ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

KÜMELEME ANALİZİNDE KÜME SAYISININ BELİRLENMESİ ÜZERİNE BİR ÇALIŞMA

Azize Celile GÜNAY ATBAŞ

İSTATİSTİK ANABİLİM DALI

ANKARA
2008

ÖZET

Yüksek Lisans Tezi

KÜMELEME ANALİZİNDE KÜME SAYISININ BELİRLENMESİ ÜZERİNE BİR ÇALIŞMA

Azize Celile GÜNAY ATBAŞ

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Yrd.Doç. Dr. Cemal ATAKAN

Kümeleme Analizinde amaç; ele alınan özellikleri bakımından birbirine benzer olan birimleri saptayarak kümelenme yapılarını oluşturmaktır. Ancak Kümeleme Analizinde küme sayısı önceden bilinmez ve bu bilinmezlik kümeleme analizinin en tartışmalı konusu olmuştur. Bu çalışmada kümeleme analizi bu yönüyle ele alınmaya çalışılmıştır. Çalışmada kümeleme yöntemlerinden tek bağlantı yöntemi, tam bağlantı yöntemi , Ward yöntemi, k-Ortalama yöntemi kullanılmıştır. Uzaklık ölçüsü olarak öklid uzaklığı ve karesel öklid uzaklığı kullanılmıştır. Uygulama bölümünde Türkiye'de 81 ilde işlenen 11 farklı suç türüne göre ceza evine giren hükümlü sayıları illerin nüfuslarına oranlanmıştır. Adam öldürme, cinsel suçlar, kişiyi hürriyetinden yoksun bırakma, hırsızlık, gasp, dolandırıcılık, uyuşturucu, sahtecilik, zimmet, kaçakçılık, orman suçlarından 2006 yılında ceza evine giren hükümlü sayıları TÜİK resmi internet sayfasından adalet istatistikleri bölümünden alınmıştır. 4 kümeleme yöntemiyle(tek bağlantı, tam bağlantı, Ward ve k-ortalama) iller bu suç türlerine göre kümelenmiştir. Çalışmada, Van ilinin tek başına bir küme olması kaçakçılık suçundan cezaevine giren hükümlü sayısının diğer illere oranla fark edilebilir oranda fazla olmasından kaynaklanabildiği sonucuna varılmıştır. Ankara, Antalya, Bursa, Denizli, Gaziantep, İstanbul, İzmir, Kayseri ve Konya gibi büyük illerin kendi aralarında küme oluşturmalarının da dikkat çekici olduğu gözlenmiştir. Kümeleme yöntemlerinden tek bağlantı, tam bağlantı yöntemleri 81 ili suç türlerine göre 5 kümeye ayırırken, Ward yöntemi ve k-ortalama yöntemi illeri 7 kümeye ayırmıştır. Oluşan kümelemenin kalitesi Küme geçerliliği indekslerinden Silhouette indeksi, Calinski Harabasz indeksi, Krzanowski Lai indeksi ve Wilk's Lambda istatistiği ile değerlendirilmiştir. Bu indekslerden Silhouette indeksi, Calinski Harabasz indeksi, Krzanowski Lai indeksleri tek bağlantı, tam bağlantı yöntemleri ile illeri 5 kümeye, Ward yöntemi ve k-ortalama yöntemi ile 7 kümeye ayırmanın kaliteli kümeleme olduğunu göstermiştir. Wilk's Lambda istatistiği ile iller, tek bağlantı yöntemine göre 10 kümeye, tam bağlantı yöntemine göre 7 kümeye, Ward yöntemi ve k-ortalama yöntemine göre 8 kümeye ayrılmıştır. Oluşan kümeler incelendiğinde, Ward yönteminin diğer yöntemlere göre daha anlamlı küme yapısı ortaya çıkardığı gözlenmiştir.

Temmuz 2008, 60 sayfa

Anahtar Kelimeler: Uzaklık Ölçüsü, Kümeleme Analizi, Hiyerarşik Kümeleme, k-Ortalama Yöntemi

ABSTRACT

Master Thesis

A STUDY ON DETERMINING THE NUMBER OF CLUSTERS IN CLUSTER ANALYSIS

Azize Celile GÜNAY ATBAŞ

Ankara University
Graduate School of Natural and Applied Sciences
Department of Statistics

Supervisor: Asst. Prof. Dr. Cemal ATAKAN

The purpose of Cluster Analysis is to determine the units similar to each other in terms of their characteristics studied, and to define their clustering structures. However, in Cluster Analysis the number of clusters is not known beforehand and this uncertainty is the most controversial issue of the Cluster Analysis. In this study, cluster analysis has been taken up considering this issue. In the study, clustering methods including single link, complete link, Ward's method and k-mean have been used. Euclidian distance and squared Euclidian distance have been used as the measure of distance. In the practical part, prisoners sentenced for 11 different types of crime in 81 provinces of Turkey have been compared to the populations of these provinces. Numbers of prisoners sentenced in 2006 for homicide, sexual offences, deprivation of personal freedom, theft, usurpation, fraud, drug abuse, forgery, embezzlement, smuggling and forest crimes have been taken from the justice statistics of the Turkish Statistical Institute published in their official web page. Using the 4 clustering methods the provinces have been clustered by the foregoing types of crime. It has been concluded that the situation of the province of Van which made a cluster by itself can stem from the fact that the number of persons imprisoned in this province for smuggling is distinguishably higher when compared to other provinces. It has also been observed that big provinces like Ankara, Antalya, Bursa, Denizli, Gaziantep, Istanbul, Izmir, Kayseri and Konya notably make up a cluster between themselves. Single link and complete link clustering methods have divided the 81 provinces into 5 clusters by types of crime, while Ward's method and k-mean divided them into 7. The quality of this clusterization has been assessed using cluster validity indices including Silhoutte index, Calinski and Harabasz index, Krzanowski and Lai index, and Wilks' Lambda statistics. Silhoutte index, Calinski and Harabasz index, and Krzanowski and Lai index have shown that dividing the provinces into 5 clusters using single link and complete link methods, and dividing the same into 7 clusters using Ward's method and k-mean method are qualified clusterings. It has been concluded that with Wilks' Lambda statistics it would be much more acceptable to divide the provinces into 10 clusters using single link method; into 7 using complete link method, and into 8 using Wards' method and k-mean. It has been observed that Ward's method is the best method.

July 2008, 60 pages

Key Words : Measure of Distance, Cluster Analysis, Hierarchical Clustering, k-Mean Method

TEŞEKKÜR

Çalışmalarımı yönlendiren, araştırmalarımın her aşamasında bilgi, öneri ve yardımlarını esirgemeyen, yoğun çalışma temposunda bana vakit ayıran danışman hocam sayın Yard.Doç.Dr. Cemal ATAKAN' a, (Ankara Üniversitesi Fen Fakültesi) akademik gelişimi destekleyen Adalet Bakanlığı Strateji Geliştirme Başkanı (Hakim), Başkanım Sayın Hüseyin YILDIRIM' a , tezimin düzeltmelerinde yardımını esirgemeyen , motivasyonumu arttıran tez koçum, sevgili dostum, Hilal YÜCEL' e, kendisi uzaklarda olsada manevi desteğini hep yanımda hissettiğim çok kıymetli dostum Serap UÇAR' a , tezime tam anlamıyla başlamama vesile olan sevgili meslektaşım Emel AKTAŞ' a , çalışmalarım süresince varlığıyla strese girmemi engelleyen, sıkıntılı anlarımda her zaman yüzümü güldüren bana destek olmak için yüksek lisansa başlayan hayat arkadaşım, eşim Barış Egemen ATBAŞ ' a , bugünlere gelmemdeki en büyük emeğe sahip Annem Jale GÜNAY' a ve gerek kardeşlerim gerek benim kişisel gelişimimizi çocukluğumuzdan beri destekleyen , bize çalışma hayatımızda başarılarıyla azmiyle örnek olan Babam (Yargıtay 9.Hukuk Dairesi Üyesi) Doç. Dr. Cevdet İlhan GÜNAY' a en içten duygularımla teşekkür ederim.

Azize Celile GÜNAY ATBAŞ
Ankara, Temmuz 2008

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ

İnsanoğlu var olduğundan bu yana etrafında bulunan nesneleri bazı özelliklerine göre sınıflara ayırma eğiliminde olmuştur. Örneğin bitkileri yenilir yenmez diye ilkel çağlarda sınıflayan insanoğlu, gün geçtikçe sınıflandırmanın amacını ve kapsamını değiştirmiştir. Sınıflandırmanın en genel amacı; benzer olanı benzemezden ayırmaktır (Everitt *et al.* 2001). Birimlerin sayısı arttıkça birimleri sınıflandırmak daha da zorlaşmış ve yeni teknikler bulmayı gerektirmiştir. Bu gereksinim sonucu kümeleme analizi kavramı ortaya çıkmıştır. Bu kavram 1960 yıllarından sonra gelişmiş ve geniş uygulama alanlarına sahip olmuştur (Anderberg 1973). Kümeleme analizi günümüzde veri madenciliği, bankacılık, pazarlama, tıp, sosyoloji, kriptoloji gibi çeşitli alanlarda bilimsel ve stratejik sonuçlar elde etmek amacıyla uygulanmaktadır.

Kümeleme analizinde amaç, benzer özellikleri gösteren birimleri aynı kümede toplayarak, özet bilgi elde etmektir. Kısa bir tanımla kümeleme analizi doğal sınıflamaları hakkında kesin bir bilgi bulunmayan durumlarda, yığına ilişkin tahminlerin yapılmasında kullanılan teknikler topluluğudur (Hartigan 1975). Burada hakkında ön bilgi bulunmayan grupları sınıflandırırken, yığını kaç kümeye ayırmak gerektiği kümeleme analizinde üzerinde durulması gereken bir konudur. Bu çalışmada kümeleme analizi bu yönüyle ele alınacaktır.

Kümeleme Analizi, Linnaeus'un 1753 yılında hayvanların ve bitkilerin sınıflandırılması üzerine yaptığı çalışmaya dayanan bir araştırma metodudur. Günümüzde bu analiz tıptan ekonomiye varan bir yelpazede uygulanmaktadır (Hofman and Jarvis 1998).

Anderberg (1973), sayısal sınıflandırma analizinden en iyi yararı elde etmek için, aşağıdaki hususlara dikkat edilmesi gerektiğini belirtmiştir;

a) Değişik sınıflandırma yöntemleri, aynı veri kümesinde farklı sonuçlar verebilir. Bu sebepten tek bir kümeleme yöntemine bağlı kalmayıp birden fazla yöntemin denenmesi daha sağlıklı sonuçlar verecektir.
b) Sayısal sınıflandırma yöntemleri hipotez tahminleri için amaçtır.
c) Sınıflandırma sonucunda, veri kümesinin karışık yapısı ortaya çıkabilir.

d) Verilerde küme yapısının olmaması veya bir tek küme olması gibi özel bir durumla da karşılaşılabilir.

