oldukça büyük bir yoğunluğa sahip olduğu için k-ortalamalar yöntemi doğal küme yapılarını bulmada başarısız olmuştur. Son olarak, Şekil 6'da küme yapıları küresel olmadığı için k-ortalamalar yöntemi doğal küme yapılarını bulmada başarısızdır [6].

Şekil 4. Farklı hacimli kümeler için k-ortalamalar [6]

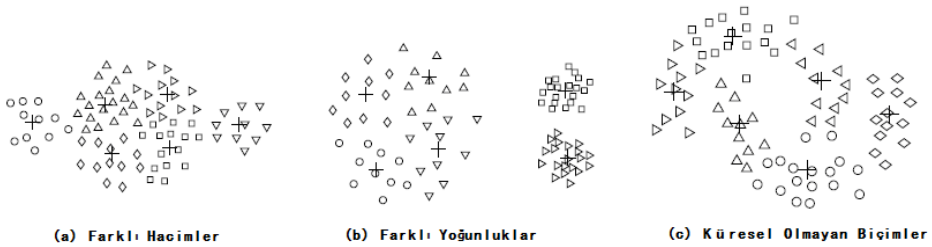


Şekil 5. Farklı yoğunluklu kümeler için k-ortalamalar [6]



Şekil 6. Küresel biçimde olmayan kümeler için k-ortalamalar [6]

Doğal küme yapılarını bulmak için kullandığımız k-ortalamalar kümeleme yöntemi yukarıda bahsedilen üç kısıttan dolayı başarısız sonuçlar verebilmektedir. Bu kısıtların yarattığı sorunlar, doğal küme yapılarının birden fazla alt kümeye ayrılmasıyla ortadan kaldırılabilir. Şekil 4, Şekil 5 ve Şekil 6 'da kullanılan veriler için çizilen Şekil 7 'de, iki ve üç doğal küme yapısından altı tane alt küme elde edilmiştir. Bu sayede k-ortalamalar yöntemiyle doğal küme yapılarının yanlış kümelenmesinin önüne geçilmiştir [6].



Şekil 7. Doğal kümelerin alt kümeleri için k-ortalamalar [6]

3.3. Küme Doğrulama (Cluster Validity)

Kümeleme analizlerinde, sonuç kümelemelerinin değerlendirilmesi kümeleme modeli geliştirme işleminin ayrılmaz bir parçasıdır. Çünkü bir veri kümesinde küme yapısı olmasa bile kümeleme algoritmaları bu veri seti içersinde istenilen sayıda küme bulacaktır. Ancak elimizdeki veri kümesinde herhangi bir küme yapısı bulunmayabilir. Bundan dolayı kümeleme algoritmalarının sonuçlarının değerlendirilmesine yönelik çeşitli sayısal küme doğrulama (cluster validity) yöntemleri geliştirilmiştir. Bu sayede kümeleme çalışmalarında, küme kalitesi ve uygun küme sayısı belirlenerek kümeleme işlemleri başarıyla tamamlanabilir [18].

Kümeleme algoritması tarafından üretilen sonuçların değerlendirilmesine yönelik içsel (external), dışsal (internal) ve göreli (relative) olmak üzere, 3 farklı kritere göre hesaplanmış, çeşitli küme doğrulama (cluster validity) yöntemleri geliştirilmiştir. Bu çalışmada göreli kritere göre hesaplanan Silhouette, Davies-Bouldin, Dunn, Calinski-Harabasz, Krzanowski ve Lai ve Hartigan küme doğrulama (cluster validity) endeksleri kullanılmıştır [6][8][17][19]. Bu endekslerle uygun küme sayısı tahmin edilmeye çalışılmıştır.

Küme doğrulama endekslerinin ideal olmayan küme biçimleri için bazı dezavantajları bulunmaktadır. Bu teknikler küme sonuçlarını değerlendirirken, kümeleri temsil edebilecek noktaları referans noktaları alarak, referans noktaları arasındaki uzaklıkları ve referans noktalarından hareketle varyans gibi parametreleri hesaplarlar. İdeal olmayan küme biçimlerini temsil edebilecek referans noktalarının seçilmesindeki zorluklardan dolayı bu tür endekslerle ideal olmayan küme biçimlerinin değerlendirilmesi her zaman doğru olmayabilir [17].

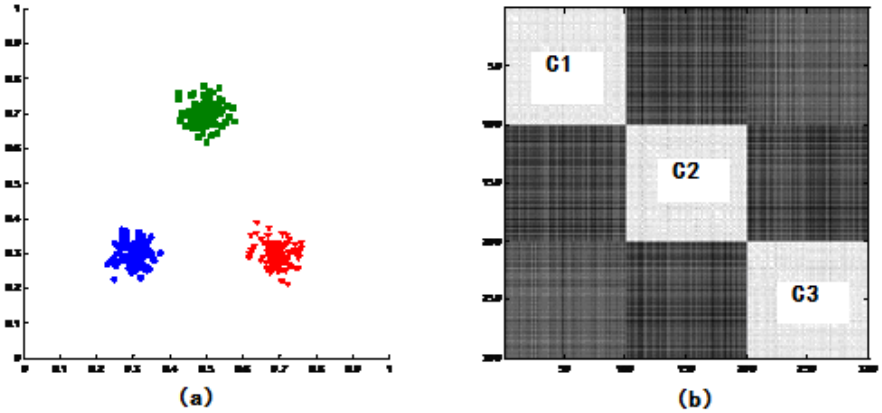
3.3.1. Görsel Küme Doğrulama (Visual Cluster Validity)

Küme yapılarını görmek adına matris grafikleri kullanılabilir. Bu sayede kendi içinde homojen kendi aralarında heterojen guruplar görsel bir şekilde gözlemlenebilir. Bu özelliğinden dolayı matris grafikleri kümeleme analizlerinden elde edilen sonuçların doğruluğunu göstermede de kullanılmaktadırlar. Matris grafikleri çok büyük veri setlerinde kullanılabilmekle birlikte veri setlerindeki küme sayıları hakkında genel bir fikir edinmeyi ve buna uygun iyileştirilmelerin yapılmasına olanak tanır.

Matris grafiklerinde görselleştirme, kümeleme algoritmalarının ürettiği sonuçlar kullanılarak, benzerlik matrisi küme etiketlerine göre sıralanarak gerçekleştirilir. Daha sonra sıralanmış benzerlik matrisindeki her bir benzerlik değeri gri ölçekte bir renge karşılık gelecek şekilde görselleştirilir. Beyaz renkler maksimum benzerliği, siyah renkler minimum benzerliği gösterir [10].

Şekil 8 (a) 'da bulunan 3 iyi ayrılmış küme noktalarını temsil eden matris grafiği Şekil 8 (b) 'deki gibidir. Şekil 8 (b) 'de C1, C2 ve C3 ile gösterilen alanlar üç farklı kümeyi temsil ederler. Birincil köşegen üzerinde bulunmayan ve C1, C2 ve C3 dışında kalan dikdörtgensel alanlar ise kümeler arası ilişkiyi gösterirler. Teoride, eğer iyi ayrılmış kümelerimiz varsa benzerlik matrisi kabaca blok diyagonal olacaktır. Değilse, benzerlik matrisindeki örüntüler

kümeler arasındaki ilişkiyi ortaya çıkaracaktır. Tüm bunlar uzaklık matrisine de uygulanabilir ancak benzerlik matrisinin büyük miktarlardaki verilerle uğraşan veri madenciliği çalışmalarında daha kaliteli sonuçlar verdiği bilinmektedir [10].



Şekil 8. Üç küme için benzerlik matrisi

Matris grafikleri uygun küme sayısının tespitine yönelik pratik bir yöntem sunmaktadır. Öncelikle veri seti çok sayıda kümeye ayrılarak matris grafikleri yardımıyla sağlanan görselleştirme ile birbirine benzer kümeler birleştirilerek uygun küme sayısı belirlenebilir [10].