Bu hususlar neticesinde bir problemin, farklı yöntemlere göre, farklı sonuçları olacağından araştırmada birden fazla yöntem kullanılarak sonuçlar değerlendirilmelidir.

Everitt (1974), birimlerin sınıflandırılmasını ilk önce biyoloji ve zooloji alanlarında yapmıştır.

Hartigan (1975) tarafından yazılan kitapta, medeniyet savaşları, İngiliz kelebeklerinin ortaya çıkış zamanı gibi çeşitli konularla ilgili kümeleme analizine ilişkin kırkın üzerinde veri setinin dahil olduğu bir çok uygulama yer almaktadır.

Kinnear and Taylor (1979), yaptıkları çalışmada kişilerin sosyo-ekonomik nitelikleri, siyasal eğilimleri esas alınarak bu özellikler itibariyle benzer olan kişilerin aynı kümelerde toplanması amacıyla tesadüfî olarak seçilen 64 kişi üzerinde 19 değişkenin değeri ölçülmüştür. Bu 19 değişken, cevaplayıcıların dünya görüşlerini, siyasal eğilimlerini, dini inançlarını ve tahsil, gelir, meslek gibi sosyo-ekonomik niteliklerini saptamak amacını güden sorularla ölçülmüştür. Örnek bireylerden toplanan 19 değişkenle ilgili veriler Öklid uzaklık fonksiyonunu kullanan Aşamalı Kümeleme Analizi kullanılarak analiz edilmiş ve çalışma sonunda cevaplayıcıların %53' nün muhafazakâr, geri kalan kısmında radikal değişikliklere karşı, eğitim ve gelir düzeyi oldukça düşük homojen kişilerden oluştuğu gözlemlenmiştir.

Cox *et al.* (1985), kümeleme analizi ile 43 çeşit sert kırmızı kış buğdayının genetik ilişkilerini belirlemeye çalışmışlardır. Analiz sonucunda, varyans-kovaryans matrisinden ($\sum$) ve pedigri analizi sonucunda hesaplanan korelasyon yardımıyla 43 çeşit sert kırmızı kış buğdayının sınıflandırılması yapılarak aralarındaki genetik ilişki belirlenmeye çalışılmıştır. Sonuç olarak 43 buğday çeşidi genetik bakımdan benzerliklerine göre 8 kümeye ayrılarak, bu buğday çeşitleri arasındaki genetik benzerlikler açıklanmaya çalışılmıştır.

Lebeda and Jendrulek (1987), gen sistemleri için genlerde parazit konukçu etkileşimine ait verilerin daha çok niteliksel yönlerden semboller (+,-) veya sözel ifadeler (evet-hayır, etkilenen-etkilenmeyen vb.) kullanılarak ifade edildiğini ve bu tip verilerin aşamalı kümeleme metotları ile analiz edilebileceğini belirtmişlerdir.

Kosaki and Juo (1988), yaptıkları çalışmada kümeleme analizi ile Nijerya'da uluslararası tarımsal araştırma enstitüsüne bağlı deneysel tarım arazilerinde kontrol edilebilir toprak değişim faktörleri temel alınarak, toprakların gruplandırılmasını yapmışlardır.

Line and Butler (1990), iki yönlü sınıflandırılmış verilerin analizi için kümeleme analizi isimli çalışmalarında, veriler homojen alt gruplar içinde tabakalar halinde tertip edilmiş ise iki yönlü sınıflandırılmış verilerin etkileşim yapısını 4 farklı kümeleme metodu ile açıklamıştır. Bu amaç için 2 yeni (araştırma genotipi x çevre interaksiyonu) metot kullanılmıştır. 1.ve 2. metotta regresyon modeli 3.ve 4. metotta ise ANOVA model kullanılmıştır.

Çakır (1994), araştırmacı TV programlarının tespiti ve program değerlendirme niteliklerinin belirlenmesi yönünde 150 kişiye anket uygulamıştır. Anket uygulamasında ilk etapta araştırmada kullanılacak TV programlarının tespitine çalışılmıştır. Böylece 8:00- 24:00 saatleri arasında en çok izlenen 7 kanalda bir hafta içinde yayınlanan 251 program tespit edilip elde edilen veriler öklid uzaklık ölçütü kullanılarak kümeleme analiz yöntemleri ile analiz edilmiştir. Sonuç olarak diğer kümeleme analiz yöntemlerine kıyasla Ward yönteminin daha anlamlı küme yapıları ortaya çıkardığına karar verilmiştir.

Yılmaz (1996), kümeleme analizi ve matematiksel programlama teknikleri isimli çalışmasında kümeleme analizi ve kümeleme analizine matematiksel bir yaklaşım amaçlamaktadır.

Franco *et al*. (1997), bitki genetik kaynakları isimli yaptıkları çalışmalarında genellikle kümeleme analizinin genetik farklılıkların belirlenmesinde kullanıldığını belirtmişlerdir.

Çalışmada; birkaç kritere göre farklı kümeleme yöntemlerinin performansını ortaya koymak, istatistiksel özelliklerden yararlanılarak germplasmlar da çoğalmalar için bir sınıflandırma metodu önermek ve önerilen sınıflama metodunun sonuçlarının temel alt setler biçiminde nasıl uygulanabileceğinin ortaya konmasıdır. Bu çalışmada Ward yöntemi kullanılmıştır.

Öztürk (1999) çalışmasında, varyans analizinde muameleleri mantıklı bir şekilde homojen gruplara ayırmanın bazen faydalı olacağını ifade etmişlerdir. Bu amaçla sık sık çeşitli karşılaştırma işlemlerinin kullanıldığını, ancak en doğrusunun kümeleme analizi tekniklerinin kullanılması olduğunu ifade etmiştir.

Yazgan (2001), yaptığı çalışmada Çukurova Üniversitesi Ziraat Fakültesi, Süt Keçiciliği Araştırma ve Uygulama Ünitesi'nde yetiştirilen kültür ırkı, yerli ırk ve melez keçiler üzerindeki araştırma sonuçlarından elde edilen 59 dişi çebiç(1 yaşına basmış keçi) üzerinden alınan canlı ağırlık, kıl rengi, pigmentasyon, kıl uzaklığı, deri kalınlığı, hemoglobin tipi, transferin tipi, glikoz akşam, kolesterol akşam, glikoz sabah, kolesterol sabah ölçüm değerlerine göre farklı uzaklık ölçüleriyle ve farklı kümeleme analiz yöntemlerini karşılaştırmıştır.

Terlemez (2001), Avrupa Birliğine üye ülkeler ve üyelik başvurusu yapan aday ülkeler arasındaki temel makro ekonomik göstergeler açısından benzerliklerini ortaya koymak, homojen ülke kümelerini belirlemek ve Türkiye'nin bu kümelerden hangisinde yer aldığını belirlemek amacıyla Kümeleme Analizi yöntemini kullanmıştır. Çalışmanın sonucunda, Türkiye ve diğer aday ülkelerin ele alınan değişkenler itibariyle enflasyon ve işsizlik sorununu ön planda tutmaları gerektiği görülmüştür. Çalışmanın bir diğer çıkarımı da, aday ülkeler içerisinde Türkiye'nin durumuna bakıldığında, en önemli makro ekonomik istikrarsızlık göstergesi olarak milli gelir, büyüme oranı, enflasyon ve cari işlemler açığı olarak belirlenmiştir.

Silahtaroğlu'nun (2004) çalışmasına göre okulların yabancı dil eğitimi veren hazırlık sınıflarında, yabancı dil bilgi ve becerilerine göre öğrencilerin farklı gruplarda toplanması, eğitim ve öğretimin aksamasına sebep olmaktadır. Veri madenciliği

kümeleme analizi yöntemini uygulayarak, birbirlerine dilbilgisi ve becerileri daha çok homojen olan öğrencilerin aynı kümede bulunmasıyla gerek soru hazırlama gerek ders işleyiş bakımından dersler daha akıcı ve aksamadan yürüyecektir. Bu amaçla dil eğitiminde kurlara ayırma işlemi veri madenciliğinde kümeleme yöntemi kullanılarak yapılacaksa birkaç hususa dikkat edilmesi gerektiği sonucuna varılmıştır. Bu hususlar, sınav soruları dil öğretiminin mümkün oldukça her yönünü (okuma-yazma, dinleme) kapsayacak şekilde olmalıdır. Sınav sonuçları optik okuyucuda okutulmalı ve kullanılacak olan kümeleme algoritması optik okuyucuya entegre edilmelidir. Bu hususlara dikkat edilerek yapılan analiz sonucunda homojen hazırlık sınıflarının eğitimin, öğretimin aksamadan yürümesi için önemli bir adım olacağı sonucuna varmıştır.

Geler (2005), çalışmasında Türkiye Cumhuriyeti sınırları içinde yer alan 80 ilin sosyo-ekonomik özelliklerini yansıtan 25 değişkeni (gelir vergisi, imalat sanayi, aile başına tarımsal hasıla, bitkisel üretim değeri, doktor başına nüfus,…) temel alarak illeri kümelere ayırmıştır. Analiz sonucunda, İstanbul ve Artvin gibi iller ile, doğu illerinin (Batman, Iğdır, Şırnak, Diyarbakır, Şanlıurfa, Van, Adıyaman, Siirt, Mardin, Bingöl, Bitlis) tutarlı bir biçimde aynı kümede olduğu sonucuna varmıştır.

Işık (2006), Kümeleme Analizi yöntemlerinden bölünmeli kümeleme yöntemi olan $k-$ortalama ve bulanık c algoritmalarını kullanarak web dokümanları üzerinde çeşitli testler uygulamış ve algoritmaların kümeleme başarılarını karşılaştırmıştır.

Dinçer ve Özdamar (1992), yaptıkları çalışmada $g_{1,}g_2, g_3$, r_{cs}, Wilk's Lambda ve Hotelling Lawley iz istatistikleri; rasgele çekilmiş 10, 20, 30, 40 ve 50 birimlik gruplar ile şartlı çekilmiş 20, 30, 40, 50 ve 60 birimlik gruplar üzerinden test edilmiştir. 6 değişkene göre değerlerin öklid uzaklıkları kullanılarak benzerlik matrisleri elde edilmiş ve küme sayısı 2 ile 5 arasında olacak şekilde $k-$ortalama yöntemi ile kümeleyip, tek bağlantı yöntemi ile bağlantıları belirlemişlerdir. Rasgele ve şartlı gruplarda, kümeleme ölçütleri gruplarda benzerlik göstermemiştir. Rasgele grupların birim sayıları artarken Wilk's Lambda ve Hotelling Lawley iz istatistikleri dışındaki kümeleme ölçütleri küme sayılarında düzensiz artış ve azalma olduğunu göstermiştir. Bu sebepten en uygun küme

sayısını belirleme ölçütü olarak Wilk's Lambda ve Hotelling Lawley iz istatistiklerinin olduğunu belirtmişlerdir.