4. UYGULAMA

4.1. Açıklama

Sosyoekonomik gelişme, gerek zaman, gerek mekân açısından farklılıklar göstermekte ve sürekli değişen bir olgu olarak kabul edilmektedir. Dolayısıyla ülkelerin gelişme çizgileri zamanla değiştiği gibi, yörelerin mevcut gelişme düzeylerinin de farklılıklar gösterdiği bilinmektedir. Çalışmada 81 ildeki 918 ilçe gelişmişlik düzeylerine göre görsel veri madenciliği yöntemleri yardımıyla kümelenmiştir.

4.2. Analizde Kullanılan Değişkenler

Ülkemizde, iller ve özellikle ilçeler itibariyle yapılacak ekonomik ve sosyal araştırmalar için ihtiyaç duyulan verilerin yeterli ölçüde ve sistematik bir şekilde temin edilmesinin ortaya koyduğu zorluklar nedeniyle, bu çalışmada kullanılan değişkenler Türkiye'deki ilçelerin gelişmişlik düzeyleri belirlenmesi için yayınlanmakta olan ve kolay ulaşılabilen 20 adet değişkenden derlenmiştir. Bu çalışmada kullanılan sosyoekonomik nitelikteki değişkenler ve bu değişkenlerin analiz aşamasında kullanılan isimleri aşağıda sıralanmaktadır [20-22]:

X1 : Toplam nüfusun yıllık ortalama artış hızı (%) (1990-2000)
X2 : Şehirleşme oranı (%) (2000)
X3 : Toplam nüfus yoğunluğu (kişi/km^2) (2000)
X4 : Ücretli çalışan kadınların toplam istihdama oranı (%) (2000)
X5 : İşsizlik (%) (2000)
X6 : Erkek okuryazar oranı (%) (2000)
X7 : Kadın okuryazar oranı (%) (2000)

X8 : Erkek yüksek okul bitirenlerin oranı (%) (2000)
X9 : Kadın yüksek okul bitirenlerin oranı (%) (2000)
X10 : Tarım kesiminde çalışanların toplam istihdama oranı (%) (2000)
X11 : İmalat sanayinde çalışanların toplam istihdama oranı (%) (2000)
X12 : İnşaat kesiminde çalışanların toplam istihdama oranı (%) (2000)
X13 : Toplam perakende ticarette çalışanların toplam istihdama oranı (%) (2000)
X14 : Ulaştırma depolamada çalışanların toplam istihdama oranı (%) (2000)
X15 : Mali kurumlarda çalışanların toplam istihdama oranı (%) (2000)
X16 : İlmi ve teknik mesleğe sahip olan kişilerin toplam istihdama oranı (%) (2000)
X17 : İşverenlerin toplam istihdama oranı (%) (2000)
X18 : İdari personel ve benzeri çalışanların toplam istihdama oranı (%) 2000
X19 : Fert başına düşen gelir (GSMH TL) (1996)
X20 : 100000 kişiye düşen banka şube sayısı (2000)

İlçelere ait veriler TÜİK tarafından yayınlanan 2000 yılı nüfus sayım sonuçlarından alınmıştır. 2000 yılı GSMH değerleri ilçe bazında bulunamadığı için çalışmada 1996 yılı GSMH değerleri kullanılmıştır. Bu sebepten dolayı çalışmada, 1996 yılında ilçe olmayan Kocaeli-Derince, Osmaniye-Hasanbeyli, Osmaniye-Sumbas, Osmaniye-Toprakkale, Düzce-Kaynaşlı ilçeleri kapsam dışı bırakılmıştır.

Kümeleme analizi hesaplama çalışmalarında MATLAB R2007a [23], SPSS 15 Evaluation [24] ve Orange [25] programları kullanılmıştır. Değişkenlerin birimlerinin farklı olması nedeniyle, veriler z dönüşümü ile standardize hale getirilmiş ve kümeleme analizi çalışmalarında standart veri matrisi kullanılmıştır.

4.3. Sapan Değer Analizi

Sapan değerler, veri madenciliği sürecinin analiz aşamasında regresyon, kümeleme analizi gibi uygulamalarda sorunlara neden olurlar. Bu nedenle sapan değerlerin veri setinden arındırılması gerekmektedir.

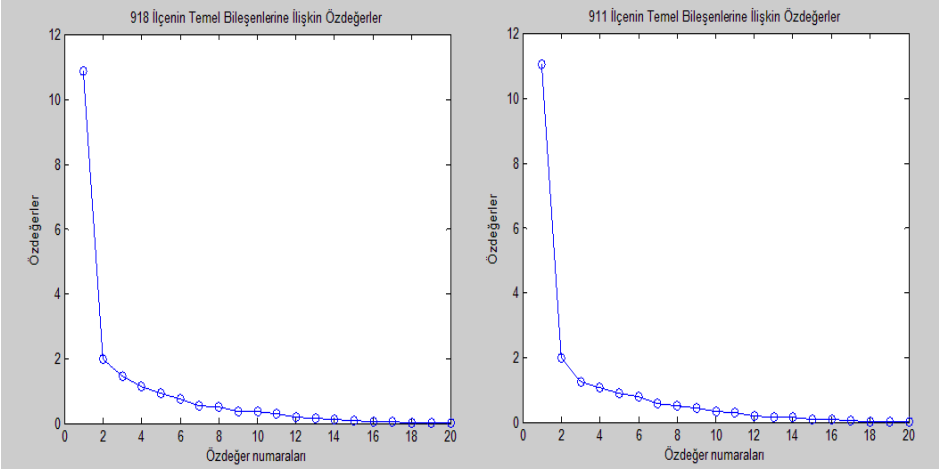
918 ilçenin 20 değişkeninden oluşan veri setinde gerek değişken sayısının fazla olması, gerekse veri birimlerinin fazla olmasından dolayı sapan değerleri ayıklamak oldukça zor bir iştir. Bunun için görselleştirme teknikleri kullanılarak sapan değerler görsel bir şekilde tespit edilip, veri setinden ayıklanabilir.

Çalışmada, değişken bazında değil de bütünü kavrayacak şekilde sapan değerleri tespit etme özelliğinden dolayı Andrews eğrileri sapan değerlerin tespitinde kullanılmıştır. Bunun için temel bileşenler analiziyle boyut indirgemesi yapılmış ve birbirinden bağımsız bileşenler elde edilmiştir. Daha sonra bu bileşenler Andrews eğrilerinde kullanılarak aşırı değerler görsel bir şekilde tespit edilmiştir. Temel bileşenler analiziyle elde edilen bileşenlerin sırası veri setinin toplam değişkenliğini açıklama oranlarıyla orantılıdır. Bu sayede Andrews eğrileri, en büyük açıklayıcılık oranına sahip bileşen en büyük etkiye sahip olacak şekilde çizilir.

918 ilçenin 20 değişkeninden oluşan veri setinin temel bileşenler analizinden elde edilen özdeğerler ve toplam varyans açıklama oranları Çizelge 1 de, bileşenlerin toplam varyansı açıklama oranları için çizilen yamaç grafiği Şekil 9 ’da, hem bütün veriyi, hem de sapan değerler çıkarılmış şekliyle verilmiştir.

Çizelge 1. Temel bileşenlere ilişkin özdeğerler

918 İlçenin Temel Bileşenlerine İlişkin Özdeğerler			911 İlçenin Temel Bileşenlerine İlişkin Özdeğerler		
Bileşen No	Özdeğerler	Toplam Varyans Açıklama Oranları	Bileşen No	Özdeğerler	Toplam Varyans Açıklama Oranları
1	10.8754	54.3771	1	11.0202	55.1008
2	1.9698	64.2261	2	1.9787	64.9943
3	1.4605	71.5287	3	1.2529	71.2588
4	1.1443	77.25	4	1.0743	76.6304
5	0.9426	81.9628	5	0.8985	81.1229
6	0.7702	85.8138	6	0.79	85.073
7	0.5423	88.5251	7	0.5701	87.9233
8	0.5118	91.0843	8	0.5064	90.4553
9	0.3727	92.9476	9	0.4438	92.6743
10	0.3616	94.7557	10	0.3435	94.3916
11	0.2869	96.1905	11	0.2949	95.8659
12	0.2065	97.2229	12	0.2042	96.8868
13	0.1418	97.9319	13	0.1593	97.6836
14	0.1228	98.546	14	0.1449	98.408
15	0.0866	98.9789	15	0.0927	98.8716
16	0.0697	99.3272	16	0.0801	99.2721
17	0.0601	99.6275	17	0.0602	99.573
18	0.0312	99.7835	18	0.0331	99.7385
19	0.0252	99.9094	19	0.0329	99.903
20	0.0181	100	20	0.0194	100

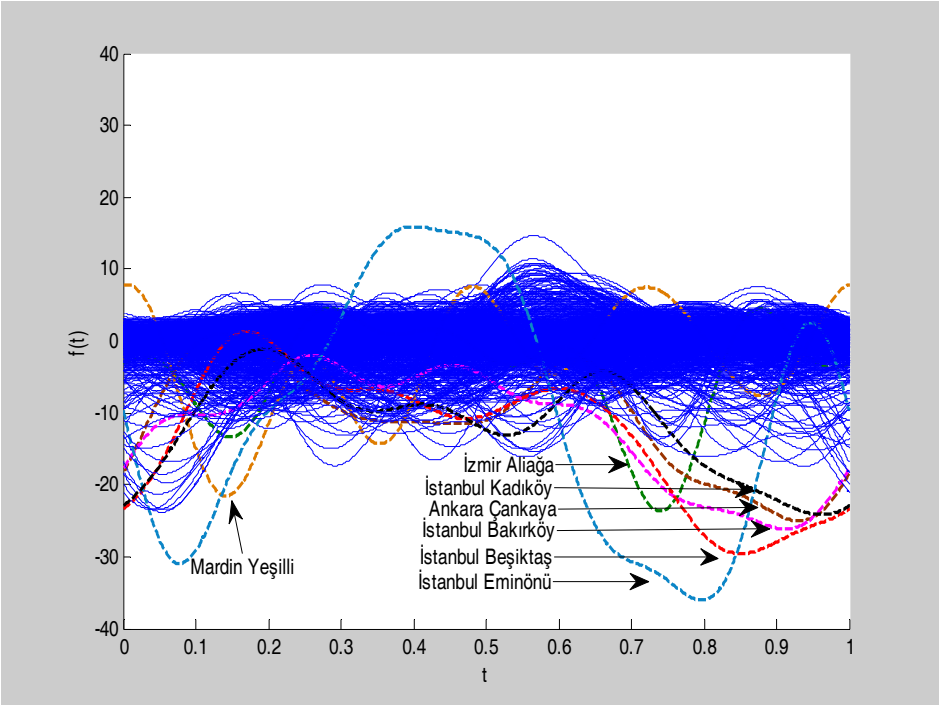


Şekil 9. Özdeğerlerin yamaç eğim grafiği

Bileşenlere ilişkin toplam varyans açıklama oranlarına ve yamaç eğim grafiklerine bakıldığında toplam varyansın yaklaşık % 95 'ini açıklayan ilk 10 temel bileşenle çalışmanın uygun olacağı düşünülmüştür. Şekil 9 'dan da gözüktüğü gibi 10. temel bileşenden sonra yamaç eğim grafiğinin eğiminin sabitleştiği gözükmektedir. İlk 10 temel bileşen kullanılarak çizilen Andrews eğrileri grafiği Şekil 10 'da ki gibidir.[†] Şekil 10 'a göre sapan değerler Eminönü, Beşiktaş, Bakırköy, Çankaya, Kadıköy, Aliağa ve Yeşilli olarak gözükmektedir. Dolayısıyla bu çalışmada Eminönü, Beşiktaş, Bakırköy, Çankaya, Kadıköy, Aliağa ve Yeşilli ilçeleri sapan değer olarak düşünülmüştür. Şüphesiz sapan değerleri çıkartılan veri seti için tekrar Andrews eğrileri grafiği çizildiğinde Y ekseninde bulunan ölçek hassasiyetinin değişmesine göre yeni sapan

[†] MATLAB R2007a Programı Andrews eğrilerini çizerken $0 < t < +1$ aralığını kullanır.

değerler tespit edilebilir. Ancak bu ikinci Andrews eğrileri grafiğinde bulunacak olan sapan değerler ilk Andrews eğrilerinde bulunan sapan değerler kadar belirgin olmayacaktır. Unutulmamalıdır ki veri görselleştirme teknikleri insan algılama yeteneklerini ve insanlar arası yorumlama farklılıklarını dikkate alarak analiz gerçekleştirilmesine olanak sağlar. Görselleştirme teknikleri ile diğer yöntemlerle fark edilmesi daha zor olan bilgiye erişilmesi ve bilginin yorumlanması kolaylaşmaktadır. Ancak grafiksel tekniklerin matematiksel sonuçlar vermemesi gibi bir dezavantajı da bulunmaktadır.



Şekil 10. 10 temel bileşen için Andrews eğrileri

4.4. Kümeleme Analizi

Bu alt bölümde tespit edilen sapan değerlerin veri setinden ayıklanmasıyla elde edilen yeni veri setinin kümelenmesi gerçekleştirilecektir.

4.4.1. Temel Bileşenler Analizi

Uygulamada ilk olarak değişkenler arasındaki bağımlılık yapısının ortadan kaldırılması ve veri boyutunun indirgenerek aynı şeyi ifade eden değişkenlerin birleştirilmesi amacıyla verilere temel bileşenler analizi uygulanmıştır. Böylece ilçelerin kümelenmesi korelasyonsuz daha az değişkenle gerçekleştirilebilecektir [16].

Değişkenler arasında anlamlı ilişkilerin olup olmadığını görmek için R korelasyon matrisini incelemek ve verilere temel bileşenler analizi uygulamanın gerekli olup olmadığını görmek, eğer değişkenler arasında ilişki varsa bunların önemli olup olmadığını anlamak için veri

setine küresellik testi uygulanmalıdır [26]. SPSS 15 Evaluation programıyla hesaplanan küresellik testi sonuçları Çizelge 2'de verilmiştir.

Çizelge 2. Küresellik test sonuçları

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,883
Bartlett's Test of Sphericity	Approx. Chi-Square	22692,117
	df	190
	Sig.	,000

Küresellik testi için;

H0: R=I (İlişki matrisi ile birim matris arasında fark yoktur. Değişkenler arasındaki ilişkiler önemsizdir.)

H1: R≠I (İlişki matrisi ile birim matris arasında fark vardır. Değişkenler arasındaki ilişkiler önemlidir [26].

Olasılık değeri olan Sig. değerine bakıldığında; 0.000<0.05 olduğundan hipotez reddedilir. Bu nedenle ilişki matrisi ile birim matris arasında fark olduğu diğer bir ifade ile değişkenler arasındaki ilişkilerin önemli olduğu 0.95 olasılıkla söylenebilir. Bu da temel bileşenler analizi uygulanmasının gerekliliğini ortaya koymaktadır.

Sapan değerleri çıkartılan 911 ilçe yeni veri setinin temel bileşenler analizinden elde edilen özdeğerler ve toplam varyans açıklama oranları Çizelge 1 'de, bileşenlerin toplam varyansı açıklama oranları için çizilen yamaç grafiği Şekil 9 'da verilmiştir.