Yılmaz ve Günayergün (2004), 'Türkiye'de Şehir Asayiş Suçları: Dağılış ve Başlıca Özellikleri' konulu çalışmasında, Türkiye şehir yerleşmelerindeki asayiş suçları coğrafi bakış ile değerlendirmişlerdir. Toplam asayiş suçları, şahsa ve mala karşı suçlar olarak sınıflandırarak ele almıştır. Çalışmasının sonucunda büyük nüfuslu sosyo-ekonomik açıdan gelişmiş, şehirleşme oranı yüksek ve göç alan Türkiye'nin batı yarısında suçların daha fazla olduğu sonucuna varmıştır. Çalışmanın analiz kısmında üssel artış yöntemi kullanılmıştır. Suç araştırmaları, suç ile suçun işlendiği 'yer' arasında anlamlı ilişkiler bulunduğunu ortaya koymaktadır (Harries1999, Appiahene-Gyamfi 2002, Karakaş 2004). Suçların farklı dağılış özellikleri göstermesi, bazı illerde bazı suçlar artarken bazı illerde azalması, suçun mekânsal açıdan ele alınmasını gerektirmektedir.

Bu çalışmanın uygulama bölümünde Türkiye'de 2006 yılında en çok işlenen ilk 11 suçtan iller bazında cezaevine giren hükümlü sayılarının şehir nüfusuna oranlı olarak (Türkiye İstatistik Kurumundan alınan verilere göre) iller kümelendirilmeye çalışılmıştır. Bu araştırma sayesinde suçlara göre kümelenen iller benzer özellikleri gösteriyorsa buralarda bu suçlara karşı benzer önlemler almak faydalı olacaktır .

Çeşitli nedenlerle meydana gelen ve sosyoloji, psikoloji, kriminoloji, hukuk gibi farklı bilim dallarınca değişik şekillerde yorumlanan suç, topluma zarar verdiği veya tehlikeli olduğu kanunlarla kabul edilen eylemler (Dönmezer 1994) ya da genel tanımıyla toplum menfaatlerini ihlal eden fiiller olarak tanımlanmaktadır (Seyhan 2002). Ortaya çıkış sebebi, gelişmesi, niteliği gibi özelliklerinden dolayı birçok bilim dalına konu olan suç olayının ana kaynağı insandır. İnsanlar belirli bir mekânda yaşadıklarından ve suçlarında bir mekânda işlenmesinden, suç olayları coğrafyanın bir inceleme konusu olmuştur. Suç araştırmaları, suç ile suçun işlendiği yer arasında anlamlı ilişkiler bulunduğunu ortaya koymuştur (Appiahene-Gyamfi 2002, Karakaş 2004). Suçların farklı dağılış özellikleri göstermesi, bazı yerlerde artarken bazı yerlerde azalması, olayın mekânsal açıdan ele alınmasını gerektirmektedir. Suçların mekânsal dağılımı incelendiğinde gelişigüzel olmadığı açıkca görülebilir. Suçların nerede, niçin ve ne

zaman meydana geldiğini anlamak, hangi suç tipleri hangi bölgelerde meydana gelmiş, demoğrafik değerler ile suçların nedenleri arasındaki ilişki, mekânlar ile suç tipleri arasındaki ilişki nedir gibi soruların yanıtlarını aramak suçla mücadele yöntemlerini geliştirebilir.

Günümüzde birçok ülkede sosyoloji, hukuk, ekonomi, kriptoloji gibi bilim dallarında suç üzerine çeşitli araştırmalar yapılmaktadır. (Karakaş 2004), çalışmasında suçun ekonomik niteliği ve ekonomiye etkisini incelemişlerdir. Diğer bir çalışmada Atasoy (2001) tarafından yapılmıştır. Çalışmada, suç haritalama bilimi ile profesyonelleri tanıştırmak ve suç haritalarının hangi soruları ne şekilde yanıtladığı ile ilgili bilgiler vererek, suçla mücadelede sağlayacağı faydalar ele alınmıştır.

Suç üzerine yapılacak çalışmalarda farklı istatistiki teknikler kullanılabilir, bu çalışmada iller suçlar bakımından kümele analiz yöntemleriyle sınıflandıralacaktır.

Bu çalışmanın amacı kümeleme analizinde küme sayısının belirlenmesi ve küme geçerliliği tekniklerinin araştırılmasıdır. Bu amaçla, kümeleme analizi teknikleri ile Türkiye'deki 81 il, illerde işlenen 11 farklı suç türüne göre incelenecek ve küme geçerliliği indeksleri ile farklı suç türlerine göre en iyi kümelemenin belirlenmesine çalışılacaktır.

Çalışmanın Birinci Bölümünde, kümeleme analizine giriş yapılarak geniş bir literatür taramasına yer verilecektir.

Çalışmanın İkinci Bölümünde kümeleme analizi ile ilgili temel tanım ve kavramlardan bahsedilerek kümeleme analizinin amacı ve kullanım alanları, kümeleme analizinde değişken seçimi, bazı uzaklık ölçüleri ve kümeleme analizi teknikleri incelenecektir. Hiyerarşik kümeleme yöntemlerinden tek bağlantı, tam bağlantı ve Ward yöntemleri, hiyerarşik olmayan kümeleme yöntemlerinden ise k – ortalama yöntemi incelenecektir.

Üçüncü Bölümde kümeleme analizinde küme sayısının belirlenmesi ve küme geçerliliği teknikleri incelenecektir. Bu amaçla çalışmanın kapsamında Silhouette indeksi, Calinski ve Harabazs indeksi, Krzanowski ve Lai indeksi ile Wilk's Lambda İstatistiği incelenmiştir. Bir örnek üzerinde yöntemlerin işleyişine bakılacaktır.

Dördüncü Bölümde, Türkiye'deki 81 il, işlenen 11 farklı suç türüne göre kümeleme analizi ile incelenecektir. İllerin 11 farklı suç türüne göre kümelendirilmesinin yanı sıra, üçüncü bölümde ele alınan küme geçerliliği indeksleri ile farklı kümeleme teknikleri ile en iyi kümeleme yöntemi üzerinde durulacaktır.

2. KÜMELEME ANALİZİ

Kümeleme analizinin çeşitli tanımları mevcuttur. Aşağıda kümeleme analizine ilişkin bazı tanımlar verilmektedir.

Kümeleme analizi, bir araştırmada incelenen birimleri aralarındaki benzerliklerine göre belirli gruplar içinde toplayarak sınıflandırma yapmayı, birimlerin ortak özelliklerini ortaya koymayı ve bu sınıflar ile ilgili genel tanımlar yapmayı sağlayan bir yöntemdir (Kaufman and Rousseuw 1990).

Kümeleme analizi, elde bulunan veri yığınını belirlenen yöntemlerle analiz ederek daha önceden etiketleri belli olmayan gruplara ayrılması işlemidir. Bu işlem sonucunda elde edilen kümeler yüksek düzeyde küme içi homojenlik ve kümeler arası heterojenlik gösterirler (Kantardzic 2003).

Kümeleme analizi, birey veya nesneleri benzerliklerine göre kümelere veya gruplara ayırmak için kullanılan bir çok değişkenli istatistik analiz tekniğidir (Tatlıdil 1996). Kümeleme analizi sonucu oluşturulan kümeler içinde aynı küme içinde yer alan birimler birbirlerine diğer kümenin içinde yer alan birimlerden daha çok benzeşirler.

X birimler üzerinden alınan gözlem değerlerinden oluşan veri matrisi olmak üzere kümeleme analizi, veri matrisini oluşturan ve doğal gruplamaları kesin olarak bilinmeyen birimleri birbirleri ile benzer olanları alt kümelere ayırmaya yardımcı olan yöntemler topluluğudur (Romesburg 1984).

Benzer birimlerin sahip oldukları karakteristiklerden yola çıkarak tanımlama amacını güden kümeleme analizi sonucunda, küme içindeki her birim önceden belirlenmiş bir kritere göre yine küme içindeki diğer birimlere çok benzer. Böylece oluşan kümelerde kümeler içi yüksek homojenlik ve kümeler arası yüksek heterojenlik sağlanır. Sınıflandırma başarılı olursa, geometrik olarak kümeler grafiğe yerleştirildiğinde küme içi birimler birbirine çok yakın iken, farklı kümelerdeki birimler birbirlerinden çok uzakta olacaktır.

Sosyal bilimler, tıp, ziraat başta olmak üzere tüm mühendislik bilimlerinde yaygın uygulama alanı bulunan kümeleme analizi, çok değişkenli varyans analizi, lojistik regresyon analizi, çok boyutlu ölçekleme gibi diğer çok değişkenli analizlerle de sıkı ilişkisi olan bir tekniktir.

Kümelemeyle günlük yaşantımızda sık sık karşılaşılabilir. Örneğin bir restorantta aynı masada oturan insanlar bir küme olarak sayılabilir. Bu tür örnekler çoğaltılabilir. Biyologların farklı hayvan türleri arasında anlamlı bir tanımlama yapmadan önce, hayvan türlerini doğru sınıflandırılması gerekir. Kısaca işin niteliği ne olursa olsun bir araştırmada ya da diğerinde bir kümeleme problemiyle karşılaşmak kaçınılmazdır (Everitt 1974).

2.1 Kümeleme Analizinin Genel Amacı

Kümeleme analizinin genel amacı, toplanan çok sayıdaki gruplanmamış gözlemlerden oluşan veriyi birimlerin benzerliklerine göre anlamlı gruplardan oluşan özel alt kümelere bölerek veriyi indirgemektir. Böylece araştırmacı en az bilgi kaybıyla, daha net ve anlaşılabilir tanımlı gözlemlere sahip olacaktır. Kümeleme analizi ilk kez 1939 yılında Tryon tarafından kullanılmıştır. 1960'lı yıllardan sonra kullanımı yaygınlaşmıştır. 1963 yılında Robert Sokal ve Peter Sneath'ın yazdığı "Sayısal Sınıflandırma İlminin Temelleri" adlı kitap bu alanda önemli bir adım olmuştur (Anderberg 1973).

Genel amacın yanı sıra kümeleme analizi için aşağıdaki özel amaçlardan da bahsedilebilir (Everitt 1974). Bu özel amaçlar;

1. Gerçek tipleri (cinsleri-ırkları) belirlemek
2. Gruplar için ön tahmin yapmak
3. Hipotezlerin testi
4. Veri yapısını netleştirmek

5. Veri indirgemek (Veriler yerine kümelerin değerlendirilmesi)

6. Aykırı değerlerin bulunması

biçiminde sıralanabilir.

Birçok araştırma alanında araştırmacıların karşılaştığı en genel sorun, gözlenen verileri anlamlı olarak nasıl organize (sınıflandırılacağı) edileceğidir. Bunun için değişik sınıflandırma teknikleri geliştirmek gerekmektedir. Başka bir ifadeyle, kümeleme analizinin en genel amacı, aynı gruba ait olan verilerin arasında maksimum benzerliğin olması, diğer gruplardaki verilerle minimum benzerlikte olmasını sağlamaktır.