Bileşenlere ilişkin toplam varyans açıklama oranlarına ve yamaç eğim grafiklerine bakıldığında toplam varyansın yaklaşık % 94 'ünü açıklayan ilk 10 temel bileşenle çalışmanın uygun olacağı düşünülmüştür. Bundan sonraki analiz aşamalarında aşırı değerlerden arındırılmış veri seti için elde edilen ilk 10 temel bileşenle çalışılmaya devam edilecektir. Bu sayede veri setleri korelasyondan arındırılmış ve daha az değişkenle kümeleme analizlerinin yapılabilmesi mümkün olacaktır. Dolayısıyla, kümeleme analizlerinde kritik öneme sahip olan iki nokta arasındaki uzaklıkların daha az anlamlı hale gelmesi değişken sayısı indirilerek önlenmiş olacaktır.

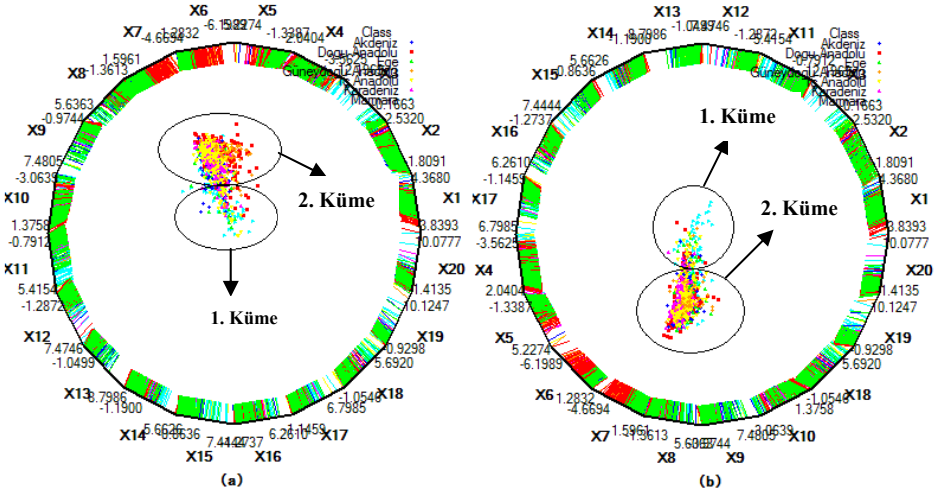
4.4.2. Uygun Küme Sayısının ve Algoritmasının Belirlenmesi

Bu alt bölümde öncelikle veri setimizdeki doğal küme yapılarının varlığı ve yapıları görsel yöntemlerle keşfedilmeye çalışılmıştır. Daha sonra sayısal yöntemlerle doğru küme sayılarının kestirilmesine gidilmiş, sayısal ve görsel yöntemlerin birlikte kullanılmasıyla uygun kümeleme algoritması seçilmiş ve kümeleme işlemlerinin kalitesi arttırılmıştır.

Şekil 12 'de ilçe veri seti için Orange programıyla çizilen değişkenleri farklı sırlanmış PolyViz grafikleri bulunmaktadır. Şekil 12 (a) ve Şekil 12 (b) 'de veri setimizde 2 kümenin bulunabileceğine yönelik ipuçları elde dilmiştir. Şüphesiz değişkenlerin farklı sıralanmasına göre PolyViz grafiğinde verilerin dağılımları ve biçimleri değişecektir. Bu durumda PolyViz grafikleri küme yapıları tespitinde sadece yol gösterici olabilmektedir.

Değişkenleri farklı sıralanmış PolyViz grafiğinde veri setimizde 2 kümenin bulunabileceğine yönelik ipuçları elde edilmektedir. PolyViz grafiklerine göre kabaca Marmara bölgesi ilçelerinin çoğunluğu bir kümeyi, geri kalan ilçelerinse diğer bir kümeyi oluşturduğu gözlenmiştir.

PolyViz grafiklerinde dikkati çeken diğer bir nokta da küme potansiyeli taşıyan veri dağılımlarının küresel olmaması, yoğunluk ve hacimlerinin farklı olmasıdır. Bu da daha sonra yapılacak kümeleme çalışmalarını olumsuz yönde etkileyecektir.



Şekil 12. Farklı sıralanmış değişkenler içi PolyViz grafiği

Doğal gurupları bilinmeyen 911 ilçenin tek bağlantılı hiyerarşik, tam bağlantılı hiyerarşik, k-ortalamalar ve SOM kümeleme yöntemlerine göre 10 küme için Silhouette (S), Davies-Bouldin (DB), Dunn (D), Calinski ve Harabasz (CH), Krzanowski ve Lai (KL) ve Hartigan (H) küme doğrulama endeksleriyle hesaplanmış uygun küme sayıları Çizelge 3 'de verilmiştir. Kümeleme yöntemleri ve küme doğrulama endeksleri Cluster Validity Analysis Platform' unda hesaplanmıştır [27].

Çizelge 3 'deki çeşitli kümeleme yöntemleri için hesaplanan endeksler, 911 ilçe veri setinin doğal küme sayılarının tespitine yönelik ortak sonuçlar vermemektedir. Endeksler çoğunluk itibariyle ilçe veri setinde 2 küme bulunabileceğini söylemektedir.

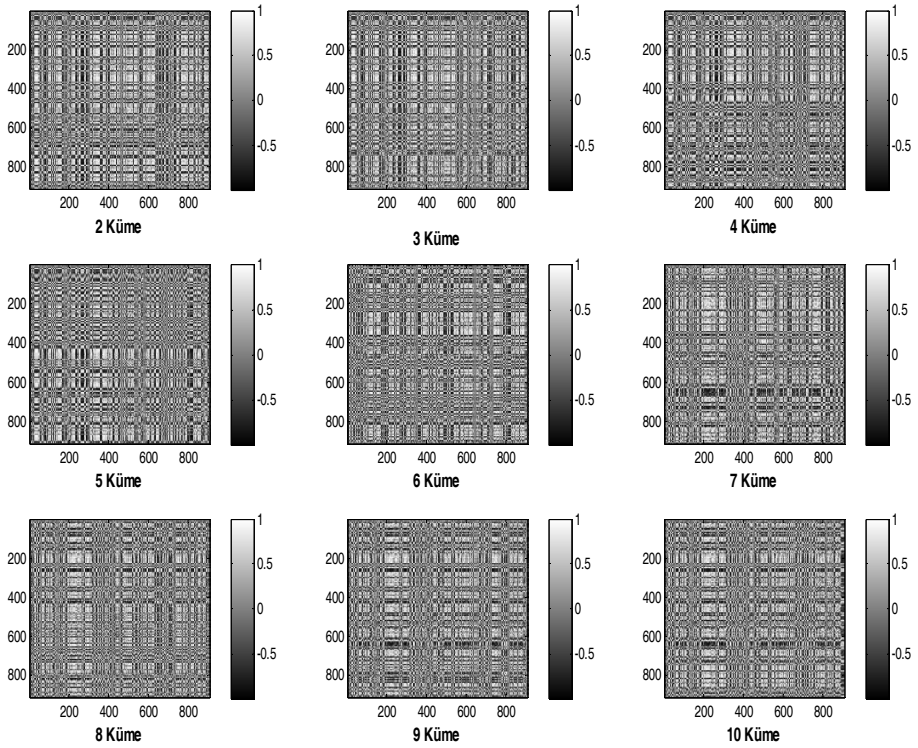
Çizelge 3. İki veri seti için küme doğrulama endeksleri

Kümeleme Yöntemleri		S	DB	D	CH	KL	H
İlçe Veri	Tek Bağlantılı	2	10	2	6	6	6
	Tam Bağlantılı	2	2	2	3	3	3
	K-Ortalamalar	2	2	2	2	2	2
	SOM	2	2	2	2	2	2

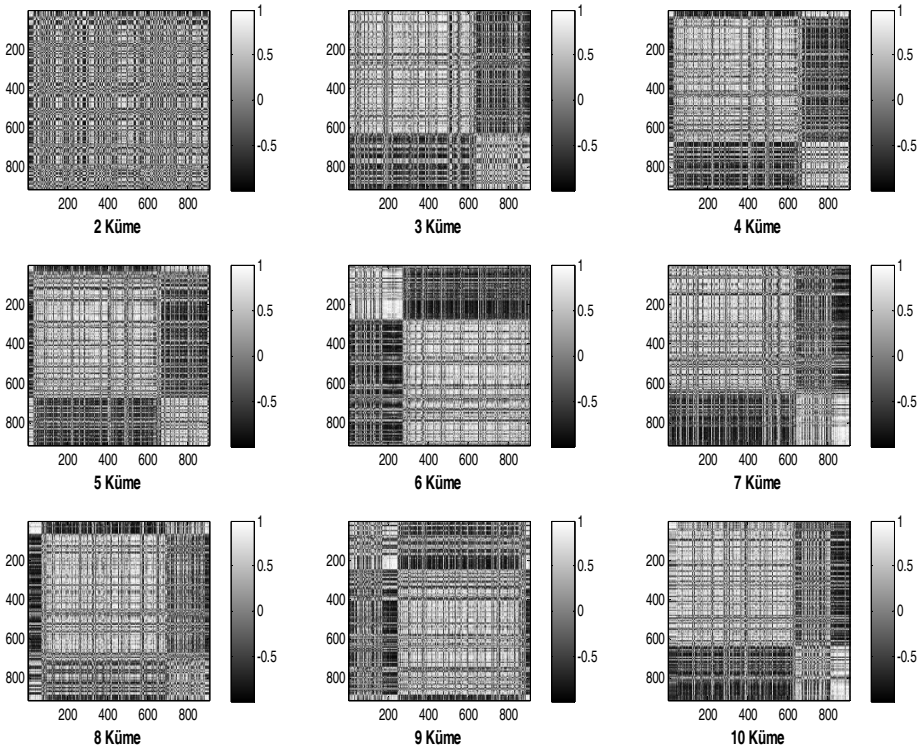
Şekil 12 'deki PolyViz grafiklerinde küme potansiyeli taşıyan veri dağılımlarının küresel olmadığı, yoğunluk ve hacimlerinin farklı olduğu tespit edilmişti. Bu da, kümeleri temsil eden belirli referans noktalarından hareketle, küme doğruluklarını hesaplayan küme doğrulama endekslerinin başarısını olumsuz yönde etkileyebilmektedir.

Uygun küme sayısının tespitine yönelik endeks hesaplamalarında net bir sonuca varılamamıştır. Ancak bu 4 kümeleme yöntemi için elde edilen sonuçlar görselleştirilerek küme kaliteleri anlaşılabilir. Bu sayede uygun küme sayısı ve kümeleme algoritması tespit edilebilir. Şekil 13, 14, 15 ve 16' da 4 farklı kümeleme yöntemiyle elde edilen sonuçların korelasyon benzerlik ölçülerine göre çizilmiş matris grafikleri bulunmaktadır. Matris grafiklerinde beyaz

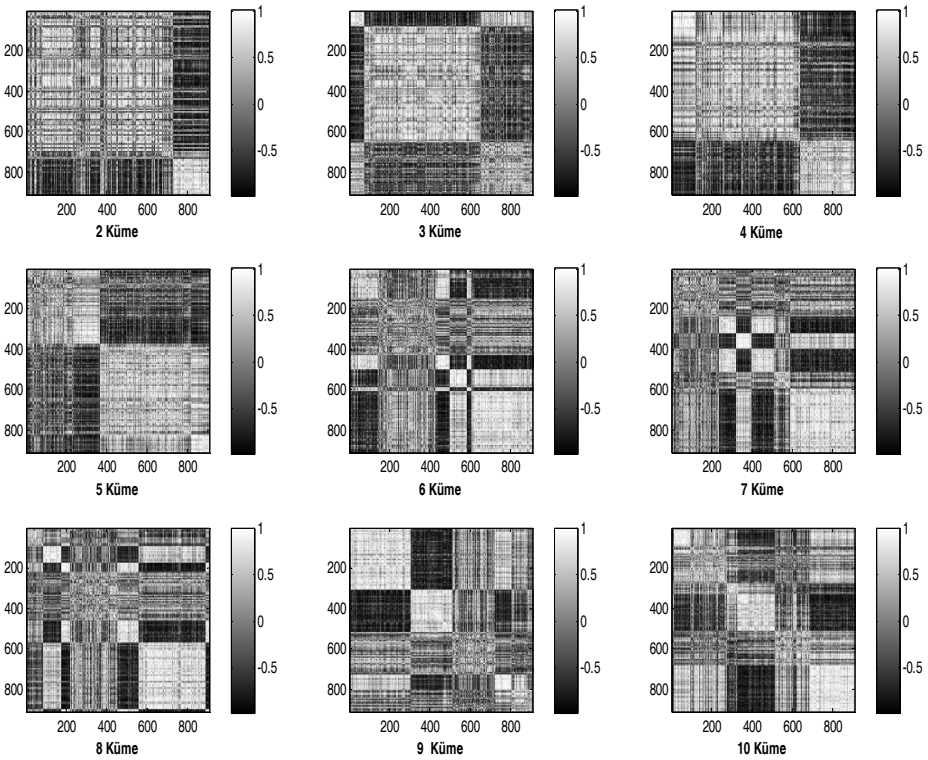
renkler ilçe veri setindeki ilçelerin birbirilerine çok benzediğini, koyu renkler ilçe veri setindeki ilçelerin birbirilerine hiç benzemediğini gösterir. Bu sayede matris grafikleriyle ilçelerin kendi içinde homojen kendi aralarında heterojen bir yapıda kümelenip kümelenmediği görsel bir şekilde anlaşılabilir.



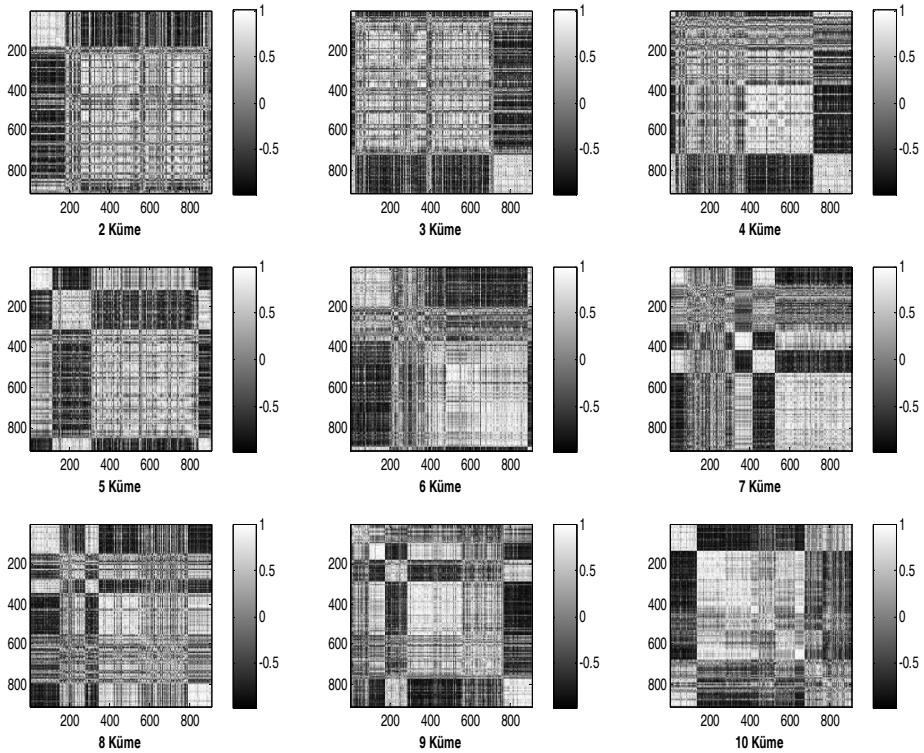
Şekil 13. Tek bağlantılı hiyerarşik kümeleme yöntemiyle 2, 3, 4, 5, 6, 7, 8, 9 ve 10 kümeye ayrılan ilçe veri setinin küme sonuçlarını gösteren matris grafikleri