Kümeleme teknikleri, araştırma problemlerinde geniş bir alanda uygulanır (Hartigan, 1975). Tıpta hastalıkları, hastalık tedavilerini sınıflandırmada kümeleme çok kullanışlı bir yöntemdir. Böylelikle "herhangi bir hastalık hangi grupta, o grubun belirtileri nelerdir ve tedavi yöntemleri nelerdir?" bunları bilmek tedavi sürecinde fayda sağlar. Tıbbın psikolojiyle ortak alanı olan psikiyatri dalında, paranoya, şizofren, manik depresif gibi önemli hastalıkların belirtilerinin doğru kümelerde teşhis edilmesi, doğru tedavi için gereklidir. Sosyal bilimlerde de kümele analizi teknikleri uygulanmaktadır. Örneğin suç istatistiklerinde suçları sınıflandırmada kullanılabilir. Kümeleme analizinin en son uygulama alanı ise veri madenciliğidir, veri madenciliğinde büyük veri yığınlarından kurtulup özet veriler elde etmek hedeflenir. Bu aşamada kümeleme analizi kullanılır.

2.2 Kümeleme Analizinin Uygulama Aşamaları

Kümeleme Analizinin uygulama aşamaları aşağıdaki gibidir;

a) Değişkenlerin seçilmesi ve veri matrisinin belirlenmesi

b) Birimlerin birbirleriyle olan benzerlik ya da uzaklıklarını gösterecek uygun bir benzerlik/uzaklık ölçüsü ile benzerlik/uzaklık matrisinin oluşturulması

c) Uygun bir kümeleme tekniği ile benzerlik/uzaklık matrisine göre birimlerin uygun sayıda kümelere ayrılması

d) Oluşturulan bu kümelerin yorumlanması, kümelerin yapılarının kurulan hipotezlerle test edilmesi ve gerekli analitik yöntemlerin uygulanması.

2.3 Kümeleme Analizinde Değişken Seçimi ve Uzaklık Ölçüleri

Kümeleme analizinde ilk adım değişken seçimi ve uygun uzaklık ölçüsünün belirlenmesidir.

2.3.1 Değişken seçimi

Kümeleme analizinde değişken kavramı çok önemlidir ve diğer çok değişkenli analiz yöntemlerinden farklıdır. Kümeleme analizinde değişkenler, gösterdikleri özellikler kullanılarak karşılaştırılır. Kümeleme analizi değişkeni, sadece nesnelerin tanımlanan özelliklerini kapsamaktadır.

2.3.2 Uzaklık ölçüleri

Benzerlikler, birimleri gruplamak ya da ayırmak için kullanılan bir takım kurallardır. Benzerlik ya da uzaklık ölçüleri tek boyutlu veya çok boyutlu yapılabilir. Her boyut birey ya da nesneleri gruplandırmak için kullanılır. Örneğin fast-food yiyecekler gruplandırmak istensin. Bunun için fast-food ürünlerinin her birinin sahip olduğu kalori miktarı, fiyatları, tatları göz önüne alınmalıdır. Çok boyutlu uzayda iki birey ya da nesne arasındaki uzaklığı hesaplamada en çok kullanılan uzaklık ölçüsü, öklid uzaklığıdır. Eğer iki veya üç boyutlu uzayda çalışılıyorsa bu ölçüm, basit olarak uzayda iki nesnenin arasındaki geometrik uzaklıktır. Bir sonraki alt kesimlerde çeşitli uzaklık ölçüleri verilecektir.

2.3.2.1 Öklid uzaklığı

Öklid uzaklığı en sık kullanılan uzaklık ölçüsüdür. Basit olarak çok boyutlu uzayda geometrik uzaklıktır ve

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2} \tag{2.1}$$

biçiminde hesaplanır (Tatlıdil 1996), burada

d_{ij} ; i. ve j. birimin birbirine olan uzaklığı

x_{ik} ; i. birimin k. değişken değeri

x_{jk} ; j. birimin k. değişken değeri

$i = 1, \ldots, n$; $j = 1, \ldots, n$ ve $k = 1, \ldots, p$ 'dir. n birim ve p değişken sayısıdır.

Karesel Öklid uzaklığı ise

$$d_{ij}^2 = \sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2 \tag{2.2}$$

biçimindedir (Tatlıdil 1996).

2.3.2.2 Minkowski uzaklığı

Uzaklıkların belirlenmesinde bir diğer kullanılan ölçüde Minkowski olarak bilinen uzaklık ölçüsüdür. Minkowski uzaklık ölçüsü

$$d_{ij} = \left[\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^q \right]^{1/q} \tag{2.3}$$

olarak tanımlanmıştır. Minkowski uzaklık ölçüsü $q = 1$ için bir sonraki kesimde verilecek olan City-Block uzaklık ölçüsüne, $q = 2$ için ise öklid uzaklık ölçüsüne eşit olacaktır. Belirtilmesi gereken bir durumda formülünden de görüleceği gibi Minkowski

uzaklık ölçüsü genel bir uzaklık ölçüsü, Öklid ve City-Block uzaklık ölçüleri ise Minkowski uzaklık ölçüsünün özel bir durumudur (Anderberg 1973).

2.3.2.3 City-Block(Manhattan) uzaklığı

City-Block uzaklık ölçüsü, birimler arasındaki mutlak uzaklıkların toplamını alarak hesaplayan bir uzaklık ölçüsüdür ve

$$d_{ij} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right| \tag{2.4}$$

biçiminde ifade edilir.

City-Block uzaklık ölçüsü uygulamada bazı sorunlara yol açmaktadır. Bu sorunlardan en belirgini City-Block uzaklık ölçüsünün değişkenler arasında korelasyon (ilişki) olmadığını varsaymasıdır. Eğer araştırma konusunda değişkenler arasında korelasyon varsa City-Block uzaklık ölçüsüyle hesaplanan uzaklık ölçüleri baz alınarak yapılan kümeleme anlamlı olmayacaktır. Sorunlardan bir diğeri de ölçüm yapılan değişkenlerin birimleri farklı olması durumunda standartlaştırılmış Karesel Öklid uzaklığıyla karşılaştırıldığında City-Block uzaklık ölçüsünün anlamlı sonuçlar vermediği görülecektir (Johnson and Wichern 1988).

2.3.2.4 Mahalanobis uzaklığı

Kullanılan diğer bir uzaklık ölçüsü de, doğrudan birleştirme yapan, standart bir yöntem olan Mahalonobis Uzaklık ölçüsüdür. İki değişken arasında bir ilişki mevcut ise, bu iki değişken arasındaki kovaryans veya korelasyonu göz önüne alan Mahalonobis uzaklığının kullanılması gerekmektedir.

p değişkenli bir analizde i ve k gözlemleri arasındaki Mahalonobis uzaklık ölçüsü;

$$Md_{ik} = \left(x_i - x_k \right)' S^{-1} \left(x_i - x_k \right) \tag{2.5}$$

biçimindedir, burada yer alan S, $p \times p$ tipinde örneklem kovaryans matrisini göstermektedir (Sharma 1996). Mahalonobis uzaklığının avantajı, aykırı noktalarıda hesaplamasıdır. Bu yönleriyle Mahalonobis uzaklığı, uzaklık ölçüleri arasında en avantajlı olanıdır denilebilir.

2.4 Kümeleme Analizi Teknikleri

Kümeleme analizinde, uygun uzaklık ölçüsü seçildikten sonraki aşama, hangi kümeleme analizi tekniğinin seçileceğidir.

Araştırmacı hangi benzerlik/uzaklık ölçüsünü kullanacağına karar verdikten sonra, kümeleme işleminin nasıl olacağına karar vermek zorundadır. Birimlerin benzerliklerine göre kümelere dâhil edilmesinde kullanılabilecek çeşitli yaklaşımlar vardır. Bu yaklaşımlardan biri, en çok benzer iki birimi aynı gruba atamakla başlayıp tüm birimlerin aynı gruba atanması ile biten hiyerarşik bir yaklaşımdır. Bir başka yaklaşım ise tüm verilerin ortalama değerlerine en yakın değerlere sahip birimlerin aynı kümeye atanmasını esas alan yaklaşımdır. En çok kullanılan bu iki yaklaşım dışında diğer yaklaşımlar da mevcuttur. Tüm yaklaşımlarda en önemli ölçüt, kümeler arası farklar ile kümeler içi benzerliklerin maksimum olmasını sağlamaktır. En çok kullanılan kümeleme algoritmaları hiyerarşik ve hiyerarşik olmayan kümeleme adı altında iki kategoride toplanmaktadır (Blashfield and Aldenferder 1978).

2.4.1 Hiyerarşik Kümeleme Yöntemleri

Yöntem, aşama sıralı kümeleme yöntemi olarak da bilinir. Gruplayıcı ve bölücü olmak üzere iki hiyerarşik yöntem mevcuttur (Hubert 1974). Gruplayıcı hiyerarşik yöntemde her birim veya her gözlem başlangıçta bir küme olarak kabul edilir. Daha sonra en yakın iki küme (veya gözlem) yeni bir kümede toplanarak birleştirilir. Böylece her adımda küme sayısı bir azaltılır. Bu süreç dendogram veya ağaç grafiği adı verilen şekille gösterilebilir.

Bölücü hiyerarşik yöntemde ise süreç gruplayıcı hiyerarşik yöntemin tam tersidir. Bu yöntemde tüm gözlemlerden oluşan büyük bir küme ile işe başlanır. Benzer olmayan gözlemler ayıklanarak daha küçük kümeler oluşturulur. Her gözlem tek başına küme oluşturana kadar işleme devam edilir (Everitt *et al*. 2001). Uygulamalarda çoğunlukla gruplayıcı hiyerarşik kümeleme yöntemi kullanılmaktadır. Gruplayıcı hiyerarşik yöntemler arasında en çok tek bağlantılı, tam bağlantılı, grup ortalama yöntemi ve Ward yöntemi kullanılmaktadır. Bu yöntemler sırasıyla aşağıdaki kesimlerde tanıtılacaktır.

2.4.1.1 Tek Bağlantı Tekniği

En yakın komşuluk olarak da bilinen tek bağlantı tekniği, uzaklıklar matrisini kullanarak birbirine en yakın (uzaklık değerleri en küçük) birey ya da nesneleri birleştirmeye dayanmaktadır (Johnson and Wichern 1988). Bu teknikte önce birbirine en yakın iki birim (gözlem) bir kümeye yerleştirilir. Daha sonra diğer en yakın uzaklık tespit edilerek ilk oluşturulan kümeye bu gözlem eklenir veya iki gözlemden oluşan yeni bir küme oluşturulur. İşlem tüm gözlemlerin bir kümeye yerleştirilmesine kadar devam eder.

Bu teknikte eğer i ve j nci birimler birleştirilmiş ise birleştirilen kümenin k ıncı küme ile ilişkisi uzaklık ölçütü olarak,

$$d_{k(i,j)} = Min\left(d_{ki}, d_{kj}\right) \tag{2.6}$$

biçiminde ifade edilmektedir.