Şekil 14. Tam bağlantılı hiyerarşik kümeleme yöntemiyle 2, 3, 4, 5, 6, 7, 8, 9 ve 10 kümeye ayrılan ilçe veri setinin küme sonuçlarını gösteren matris grafikleri



Şekil 15. K-ortalamalar kümeleme yöntemiyle 2, 3, 4, 5, 6, 7, 8, 9 ve 10 kümeye ayrılan ilçe veri setinin küme sonuçlarını gösteren matris grafikleri



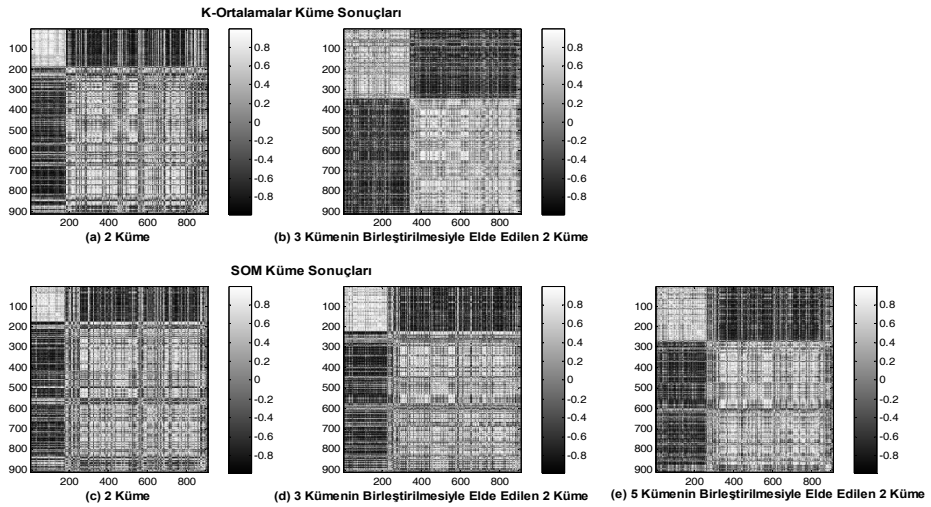
Şekil 16. SOM kümeleme yöntemiyle 2, 3, 4, 5, 6, 7, 8, 9 ve 10 kümeye ayrılan ilçe veri setinin küme sonuçlarını gösteren matris grafikleri

Şekil 13 ve Şekil 14 ’de tek ve tam bağlantılı hiyerarşik kümeleme yöntemleriyle elde edilen 2, 3, 4, 5, 6, 7, 8, 9 ve 10 kümenin sonuçları matris grafikleriyle gösterilmiştir. Şekillere göre tek ve tam bağlantılı hiyerarşik kümeleme yöntemiyle kendi içinde homojen kendi aralarında heterojen küme yapıları net bir şekilde gözlenememiştir. Dolayısıyla tek ve tam bağlantılı hiyerarşik kümeleme yöntemlerinin ilçeleri kümelemede başarısız olduğu söylenebilir. Şekillerden tam bağlantılı kümeleme yönteminin tek bağlantılı kümeleme yöntemine göre daha belirgin kümeler elde ettiği de söylenebilir.

Şekil 15 ve Şekil 16 ’da k-ortalamalar ve SOM kümeleme yöntemleriyle elde edilen 2, 3, 4, 5, 6, 7, 8, 9 ve 10 kümenin sonuçları matris grafikleriyle gösterilmiştir. Şekillere göre k-ortalamalar ve SOM kümeleme yöntemleriyle sadece 2 ’ye ayrılan ilçe veri seti, kendi içinde homojen kendi aralarında heterojen bir yapı sergilemektedir. Dolayısıyla k-ortalamalar ve SOM kümeleme yöntemlerine göre ilçe veri seti için en uygun küme sayısı 2’dir. Bu iki küme için çizilen matris grafiklerine baktığımızda küme içi homojenliği ve kümeler arası heterojenliği bozan bazı noktaların bulunduğu gözlenmiştir. Yani k-ortalamalar ve SOM kümeleme yöntemleriyle 2’ ye ayrılan ilçe veri setinin iyi kümelenmesini bozan bazı ilçelerin bulunduğu anlaşılmıştır.

Şekil 12 ’deki PolyViz grafiklerinde küme potansiyeli taşıyan veri dağılımlarının küresel olmadığı, yoğunluk ve hacimlerinin farklı olduğu tespit edilmişti. Bu da belirli bir merkeze göre küresel biçimde kümeleme yapan k-ortalamalar ve SOM kümeleme yöntemlerinin başarısını olumsuz yönde etkilemektedir. Veri seti 2 ’den fazla alt kümeye ayrılarak bu başarısız

kümelenmenin önüne geçilebilir. K-ortalamalar ve SOM kümeleme yöntemlerinin farklı küme sonuçları için çizilen Şekil 15 ve Şekil 16 'deki matris grafiklerine bakıldığında k-ortalamalar için 3, SOM için 3 ve 5 kümeye ayrılan ilçe veri setinin kendi içinde homojen kümelenmeler sergilediği gözlenmiştir. Bu 3 kümeden 2 tanesinin ve 5 kümeden de 4 tanesinin birbirine benzediği gözlenmiştir. Bu benzeyen kümelerin birleştirilmesi suretiyle 3 ve 5 küme, 2 kümeye indirgenerek k-ortalamalar ve SOM kümeleme yöntemlerinin başarısı arttırılabilir.



Şekil 17. K-ortalamalar ve SOM kümeleme yöntemleriyle 2, 3 ve 5 kümeye ayrılarak 2 kümeye indirgenen ilçe veri setinin küme sonuçlarını gösteren matris grafikleri

Şekil 17 (a), (c) 'de k-ortalamalar ve SOM kümeleme yöntemleriyle 2 kümeye ayrılan ilçe veri setini, Şekil 17 (b), (d), (e) 'de k-ortalamalar ve SOM kümeleme yöntemleriyle önce 3 ve 5 kümeli daha sonra kümelerin birleştirilmesiyle elde edilen 2 kümeli ilçe veri setini gösteren matris grafikleri bulunmaktadır. Şekil 17 (b), (d), (e) 'deki matris grafikleri Şekil 17 (a), (c) 'deki matris grafiklerine göre kendi içinde daha homojen, kendi aralarında daha heterojen bir yapı sergilemektedir. Dolayısıyla her iki kümeleme yöntemi içinde Şekil (b), (d), (e) 'nin elde edilmesinde kullanılan kümeleme yaklaşımının daha başarılı olduğu sonucu ortaya çıkmaktadır. Şekil 17 'deki şekiller incelendiğinde k-ortalamalar kümeleme yönteminin SOM kümeleme yöntemine göre daha başarılı kümeler elde ettiği de söylenebilir. Dolayısıyla 911 ilçe veri setinin kümelenmesinde en başarılı kümeleme yöntemi k-ortalamalar kümeleme yöntemidir.

4.5. Küme Sonuçları

Çalışmanın bundan sonraki aşamalarında k-ortalamalar kümeleme yönteminin bulduğu 2 küme denilince önce 3 kümeye ayrılmış daha sonra benzer kümelerin birleştirilmesiyle elde edilen 2 küme akla gelecektir.