Eşitlikte;

$d_{k(i,j)}$; k.kümenin daha önce oluşan i. ve j. kümelerle olan uzaklığını,

d_{kj} ; k'ıncı kümenin j'inci kümeye olan uzaklığını,

d_{ki} ; k'ıncı kümenin i'nci küme ile olan uzaklığını göstermektedir.

2.4.1.2 Tam Bağlantı Tekniği

Bu yöntem, en uzak komşuluk olarak da bilinmektedir. Tek bağlantı tekniğine çok benzemekle birlikte bu teknikteki tek farklılık her kümedeki eleman çiftleri arasındaki uzaklığın maksimum olanının ele alınmasıdır. Bu tekniğe tam bağlantı tekniği denmesinin nedeni, bir küme içindeki tüm birimlerin birbirlerine maksimum uzaklık veya minimum yakınlığa bağlı olmasıdır (Green 1989). Tam bağlantı tekniğindeki uzaklıklar,

$$d_{k(i,j)} = Max\left(d_{ki}, d_{kj}\right) \tag{2.7}$$

biçiminde gösterilmektedir.

2.4.1.3 Ortalama Bağlantı Tekniği

Bu teknikte de işleme tek bağlantı ve tam bağlantı tekniklerinde olduğu gibi başlanır. Ancak kümeleme kriteri olarak bir küme içindeki birim ile diğer küme içindeki birimler arasındaki ortalama uzaklıklar kullanılır. Ortalama bağlantı tekniğinde kümeler küçük varyanslar ile birbirlerine bağlıdır. Bu teknik tek bağlantı ve tam bağlantı teknikleri arasında sonuçlar vermesi nedeniyle bir alternatif yöntem olarak önerilmektedir (Hubert 1974).

2.4.1.4 Ward Yöntemi

Ward yönteminde**,** grup bağlantılarından çok grup içi kareler toplamı işlenmektedir (Chatfield and Colins 1980). Yönteme her birinin içinde tek bir birim bulunan n tane küme ile başlanır. Yöntemin ilk basamağında her gözlem bir küme olduğundan Hata Kareler Toplamı sıfır olmaktadır (Everitt 1974). Her aşamada iki alt küme bir sonraki seviyeyi oluşturmak için birleştirilir. Bu durumda k(k-1) alt grup olduğu varsayılır. k kümesinde yer alan n_i noktanın k kümesinin ortalamalar vektörüne olan öklid uzaklıkları toplamı,

hata kareler toplamıdır ve W_k olarak ifade edilir;

$$W_k = \sum_{i=1}^{p}\sum_{j=1}^{n_k}(x_{ijk} - \bar{x}_{ik})^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{n_k} x_{ijk}{}^2 - n_k \sum_{i=1}^{p} \bar{x}_{ik}{}^2 \qquad (2.8)$$

biçiminde hesaplanır. Burada W_k değeri k=1,2,...,n kümelerde hesaplanarak küme içi hata kareler toplamı,

$$W = \sum_{i=1}^{n} W_k \qquad (2.9)$$

şeklinde hesaplanır ve W 'de en küçük artışa sahip olan p ve q kümeleri birleştirilerek t kümesi elde edilir. W ' deki bu artış;

$$DW_{pq} = W_t - W_P - W_q \qquad (2.10)$$

eşitliği ile hesaplanır. Böylece n birim (n-1) kümeye ayrılmış olur. Böylelikle küme sayısı k=1 oluncaya kadar W artış değerleri bulunarak birimlerin aşamalı biçimde birbirlerine bağlanmaları sağlanır.

Analiz sonucunda Ward Yöntemi kullanıldığında, birimler değişik seviyelerde başarılı bir şekilde bir araya geldikleri "dendogram" adlı şemada gösterilir (Dibb 1998). Ward yöntemi aykırı noktalara duyarlı bir yöntemdir (Everitt *et al.* 2001).

2.4.2 Hiyerarşik olmayan kümeleme yöntemleri

Bazı durumlarda küme sayısı önceden bellidir ve araştırmacı bu küme sayısına göre çözümler üretmek durumundadır. Küme sayısı konusunda ön bilgi varsa veya araştırmacı anlamlı olacak küme sayısına karar vermiş ise bu durumda, çok uzun zaman alan hiyerarşik yöntemler yerine hiyerarşik olmayan yöntemler kullanılmaktadır (Anderberg 1973). Hiyerarşik olmayan yöntemlerde prosedür şu şekildedir; ilk kısım veya belirli bir başlama noktası ile başlanır. Şayet ilk kısım ile başlanırsa çalışılan ana yığından belirli bir örneklem seçilir ve ilk kısmı elde edebilmek için küme üyeleri düzeltilir.

Hiyerarşik olmayan kümeleme yöntemlerinin temel dezavantajı küme sayılarının daha önceden belirlenmesi ve küme seçimlerinin keyfi olmasıdır (Blashfield and Aldenferder 1978).

Birimlerin birbirleri ile olan benzerliklerini benzer grup ortalama vektörleri tanımlayarak kendi içinde homojen ve aralarında heterojen gruplamaları belirlemeyi amaçlayan kümeleme yöntemlerine hiyerarşik olmayan kümeleme yöntemleri denilmektedir (Hartigan 1975). Hiyerarşik olmayan kümeleme yöntemleri başlığı altında bir çok teknikten söz edilebilir ancak bunlardan en sık kullanılan iki tanesi k-ortalama yöntemi ve en çok olabilirlik yöntemidir, bu yöntemler alt kesimlerde açıklanacaktır.

2.4.2.1 k – Ortalama yöntemi

MacQueen "k – ortalama" terimini her bir birimin en yakın merkezli (ortalama) kümeye atanması süreci anlamında kullanmıştır. k – ortalama tekniği, gözlemleri kümelerin önceden belirlenen sayısına göre gruplandırmakla işleme başlar. Böylece her biri tek gözlemden oluşan k tane küme ile işleme başlanır ve her bir yeni gözlem en yakın ortalamalı gruba eklenir. Gruba yeni bir gözlem eklendikten sonra küme ortalaması yeniden hesaplanır. Bu süreç tüm gözlemler gruplara atanıncaya kadar devam eder. Tüm gözlemler gruplara atandıktan sonra atandıkları küme ortalamasından daha yakın küme ortalaması varsa, gözlemlerin yerleri değiştirilmektedir. Amaç diğer kümeleme yöntemlerinde olduğu gibi, gerçekleştirilen kümeleme işlemi sonucunda elde edilen kümelerin, küme içi benzerliklerinin maksimum, kümeler arası benzerliklerinin ise minimum olmasını sağlamaktır. Küme benzerliği, kümenin ağırlık merkezi kabul edilen bir birim ile kümedeki diğer birimler arasındaki uzaklıkların ortalama değeri ile ölçülmektedir (Han and Kamber 2001).

Yöntemin işleyişi aşağıda özetlenmiştir:

Adım 1: İlk k gözlem, her biri bir elemanlı küme olarak alınır ve bunların her biri birer küme ortalaması olarak kabul edilir. Tüm birimlerin küme ortalamalarına olan uzaklıkları hesaplanır.

Adım 2: Geriye kalan $(n-k)$ birimin her biri en yakın küme ortalaması olan kümeye atanır. Her atamadan sonra küme ortalaması yeniden hesaplanır.

Adım 3: Bütün birimler 2.adımda k kümeye atandıktan sonra küme ortalamaları yeni çekirdek nokta olarak alınır ve en yakın ortalamaya göre atama işlemi küme elemanları yerlerinin değişmez olmasına kadar tekrarlanır.

2.4.2.2 En çok olabilirlik yöntemi

Diskriminant analizinde de kullanılan en çok olabilirlik yönteminde her bir gözlem en büyük olabilirlik değerini verecek biçimde daha önceden belirlenen kümelere atanır. Kuramsal dayanağı güçlü olmakla birlikte en çok olabilirlik yöntemi yaygın olarak kullanılmamaktadır (Tatlıdil 1996).

3. KÜME SAYISININ BELİRLENMESİ ve KÜME GEÇERLİLİĞİ TEKNİKLERİ

Bu bölümde kümeleme analizinde küme sayısına karar verme ve oluşan kümelemenin kalitesini değerlendirmek için küme geçerliliği (cluster validation) teknikleri tanımlanacaktır.

3.1 Küme Sayısına Karar Verme

Kümeleme Analizinin en kritik konusu küme sayısına karar vermektir. Araştırmacının küme sayısına karar vermede öznelliği minimize etmesi gerekmektedir. Ancak günümüzde yayınlanan birçok makalede bu konuda kesin bulunmuş sonuçlar yoktur. İlk önerilen yaklaşımlardan en çok bilinen eşitlik,

$$k = \left(n/2\right)^{1/2} \tag{3.1}$$

biçiminde hesaplanmaktadır. Burada k küme sayısı, n birim sayısını göstermektedir. Küçük örneklemli araştırmalarda kullanılması tavsiye edilir. Büyük örneklemli araştırmalarda kullanılması durumunda sağlıklı sonuçlara ulaşılması zorlaşır (Everitt 1974).

Diğer bir yöntem ise Mariott tarafından önerilmiştir. Bu yüzden M harfi ile gösterilir.

$$M = k^2 \left|W\right| \tag{3.2}$$

biçimindedir (Marriot 1971). Burada W , grup içi kareler toplamı matrisidir ve

$$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)(X_{ij} - \bar{X}_j)' \tag{3.3}$$

n_j ; j. kümedeki birim sayısı

k ; küme sayısı

X_{ij} ; j. kümedeki i. birim değerleri

$\overline{X}_{j;}$; j. kümenin örneklem ortalama vektörü

biçiminde hesaplanır. M değerini minimum yapan k değeri uygun küme sayısı olarak alınmaktadır. Everitt (1974) farklı ölçümlerle ilgili yaptığı çalışmada, k^2W ifadesinin diğer yöntemlere göre daha iyi sonuç verdiğini ortaya koymuştur.

Calinski ve Harabasz (1974) tarafından önerilen bir diğer kriter :

$$C = \frac{trB}{k-1} / \frac{trW}{n-k} \tag{3.4}$$

şeklindedir. Burada W , (3.3)'de verilmiştir. B ise gruplar arası kareler ve çarpımlar toplamı matrisini ifade eder ve

$$B = \sum_{j=1}^{k} n_j (\overline{X}_j - \overline{X})(\overline{X}_j - \overline{X})' \tag{3.5}$$

biçiminde hesaplanır.

$C = \frac{trB}{k-1} / \frac{trW}{n-k}$, değerini maksimum yapan k değeri uygun küme sayısını verir (Seber 1984).

Küme sayısına karar vermede kullanılan bir diğer yaklaşım ise Lewis ve Thomas tarafından literatüre katılmıştır. Lewis ve Thomas' a göre küme sayısına karar vermek için iki kriter vardır. Kümeler, toplam varyansın %80'ini açıklamalı ve varyansta %5'e kadarlık bir artış durumunda yeni bir küme ilave edilebilmektedir (Ruiz 1998).