Çizelge 4. K-ortalamalar kümeleme yöntemiyle kümelenen ilçelerin istatistikleri

K-Ortalamalar Küme Sonuçları İçin Tanımlayıcı İstatistikler												
Kümeler	1. Küme			2. Küme			Kümeler	1. Küme			2. Küme	
Değişkenler	İlçe Sayısı	Ortalama	Standart Sapma	İlçe Sayısı	Ortalama	Standart Sapma	Değişkenler	İlçe Sayısı	Ortalama	Standart Sapma	İlçe Sayısı	Ortalama	Standart Sapma
X1	341	18.67	22.43	570	-1.47	19.28	X11	341	12.62	12.62	570	2.96	2.71
X2	341	64.79	19.49	570	35.34	12.87	X12	341	5.37	5.37	570	2.56	1.47
X3	341	1178.93	4563.43	570	54.63	57.26	X13	341	9.62	9.62	570	2.97	1.46
X4	341	33.52	7.76	570	46.02	4.91	X14	341	3.49	3.49	570	1.32	0.63
X5	341	9.13	4.81	570	5.14	3.14	X15	341	2.72	2.72	570	0.69	0.35
X6	341	95.09	2.44	570	90.12	5.31	X16	341	7.31	7.31	570	3.15	0.94
X7	341	83.26	7.66	570	71.10	11.76	X17	341	1.41	1.41	570	0.49	0.22
X8	341	6.41	2.77	570	3.32	0.99	X18	341	5.63	5.63	570	1.89	0.62
X9	341	3.65	2.56	570	1.15	0.56	X19	341	238.70	238.70	570	131.60	145.02
X10	341	45.63	20.95	570	78.83	8.40	X20	341	11.33	11.33	570	7.60	4.84

Çizelge 5. Küme sonuçları için ANOVA

K.Ortalamalar Küme Sonuçları İçin ANOVA							K.Ortalamalar Küme Sonuçları İçin ANOVA						
Değişkenler	İstatistikler	Sum of Squares	df	Mean Square	F	Sig.	Değişkenler	İstatistikler	Sum of Squares	df	Mean Square	F	Sig.
X1	Between Groups	86537.63	1	86537.63	205.55	0.00	X11	Between Groups	19939.97	1	19939.97	432.33	0.00
	Within Groups	382695.55	909	421.01				Within Groups	41924.85	909	46.12		
	Total	469233.18	910					Total	61864.82	910			
X2	Between Groups	184979.99	1	184979.99	752.78	0.00	X12	Between Groups	1683.45	1	1683.45	383.46	0.00
	Within Groups	223368.98	909	245.73				Within Groups	3990.59	909	4.39		
	Total	408348.97	910					Total	5674.04	910			
X3	Between Groups	269699492.43	1	269699492.43	34.62	0.00	X13	Between Groups	9426.90	1	9426.90	691.43	0.00
	Within Groups	7082329503.65	909	7791341.59				Within Groups	12393.30	909	13.63		
	Total	7352028996.08	910					Total	21820.20	910			
X4	Between Groups	33366.79	1	33366.79	887.95	0.00	X14	Between Groups	1010.20	1	1010.20	741.70	0.00
	Within Groups	34157.96	909	37.58				Within Groups	1238.06	909	1.36		
	Total	67524.75	910					Total	2248.26	910			
X5	Between Groups	3386.96	1	3386.96	228.88	0.00	X15	Between Groups	876.00	1	876.00	470.33	0.00
	Within Groups	13451.58	909	14.80				Within Groups	1693.03	909	1.86		
	Total	16838.54	910					Total	2569.02	910			
X6	Between Groups	5273.47	1	5273.47	265.27	0.00	X16	Between Groups	3697.38	1	3697.38	815.66	0.00
	Within Groups	18070.52	909	19.88				Within Groups	4120.48	909	4.53		
	Total	23343.99	910					Total	7817.86	910			
X7	Between Groups	31533.23	1	31533.23	290.61	0.00	X17	Between Groups	180.03	1	180.03	719.47	0.00
	Within Groups	98634.21	909	108.51				Within Groups	227.46	909	0.25		
	Total	130167.44	910					Total	407.49	910			
X8	Between Groups	2049.56	1	2049.56	588.31	0.00	X18	Between Groups	2997.08	1	2997.08	700.40	0.00
	Within Groups	3166.80	909	3.48				Within Groups	3889.70	909	4.28		
	Total	5216.36	910					Total	6886.78	910			
X9	Between Groups	1335.90	1	1335.90	502.71	0.00	X19	Between Groups	2447353.86	1	2447353.86	95.39	0.00
	Within Groups	2415.60	909	2.66				Within Groups	23321869.97	909	25656.62		
	Total	3751.50	910					Total	25769223.83	910			
X10	Between Groups	235254.37	1	235254.37	1128.75	0.00	X20	Between Groups	2975.15	1	2975.15	79.78	0.00
	Within Groups	189453.81	909	208.42				Within Groups	33900.26	909	37.29		
	Total	424708.18	910					Total	36875.41	910			

Çizelge 4 'de k-ortalamalar kümeleme yöntemiyle 2 kümeye ayrılan ilçelerin istatistikleri bulunmaktadır. İstatistiklerden görüldüğü gibi 1. kümenin X4 (ücretli çalışan kadınların toplam istihdama oranı) ve X10 (tarım kesiminde çalışanların toplam istihdama oranı) değişken ortalamaları haricindeki bütün değişken ortalamaları 2. kümenin değişken ortalamalarından yüksektir. Kümeleme yaptığımız değişkenlerden sadece X10 ve X5 (işsizlik oranı) değişkenleri gelişmişlikle ters orantılıdır. Dolayısıyla X4 ve X5 değişkenleri dışında 1. küme 2. kümeye göre daha gelişmiş durumdadır. Buradan hareketle 1. kümeye gelişmiş, 2. kümeye daha az gelişmiş ilçeler topluluğu diyebiliriz.

Çizelge 4 'de kümelerin değişken ortalamalarının bir birinden ayrıştığı gözlenir. ANOVA yaklaşımı ile de bu değişken ortalamalarının istatistikî olarak birbirlerinden farklı olduğu tespit edilebilir. Çizelge 5 'de k-ortalamalar yöntemiyle kümelenen ilçelerin değişkenleri için hesaplanan ANOVA tablosu bulunmaktadır. Tabloya göre 0.05 anlamlılık düzeyine göre kümelerde bulunan tüm değişken ortalamaları birbirinden istatistikî olarak faklıdır (değişkenler farklı ana kütlelerden gelmektedir) Dolayısıyla küme yapıları tüm değişkenler bazında farklılık göstermektedir.

Çizelge 6. K-ortalamalar kümeleme yöntemine göre bölgelere göre kümeler

İlçeler	K-Ortalamalar					
Bölgeler	Küme 1	Küme 2	Küme 1 (%)	Küme 2 (%)	Küme Sınıfı	Genel Toplam
Akdeniz	54	50	51.92%	48.08%	1. Küme	104
Doğu Anadolu	28	136	17.07%	82.93%	2. Küme	164
Ege	44	72	37.93%	62.07%	2. Küme	116
Güneydoğu Anadolu	8	41	16.33%	83.67%	2. Küme	49
İç Anadolu	63	89	41.45%	58.55%	2. Küme	152
Karadeniz	48	143	25.13%	74.87%	2. Küme	191
Marmara	96	39	71.11%	28.89%	1. Küme	135
Genel Toplam	341	570	37.43%	62.57%	2. Küme	911

Çizelge 6 'da k-ortalamalar kümeleme yöntemiyle bulunan bölgelere göre kümeler verilmiştir. Tablolardan da gözüktüğü gibi Doğu, Güneydoğu Anadolu ve Karadeniz bölgelerinin ilçelerinin çoğunluğu 2. kümede yer almaktadır. Marmara ve Akdeniz bölgesinin ilçeleri ise çoğunlukla 1. kümede yer almaktadır.

5. SONUÇ VE ÖNERİLER

Bu çalışmada, Türkiye'de 918 ilçeye ait 20 değişkenle yapılan, veri madenciliği çerçevesinde kümeleme analizinde, görselliğin yeri ve önemi vurgulanmaya çalışılmıştır. Sadece sayısal değerlendirmelerle yapılan kümeleme analizlerinde yanılmalar olabileceğini ve doğru kümeleme analizlerinin görsellikle pekiştirilmesi gereği bu makalede vurgulanmıştır. Bu amaca dönük olarak sapan değer, potansiyel küme sayı ve yapıları, uygun kümeleme algoritma tercihleri ve nihayet küme doğrulama çalışmaları bu çerçevede tartışılıp ele alınmıştır.

Yapılan uygulamada, sosyoekonomik özelliklere göre 81 ildeki 918 ilçe, görsel yöntemlerin de desteğiyle, Andrews eğrileri, PolyViz ve Matris grafikleri kullanılarak k-ortalamalar, tek ve tam bağlantılı hiyerarşik ve SOM kümeleme yöntemleriyle kümelenmiştir**.** Sonuç itibariyle görsel teknikler ve k-ortalamalar yöntemi tercihiyle ana kütle gelişmiş ve daha az gelişmiş özelliklerden oluşan iki kümeye başarılı bir şekilde ayrılabileceği görülmüştür. Burada görsel tekniklerin sapan değerlerin ayıklanmasında, potansiyel küme yapılarının ve sayılarının doğru belirmesinde önemli katkıları olduğu gösterilmiştir.

Araştırmanın diğer ayağı olan küme özelliklerinin yorumlanması aşamasında ise gerçeklerle bu ayırımın iyi bir şekilde bağdaştığı belirlenmiştir. Bu bağlamda kümeleme sonucunda gelişmiş ilçeler kümesinde en fazla Marmara bölgesinin ilçeleri bulunmaktadır. Gelişmemiş ilçeler kümesinde ise en fazla Doğu ve Güneydoğu Anadolu bölgelerinin ilçeleri bulunmaktadır. Genel bir değerlendirme yapıldığında sosyoekonomik özelliklere göre Türkiye'de bölgeler itibariyle homojen bir dağılım olmadığı gözlenmiştir. Marmara Bölgesinin tek başına homojen bir yapı sergilediği ve Türkiye'nin diğer bölgelerinden daha gelişmiş olduğu tespit edilmiştir.

En gelişmiş bölge olan Marmara bölgesinde sanayinin yaygın olması, okullaşma ve buna bağlı olarak okuryazar oranının yüksek olması gelişmişlikte önemli bir faktör olduğu tespit edilmiştir. Akdeniz bölgesinin de birinci kümede yer aldığı görülmüştür. Diğer kümede yer alan, az gelişmiş bölge olan, Doğu ve Güneydoğu Anadolu Bölgelerine yapılan yatırımların azlığı, kız çocuklarının okula gönderilmemesi gerçeği ve okullaşma oranının da düşük olması bu neticelerin elde edilmesinde önemli faktörler olarak ortaya çıkmaktadır.

Görsel tekniklerle diğer kümeleme analizlerini birleştiren çalışmaların artırılması ve bunun bir standart olarak ele alınma gereği, bu alandaki görsel tekniklerin çoğaltılması bu çalışmanın en önemli bulgusunu oluşturmaktadır.

REFERENCES / KAYNAKLAR

[1] Andrews D. F., "Plots Of High-Dimensional Data", Biometrics, Cilt 28, No 1, 125-136, 1972.

[2] Everitt B. S. and Nicholls P., "Visual Techniques For Representing Multivariate Data", The Statistician, Cilt 24, No 1, 37-49, 1975.

[3] Chen K. and Liu L., "IVIBRATE: Interactive Visualization-Based Framework for Clustering Large Datasets", ACM Transactions on Information Systems, Cilt 24, No 2, 245-294, 2006.

[4] Xu R. and Wunsch D., "Survey Of Clustering Algorithms", IEEE Transactions on Neural Networks, Cilt 16, No 3, 645-678, 2005.

[5] Han J. and Kamber M., "Data Mining: Concepts and Techniques", 1. Baskı, Morgan Kaufmann, San Francisco, 2000, 270-296.

[6] Tan P., Steinbach M. and Kumar V., "Introduction To Data Mining", 1. Baskı, Addison Wesley, USA., 2006, 487-532, 541-542 and 569-642.

[7] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "Cluster validity methods: part I", SIGMOD Record., Cilt 31, No 2, 40-45, 2002.

[8] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "Cluster validity methods: part II", SIGMOD Record., Cilt 31, No 3, 19-27, 2002.

[9] Bezdek J. C. and Hathaway R. J., "VAT: A Tool For Visual Assesment Of (Cluster) Tendency", Proceedings of the 2002 International Joint Conference on, Honolulu, HI, USA, May, 2002, 2225-2230.

[10] Bilgin T. T., "Çok Boyutlu Uzayda Görsel Veri Madenciliği İçin Üç Yeni Çatı Tasarımı ve Uygulamaları", Doktora Tezi, Fen Bilimleri Enstitüsü, Marmara Üniversitesi, İstanbul, 2007.

[11] Ankerst M., "Visual Data Mining", Doktora Tezi, Institute for Computer Science, University of Munchen, 2000.

[12] Keim D. A., "Information Visualization and Visual Data Mining", IEEE on Transactions on Visualizations and Computer Graphics, Cilt 8, No 1, 1-8, 2002.

[13] Bartke K., "2D, 3D and Hight-Dimensional Data and Information Visualization", University of Hannover, Institut für Wirtschaftsinformatik. Available from: http://www.iwi.uni-hannover.de/lv/seminar_ss05/bartke/Assets/Paper.pdf |accessed September 24,2008|.

[14] Martinez W. L. and Martinez A. R., "Exploratory Data Analysis with MATLAB", 1. Baskı, Boca Raton: CRC Press, USA., 2005, 321-331 ve 337-343

[15] İnal E., "Verinin Keşfedilmesi", Gebze İleri teknoloji Enstitüsü, Available from: http://www.bilmuh.gyte.edu.tr/BIL454/ |accessed September 24,2008|.

[16] Özdamar K., "Paket Programlar İle İstatistiksel Veri Analizi 2", 5. Baskı, Kaan Kitabevi, Eskişehir, 2004, 213-230 ve 279-292

[17] Kovacs F., Legany C. and Babos A., "Cluster Validity Measurement Techniques", 6th International Symposium of Hungarian Researchers on Computational Intelligence, Budapest, Hungary, November, 2005

[18] Toledo M. D. G., "A Comparison In Cluster Validation Techniques ", Yüksek Lisans Tezi, Mathematics Department, University Of Puerto Rico, 2005.

[19] Volkovich Z., Barzily Z. and Morozensky L., "A Statistical Model Of Cluster Stability", Pattern Recognition, Cilt 41, No 7, 2174-2188, 2008

[20] DİE, "Genel Nüfus Sayımı, İdari Bölünüş", İstanbul, 2000

[21] Türkiye Bankalar Birliği, "Banka Şubeleri Sorgulama", Available from: http://www.tbb.org.tr/net/subeler/, |accessed November 12,2008|.

[22] Devlet İstatistik Enstitüsü, "İlçeler İtibariyle Gayri Safi Yurtiçi Hasıla", Available from: http://www.die.gov.tr/TURKISH/SONIST/GSYIH/241097t1.htm, |accessed November 12,2008|.

[23] The MathWorks Inc, Available from: http://www.mathworks.com/, |accessed November 12,2008|.

[24] SPSS Inc, Available from: http://www.spss.com/, |accessed November 12,2008|.

[25] Demsar J., Zupan B. and Leban G., Faculty of Computer and Information Science, University of Ljubljana, Available from: http://www.ailab.si/orange/, |accessed November 12,2008|.

[26] Saraçlı S., Yılmaz V. ve Kaygısız Z., "Türkiye' de Beşeri Kalkınmışlığın Coğrafi Dağılımının Çok Değişkenli İstatistiksel Tekniklerle İncelenmesi", 3. Ulusal Bilgi, Ekonomi ve Yönetim Kongresi, Eskişehir, Kasım, 2004, 21-28.

[27] Wang, K.J., Cluster Validation ToolBox CVAP, Available from: http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=14620 |accessed November 12,2008|.

[28] Ding Y. and Harrison R. F., "Relational visual cluster validity (RVCV)", Pattern Recognition Letters, Cilt 28, No 15, 2071-2079, 2007.