Küme sayısının belirlenmesinde kullanılan ölçütlerden Wilk's Lamda ölçütü, birim sayısı 30'un üzerinde olduğunda duyarlılığı diğer ölçütlere göre en yüksek düzeye ulaşmaktadır (Dinçer ve Özdamar 1992). Everitt (1979) ve Anderberg (1973) gruplararası varyansın grup içine göre maksimum olduğu durumları belirleme açısından Wilk's Lamda ölçütünün çok duyarlı olduğunu belirtmişlerdir .

Kümeleme analizi yöntemleri, oluşan kümelenmeleri doğrulamak için çok değişkenli bir yöntem olan Temel Bileşenler Analizinden faydalanabilir. Temel Bileşenler Analizi

yönteminin kullanılmasıyla oluşturulan Temel bileşen skoru grafiğinden yorumlar yapılarak, bu grafik yardımıyla kümeleme analizi sonuçlarının doğruluğu denetlenebilir. Küme sayısına karar verme teknikleri literatüre katılmış olsa da araştırmacı, araştırma sonucu oluşan küme yapılarını incelemeli ve oluşan kümeleri farklı hipotez testleriyle test etmelidir.

Küme sayısına karar verdikten sonra oluşan küme yapılarının kalitesini değerlendirmek için (cluster validation) küme geçerliliği teknikleri alt kesimlerde tanımlanacaktır.

3.2 Küme Geçerliliği Teknikleri

Kümeleme analizi uygulamalarında doğru küme sayısı çoğunlukla bilinemez. Kümeleme analizinin sonuçlarının kalitesini değerlendirmek için küme geçerliliği tekniklerine ihtiyaç vardır. Bu teknikler arasında en çok kullanılanları Silhouette indeksi, Calinski ve Harabazs indeksi, Krzanowski ve Lai indeksi olarak sayılabilir. Bu indeksler küme içi değişim ve kümeler arası değişim ölçüleri arasındaki ilişkilere dayanmaktadır. İyi bir kümeleme ile aynı kümede yer alan birimler arasındaki benzerliğin olabildiğince fazla, farklı kümelerde yer alan birimler arasındaki benzerliğin ise olabildiğince az olması amaçlanmaktadır. Küme geçerliliği teknikleri küme içi ve kümeler arası değişimlerin farklı açılardan değerlendirilmesi ile birbirinden ayrılmaktadır (Bolshakova And Azuaje). Bu çalışmada bu indekslerin yanı sıra çok değişkenli hipotezlerin değerlendirilmesinde göz önüne alınan Wilks Lamda ölçütü de küme sayısının belirlenmesinde kullanılacaktır. Küme geçerliliği teknikleri hipotetik olarak alınan bir örnek üzerinde incelenecektir.

3.2.1 Silhouette İndeksi

Rousseeuw (1987), her bir birimin kendi kümesine uygunluğunu tanımlayacak bir Silhouette indeksi önermiştir. $a(i)$; $i.$ birimin kendi kümesindeki tüm noktalara olan ortalama uzaklıklarını (benzerliğini) ve $b(i)$; $i.$ birimin diğer kümelerdeki tüm

noktalara olan ortalama uzaklıkların minimumunu göstersin. Buradan *i.* birim için Silhouette indeksi;

$$sil(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)} \tag{3.6}$$

olarak tanımlanır. Eğer $sil(i)$ değeri 1'e yaklaşırsa *i.* birimin atandığı kümeye daha uyduğu, $sil(i)$ değeri 0'e yaklaşırsa veya negatif olursa *i.* birimin atandığı kümeye uygun olmadığı sonucuna varılır. Negatif değerler yalnızca bir birim en uygun kümesine atanamadığında ortaya çıkar. Tüm kümelemenin kalitesi (geçerliliği) için bir doğal bir ölçü;

$$sil(C) = \frac{1}{n}\sum_{s_i \in S} sil(i) \tag{3.7}$$

tüm birimler için ortalama Silhouette değeri olarak tanımlanabilir. Bu kritere göre, maksimum ortalama Silhouette değerine ulaşılan küme sayısı uygun küme sayısı olarak alınır (Rousseeuw 1987).

10 birim ve 3 değişkenli bir düşüncesel örnek dikkate alınsın. $k-$ ortalama tekniği ile 'k' küme sayısı olmak üzere; $k = 2,3,4$ için kümeleri oluşturduğu varsayılsın. Birimlere ilişkin değerler ve kümeleme sonuçları Çizelge 3.1'de özetlenmektedir.

Çizelge 3.1 Hipotetik örneğe ilişkin değerler

Birim	x_1	x_2	x_3	$k=2$	$k=3$	$k=4$
A1	20	20	18	2	1	3
A2	16	24	21	1	2	3
A3	18	31	25	1	3	4
A4	13	22	26	1	2	1
A5	18	24	17	2	1	3
A6	18	23	25	1	2	1
A7	29	23	19	2	1	2
A8	24	25	18	2	1	2
A9	20	26	23	1	3	4
A10	19	20	25	1	2	1

Birimler arasındaki Öklid uzaklıkları Çizelge 3.2'de özetlenmiştir.

Çizelge 3.2 Hipotetik örnekteki birimler arasındaki öklid uzaklıkları

Birim	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
A1	0.00	6.40	13.19	10.82	4.58	7.87	9.54	6.40	7.81	7.07
A2	6.40	0.00	8.31	6.16	4.47	4.58	13.19	8.60	4.90	6.40
A3	13.19	8.31	0.00	10.34	10.63	8.00	14.87	11.00	5.74	11.05
A4	10.82	6.16	10.34	0.00	10.49	5.20	17.49	13.93	8.60	6.40
A5	4.58	4.47	10.63	10.49	0.00	8.06	11.22	6.16	6.63	9.00
A6	7.87	4.58	8.00	5.20	8.06	0.00	12.53	9.43	4.12	3.16
A7	9.54	13.19	14.87	17.49	11.22	12.53	0.00	5.48	10.30	12.04
A8	6.40	8.60	11.00	13.93	6.16	9.43	5.48	0.00	6.48	9.95
A9	7.81	4.90	5.74	8.60	6.63	4.12	10.30	6.48	0.00	6.40
A10	7.07	6.40	11.05	6.40	9.00	3.16	12.04	9.95	6.40	0.00

$k = 2$ durumunda, tüm birimler için Silhouette değerini hesaplansın.

Çizelge 3.1 incelendiğinde $k = 2$ durumu için;

(A1,A5,A7,A8) birimleri kendi aralarında bir küme ve (A2,A3,A4,A6,A9,A10) birimleri ise diğer kümeyi oluşturduğu kabul edilsin.

$a(i)$; i. birimin kendi kümesindeki tüm noktalara olan ortalama uzaklıkları (benzerliğini) olmak üzere;

$$a(A1) = \frac{1}{3}(4.58 + 9.54 + 6.40) = \frac{1}{3}(20.52) = 6.84$$

$b(i)$; i. birimin diğer kümelerdeki tüm noktalara olan ortalama uzaklıkların minimumu olmak üzere;

$$b(A1) = \frac{1}{6}(6.40 + 13.19 + 10.82 + 7.87 + 7.81 + 7.07) = \frac{1}{6}(53.16) = 8.86$$

olmak üzere A1 birimi için Silhouette değeri eşitlik (3.6) 'dan

$$sil(A1) = \frac{8.86 - 6.84}{\max(6.84, 8.86)} = \frac{2.02}{8.86} = 0.228$$

biçiminde elde edilir. Aynı şekilde

$$a(A2) = \frac{1}{5}(8.31 + 6.16 + 4.58 + 4.90 + 6.40) = \frac{1}{5}(30.35) = 6.07$$

$$b(A2)=\frac{1}{4}(6.40+4.47+13.19+8.60)=\frac{1}{4}(32.66)=8.165$$

olmak üzere A2 birimi için Silhouette değeri

$$sil(A2)=\frac{8.165-6.07}{\max(6.07,8.165)}=\frac{2.095}{8.165}=0.256$$

biçiminde elde edilir. Benzer biçimde A3, A4,...,A10 birimleri için Silhouette değerleri;

$$sil(A3)=\frac{12.423-8.69}{\max(8.69,12.423)}=\frac{3.733}{12.423}=0.300$$

$$sil(A4)=\frac{13.183-7.34}{\max(7.34,13.183)}=\frac{5.843}{13.183}=0.443$$

$$sil(A5)=\frac{8.213-7.32}{\max(7.32,8.213)}=\frac{0.893}{8.213}=0.108$$

$$sil(A6)=\frac{9.473-5.012}{\max(5.012,9.473)}=\frac{4.461}{9.473}=0.470$$

$$sil(A7)=\frac{13.403-8.746}{\max(8.746,13.403)}=\frac{4.657}{13.403}=0.347$$

$$sil(A8)=\frac{9.898-6.013}{\max(6.013,9.898)}=\frac{3.885}{9.898}=0.392$$

$$sil(A9)=\frac{7.805-5.952}{\max(5.952,7.805)}=\frac{1.853}{7.805}=0.237$$

$$sil(A10)=\frac{9.515-6.682}{\max(6.682,9.515)}=\frac{2.833}{9.515}=0.298$$

elde edilir. Buradan $k=2$ durumunda Silhouette değeri eşitlik (3.7)' den

$$sil(k=2)=\frac{1}{10}(0.228+0.256+0.300+\ \ldots\ +0.298)=\frac{1}{10}(3.079)=0.3079$$

olarak hesaplanır.

$k = 3$ durumunda, tüm birimler için Silhouette değerini bulmak için Çizelge 3.1'de yer alan birimlerden (A1,A5,A7,A8) 1. kümenin birimleri, (A2,A4,A6,A10) 2. kümenin birimleri ve (A3,A9) ise 3. kümenin birimleri biçiminde belirlensin.

$$a(A1) = \frac{1}{3}(4.58 + 9.54 + 6.40) = \frac{1}{3}(20.52) = 6.84$$

$$b(A1) = \min\left\{\frac{1}{4}(6.40 + 10.82 + 7.87 + 7.07), \frac{1}{2}(13.19 + 7.81)\right\} = \min\left\{\frac{1}{4}(32.16), \frac{1}{2}(21)\right\}$$

$$= \min\{8.04, 10.5\} = 8.04$$

olmak üzere A1 birimi için silhouette değeri;

$$sil(A1) = \frac{8.04 - 6.84}{\max(6.84, 8.04)} = \frac{1.2}{8.04} = 0.149$$

ve

$$a(A2) = \frac{1}{3}(6.16 + 4.58 + 6.40) = \frac{1}{3}(17.14) = 5.713$$

$$b(A2) = \min\left\{\frac{1}{4}(6.40 + 4.47 + 13.19 + 8.60), \frac{1}{2}(8.31 + 4.90)\right\} = \min\left\{\frac{1}{4}(32.66), \frac{1}{2}(13.21)\right\}$$

$$= \min\{8.165, 6.605\} = 6.605$$

olmak üzere A2 birimi için Silhouette değeri

$$sil(A2) = \frac{6.605 - 5.713}{\max(5.713, 6.605)} = \frac{0.892}{6.605} = 0.135$$

olarak elde edilir. Benzer biçimde A3, A4,..., A10 birimleri için Silhouette değerleri;

$$sil(A3) = \frac{9.425 - 5.74}{\max(5.74, 9.425)} = \frac{3.685}{9.425} = 0.391$$

$$sil(A4) = \frac{9.47 - 5.92}{\max(5.92, 9.47)} = \frac{3.55}{9.47} = 0.375$$

$$sil(A5) = \frac{8.005 - 7.32}{\max(7.32, 8.005)} = \frac{0.685}{8.005} = 0.085$$

$$sil(A6) = \frac{6.06 - 4.313}{\max(4.313, 6.06)} = \frac{1.747}{6.06} = 0.288$$

$$sil(A7) = \frac{13.8125 - 8.747}{\max(8.747, 13.8125)} = \frac{5.066}{13.8125} = 0.366$$

$$sil(A8) = \frac{8.74 - 6.013}{\max(6.013, 8.74)} = \frac{2.727}{8.74} = 0.312$$

$$sil(A9) = \frac{6.005 - 5.74}{\max(5.74, 6.005)} = \frac{0.265}{6.005} = 0.044$$

$$sil(A10) = \frac{8.725 - 5.32}{\max(5.32, 8.725)} = \frac{3.405}{8.725} = 0.390$$

biçiminde elde edilir. Buradan, $k = 3$ için ortalama Sillhouette değeri

$$sil(k = 3) = \frac{1}{10}(0.149 + 0.135 + 0.391 + \ \ldots\ + 0.390) = \frac{1}{10}(2.535) = 0.2535$$

olarak hesaplanır.

$sil(k = 2)$ ve $sil(k = 3)$ değerleri incelendiğinde 2 ve 3 küme arasından örnek için $sil(k = 2) = 0{,}3079$ değeri $sil(k = 3) = 0{,}2535$ değerinden daha büyük olduğu için küme sayısının 2 olması daha uygun olabilir.

3.2.2 Calinski ve Harabazs İndeksi

Calinski ve Harabazs (1974), k kümeye sahip bir kümelemenin kalitesini (geçerliliğini) değerlendirebilmek için

$$CH(k) = \frac{BSS(k)/(k-1)}{WSS(k)/(n-k)} \tag{3.8}$$

indeksini önermişlerdir. Burada,

$$WSS(k) = \frac{1}{2}\sum_{l=1}^{k}\sum_{i,j \in C_l} d(i, j) \tag{3.9}$$

$$BSS(k) = \frac{1}{2}\sum_{l=1}^{k}\sum_{\substack{i \in C_L \\ j \notin C_l}} d(i, j) \tag{3.10}$$

olmak üzere, $WSS(k)$ ve $BSS(k)$ sırasıyla, kümeler içi ve kümeler arası kareler toplamlarıdır. $WSS(k)$ ve $BSS(k)$ hesaplanırken karesel uzaklık ölçüleri

kullanılmaktadır. Bu kritere göre, maksimum CH indeks değerine ulaşılan küme sayısı, uygun küme sayısı olarak alınır (Calinski Harabasz 1974).

Çizelge 3.1' de yer alan 10 birim için Karesel Öklid uzaklıkları Çizelge 3.3' de verilmiştir.

Çizelge 3.3 $k = 2$ Kümeye göre karesel öklid uzaklıkları

Birim	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
A1	0	41	174	117	21	62	91	41	61	50
A2	41	0	69	38	20	21	174	74	24	41
A3	174	69	0	107	113	64	221	121	33	122
A4	117	38	107	0	110	27	306	194	74	41
A5	21	20	113	110	0	65	126	38	44	81
A6	62	21	64	27	65	0	157	89	17	10
A7	91	174	221	306	126	157	0	30	106	145
A8	41	74	121	194	38	89	30	0	42	99
A9	61	24	33	74	44	17	106	42	0	41
A10	50	41	122	41	81	10	145	99	41	0

Burada, (A1,A5,A7,A8) bir küme ve (A2,A3,A3,A6,A9,A10) diğer bir küme olarak belirlenmiştir. Çizelge 3.3'den $WSS(k)$ ve $BSS(k)$ değerleri aşağıdaki gibi elde edilebilir.

WSS değerinin elde edilişi;

İlk küme içi uzaklık değerlerinin toplamı, Çizelge (3.3)' den karesel öklid uzaklıkları olmak üzere;

$$WSS(k=2) = \frac{1}{2}[\sum_{i=1,5,7,8}\sum_{j=1,5,7,8} d(i,j) + \sum_{i=2,3,4,6,9,10}\sum_{j=2,3,4,6,9,10} d(i,j)]$$

Buradan ilk küme içi uzaklık değerlerinin toplamı;

$$\sum_{i=1,5,7,8}\sum_{j=1,5,7,8} d(i,j) = (21+91+41)+ (21+126+38)+(91+126+30)+(41+38+30)$$

$$=153+185+247+109=694$$

İkinci küme içi uzaklık değerlerinin toplamı;

$$\sum_{i=2,3,4,6,9,10}\sum_{j=2,3,4,6,9,10} d(i,j)=(69+38+21+24+41)+(69+107+64+33+122)+(38+107+27+74+41)$$

$$+(21+64+27+17+10)+(24+33+74+17+41)+(41+122+41+10+41)$$

=1933+395+287+139+189+255=1458

Buradan $WSS\left(k=2\right)=\frac{1}{2}\left(694+1458\right)=1076$

biçiminde elde edilir.

$BSS(k)=\frac{1}{2}[\sum_{i=1,5,7,8}\sum_{j=2,3,4,6,9,10}d(i,j)]$ olmak üzere;

BSS değeri hesaplanırken bir kümedeki birim ile diğer kümelerde yer alan birimler arasındaki benzerlik ölçüleri dikkate alınır.

$\sum_{i=1,5,7,8}\sum_{j=2,3,4,6,9,10}d(i,j)$=(41+174+117+62+61+50)+(41+20+174+74)+(174+113+221+121)
+(117+110+306+194)+((20+113+110+65+44+81)+(62+65+157+89)
+(174+221+306+157+106+145)+(74+121+194+89+42+89)
+(61+44+106+42)+(50+81+145+99)
=505+309+629+727+433+373+1109+619+253+375=5332

$BSS\left(k=2\right)=\frac{1}{2}\left(5332\right)=2666$

olarak elde edilir. Buradan

$CH\left(k=2\right)=\frac{2666/\left(2-1\right)}{1076/\left(10-2\right)}=19.821$

biçiminde elde edilir.

$k=3$ için indeks değerini hesaplamak için Çizelge 3.3'den (A1,A5,A7,A8) birimleri bir kümeyi, (A2,A4,A6,A10) birimleri ikinci kümeyi, (A3,A9) birimleri üçüncü kümeyi oluştursunlar.

Buradan birinci küme içi uzaklık değerlerinin toplamı Çizelge (3.3)'den;

$\sum_{i=1,5,7,8}\sum_{j=1,5,7,8}d(i,j)=\left(21+91+41\right)+\left(21+126+38\right)+\left(91+126+30\right)+\left(41+38+30\right)$

$=\left(153\right)+\left(185\right)+\left(247\right)+\left(109\right)=694$

İkinci küme içi uzaklık değerlerinin toplamı;

$$\sum_{i=2,4,6,10}\sum_{j=2,4,6,10} d(i,j) = (38+21+41)+(38+27+41)+(21+27+10)+(41+41+10)$$

$$= (100)+(106)+(58)+(92) = 356$$

Üçüncü küme içi uzaklık değerlerinin toplamı;

$$\sum_{i=3,9}\sum_{j=3,9} d(i,j) = (33+33)$$

$$= 66$$

olmak üzere

$$WSS(k=3) = \frac{1}{2}(694+356+66) = \frac{1}{2}(1156) = 558$$

biçiminde elde edilir.

BSS değerinin elde edilişi;

$$\sum_{i=1,5,7,8}\sum_{j=2,4,6,10} d(i,j) + \sum_{i=1,5,7,8}\sum_{j=3,9} d(i,j) + \sum_{i=2,4,6,10}\sum_{j=1,5,7,8} d(i,j) + \sum_{i=2,4,6,10}\sum_{j=3,9} d(i,j) + \sum_{i=3,9}\sum_{j=1,5,7,8} d(i,j) + \sum_{i=3,9}\sum_{j=2,4,6,10} d(i,j)$$

$$= \left[(42+117+62+50)+(174+61)\right]+\left[(41+20+174+74)+(69+24)\right]+$$

$$\left[(174+113+221+121)+(69+107+64+122)\right]+\left[(117+110+306+194)+(107+64)\right]+$$

$$\left[(20+110+65+81)+(113+44)\right]+\left[(62+65+157+89)+(64+17)\right]+$$

$$\left[(174+306+157+145)+(221+106)\right]+\left[(74+194+89+99)+(121+42)\right]+$$

$$\left[(61+44+106+42)+(24+74+17+41)\right]+\left[(50+81+145+99)+(122+41)\right]$$

$$= [270+235]+[309+93]+[629+362]+[727+181]+[276+157]+$$

$$[373+81]+[782+327]+[456+163]+[253+156]+[375+163]$$

$$= 6368$$

olmak üzere

$$BSS(k=3) = \frac{1}{2}(6368) = 3184$$

biçiminde elde edilir. Buradan

$$CH\left(k=3\right)=\frac{3184/\left(3-1\right)}{558/\left(10-3\right)}=19.97$$

olarak elde edilir.

Calinski ve Harabasz indeksi için 2 veya 3 kümeye ayırmak söz konusu olduğunda $CH(k=3)=19{,}97$ değeri $CH(k=2)=19{,}82$ değerinden daha büyük olduğu için 3 kümeye ayırmak daha uygun olacaktır.

3.2.3 Krzanowski ve Lai indeksi

Krzanowski and Lai (1985), küme içi kareler toplamı (WSS) değerinin azalışı tabanında bir indeks önermişlerdir. Öncelikle Krzanowski ve Lai

$$DIFF\left(k\right)=\left(k-1\right)^{2/p}WSS\left(k-1\right)-k^{2/p}WSS\left(k\right) \tag{3.11}$$

istatistiğini tanımlamışlar ve bu istatistiğe bağlı olarak

$$KL\left(k\right)=\left|\frac{DIFF\left(k\right)}{DIFF\left(k+1\right)}\right| \tag{3.12}$$

indeksini tanımlamışlardır. $DIFF\left(k\right)=k^{2/p}WSS\left(k\right)$ (3.13)

alınarak da KL indeksi hesaplanmaktadır. KL indeks değerinin maksimum olması amaçlanmaktadır. k, KL indeksi için uygun küme sayısı olsun. KL indeksinde uygun küme sayısının yani k'nin belirlenmesindeki temel düşünce, $WSS\left(k\right)$ değerinin uygun küme sayısına kadar hızlı bir şekilde azaldığı ve uygun küme sayısından sonra yavaş bir şekilde azaldığıdır. Yani $WSS\left(k\right)$ değerinin en hızlı azalış değerine ulaştığı küme sayısı uygun küme sayısı olarak alınmaktadır.

Çizelge 3.1'deki Hipotetik örnek için, $DIFF\left(k\right)=k^{2/p}WSS\left(k\right)$ alarak KL indeks değerini hesaplansın.

Bölüm 3.2.2' de

$WSS\left(2\right)=WSS\left(k=2\right)=1076$

$WSS\left(3\right)=WSS\left(k=3\right)=558$

olarak bulunmuştur. Eşitlik (3.13)' den

$$DIFF(2) = DIFF(k=2) = 2^{2/3} WSS(2) = \left(2^{2/3}\right)(1076) = 1708.044$$

$$DIFF(3) = DIFF(k=3) = 3^{2/3} WSS(3) = \left(3^{2/3}\right)(558) = 1160.687$$

olmak üzere

$$KL(2) = KL(k=2) = \left|\frac{DIFF(2)}{DIFF(3)}\right| = \left|\frac{1708.044}{1160.687}\right| = 1.471$$

olarak elde edilir.

Buradan

$k=4$ için , Çizelge 3.1'den 1. küme (A4,A6,A10), 2.küme (A7,A8), 3.küme (A1,A2,A5) ve 4. küme (A3,A9) biçiminde alındığında;

WSS değerinin k=4 için elde edilişi;

İlk küme (A4,A6,A10) birimleri için küme içi uzaklık değerlerinin toplamı;

$$\sum_{i=4,6,10} \sum_{j=4,6,10} d(i,j) = (27+41)+ (27+10)+ (41+10)=156$$

İkinci küme(A7,A8) birimleri için küme içi uzaklık değerlerinin toplamı;

$$\sum_{i=7,8} \sum_{j=7,8} d(i,j) = (30+30)=60$$

Üçüncü küme (A1,A2,A5) birimleri için küme içi uzaklık değerlerinin toplamı;

$$\sum_{i=1,2,5} \sum_{j=1,2,5} d(i,j) = (41+21)+(41+20)+(21+20)=164$$

Dördüncü küme (A3,A9) birimleri için küme içi uzaklık değerlerinin toplamı;

$$\sum_{i=3,9} \sum_{j=3,9} d(i,j) = (33+33)= 66$$

Buradan WSS(k=4)= $\frac{1}{2}(156+60+164+66)=223$

$$DIFF(4) = DIFF(k=4) = 4^{2/3} WSS(4) = (4^{2/3})223 = 314.43 \text{ dir.}$$

Böylece,

$$KL(3) = KL(k=4) = \left|\frac{DIFF(3)}{DIFF(4)}\right| = \left|\frac{1160.687}{314.43}\right| = 3.691$$

olarak elde edilir.

$KL(2) = 1{,}471$ ve $KL(3) = 3{,}691$ değerleri incelendiğinde 2 ve 3 küme arasında $KL(3)$ değeri daha büyük olduğundan bu örnek için birimleri 2 kümeye ayırmaktansa 3 kümeye ayırmanın daha uygun olacağı söylenebilir ancak diğer küme sayılarındaki Krzanowski Lai indeks değerlerine bakarak maksimum indeks değeri uygun küme sayısı olarak kabul edilecektir.

3.2.4 Wilk's Lambda istatistiği

Çok Değişkenli Varyans Analizi (MANOVA) iki ve daha fazla gruba ait çok değişkenli verilerin aynı ortalama vektörlü çok değişkenli normal dağılımdan gelip gelmediğinin testinde kullanılmaktadır (Johnson and Wichern 1988). Veriler kümelere ayrıldıktan sonra, elde edilen kümelere MANOVA analizi uygulanabilir ve eğer iyi bir kümeleme yapılmış ise MANOVA analizinde de küme ortalamalarının birbirinden istatistiksel olarak farklı olması beklenir. Oluşturulacak kümeler için hesaplanan Wilks Lambda ölçütünün değerleri küme sayısının belirlenmesinde kullanılabilir (Tatlıdil 1996). Wilks değeri 0 ile 1 arasında değerler alabilen bir istatistiktir. Küme sayısı arttıkça Wilks istatistiğinin değeri azalmakta ve sıfıra yaklaşmaktadır. Wilks Lambda Ölçütünün değerinin tam olarak 0 değerine ulaşmamakla birlikte sıfıra en çok yaklaştığı durumdaki grup sayısı uygun küme sayısı olarak belirlenebilir. Uygulama çalışmasında Wilks lambda ölçüt değerinin 0,001 değerinin altında düştüğü ilk küme sayısı küme sayısı olarak belirlenmiştir.

MANOVA uygulanabilmesi için; verilerin çok değişkenli normal dağılımdan seçilmesi, gruplardaki kovaryansların homojen olması gibi varsayımlara ihtiyaç vardır (Krzanowski 1993).

Varyansların homojenliği testi Box'ın M testi ile yapılabilir. Varyansların homojenliği gruplardaki birimlerin alındıkları yığınların varyans-kovaryans matrislerinin istatistiksel olarak aynı olması anlamına gelmektedir. Varyansların homojenliği için hipotezler Σ_i $(i = 1, \ldots, k)$ $i.$ grubtaki birimlerin seçildikleri yığının varyans-kovaryans matrisi olmak üzere aşağıdaki şekilde oluşturulmaktadır.

$H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$

$H_1 : \Sigma_i \neq \Sigma_j \quad ; \; i, j = 1, \ldots, k \; , \; i \neq j$

Bu hipotezin test edilmesinde kullanılan Box'ın M istatistiği ise

$$M = \sum_{i=1}^{k} (n_i - 1) \ln |S| - \sum_{i=1}^{k} (n_i - 1) \ln |S_i| \qquad (3.14)$$

şeklinde hesaplanmaktadır (Hawkins 1982). Burada n_i, $i.$ gruptaki birim sayısı, S_i, $i.$ gruptaki birimlerden hesaplanan örneklem varyans-kovaryans matrisi ve

$$S = \frac{\sum_{i=1}^{k} (n_i - 1) S_i}{\sum_{i=1}^{k} (n_i - 1)} \qquad (3.15)$$

ise tüm gruplar için örneklem ortak varyans-kovaryans matrisini göstermektedir. Box'ın M istatistiğinin değerlendirilebilmesi için ;

$$C = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \left(\sum_{i=1}^{k} \frac{1}{(n_i - 1)} - \frac{1}{\sum_{i=1}^{k} (n_i - 1)} \right) \qquad (3.16)$$

gibi bir C çarpanı hesaplanır. Buradan MC istatistiğinin $(p)(p+1)(k-1)/2$ serbestlik dereceli χ^2 dağılımına sahip olduğu gösterilmiştir (Harris 1975). Hesaplanan MC istatistiği χ^2 tablo değerinden küçük ise $H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$ şeklindeki hipotez red edilemez. Gruplardaki varyans-kovaryans matrislerinin homojen olması MANOVA yönteminin uygulanabilmesi anlamına gelmektedir.

MANOVA iki ve daha fazla gruba ait çok değişkenli verilerin aynı ortalama vektörlü çok değişkenli normal dağılımdan gelip gelmediğinin testinde kullanılmaktadır. Varyans-kovaryans matrisleri homojen olan k tane grup için $\underline{\mu}_i$ $(i = 1, \ldots, k)$ $i.$ gruptaki birimlerin seçildikleri yığının ortalama vektörü olmak üzere, hipotezler aşağıdaki şekilde oluşturulmaktadır.

$H_0 : \underline{\mu}_1 = \underline{\mu}_2 = \ldots = \underline{\mu}_k$

$H_1 : \underline{\mu}_i \neq \underline{\mu}_j \quad ; i, j = 1, \ldots, k \, , \; i \neq j$

Bu hipotezin test edilmesinde kullanılan başlıca yöntemler, Roy'un en büyük karakteristik kök yöntemi, Hotelling-Lawley iz yöntemi ve Wilks'in olabilirlik oran yöntemidir. Burada Wilks'in olabilirlik oran yöntemi kullanılacaktır.

Wilks'in olabilirlik oran istatistiği, T ; genel çarpımlar ve kareler toplamı matrisi, B ; gruplar arası çarpımlar ve kareler toplamı matrisi ve W grup içi çarpımlar ve kareler toplamı matrisi olmak üzere,

$$\Delta = \frac{|W|}{|W+B|} = \frac{|W|}{|T|} \tag{3.17}$$

şeklinde hesaplanmaktadır (Tatlıdil 1996). Burada

Daha önce (3.3) ve (3.5) eşitliklerinde verildiği gibi $B = \sum_{j=1}^{k} n_j (\overline{X}_j - \overline{X})(\overline{X}_j - \overline{X})'$ ve

$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_j)(X_{ij} - \overline{X}_j)'$ biçimindedir.

Burada X_{ij}; j. kümedeki i. birim değerleri

$\overline{X_j}$; j. kümenin örneklem ortalama vektörü

$\overline{X}$; Tüm birimlerin örneklem ortalama vektörü

n_j; j. kümedeki birim sayısı

0 ile 1 arasında değerler alan bu test istatistiğinde, Δ'nın 0'a yakın olması $H_0 : \underline{\mu}_1 = \underline{\mu}_2 = \ldots = \underline{\mu}_k$ hipotezinin reddedileceğini, 1'e yakın olması ise kabul edileceğini gösteren bir işarettir. Buradan n'nin yeterince büyük olması durumunda,

$$L = -\left(n - 1 - \left(p + k\right)/2\right) \ln \Delta \tag{3.18}$$

istatistiğinin $(p)(k-1)$ serbestlik dereceli yaklaşık χ^2 dağılımına sahip olduğu gösterilmiştir. Hesaplanan L istatistiği χ^2 tablo değerinden büyük ise birimlerin aynı ortalama vektörlü yığınlardan geldiğini iddia eden H_0 hipotezi reddedilir (Johnson and Wichern 1988).

H_0 hipotezi reddedilirse birimlerin ortalama vektörlerinin en az ikisinin birbirinden farklı olduğu sonucuna varılır. H_0 hipotezi reddedilirse birbirinden farklı olan grupların incelemesi Hotelling T^2 yöntemi ile yapılabilmektedir. Hotelling T^2 istatistiği çok değişkenli iki kitle ortalama vektörünün eşitliğinin test edilmesinde kullanılan bir istatistiktir (Hawkins 1982). n_1 birinci gruptaki birim sayısı, n_2 ikinci gruptaki birim sayısı, S_1 birinci gruptaki birimlerden hesaplanan örneklem varyans-kovaryans matrisi, S_2 ikinci gruptaki birimlerden hesaplanan örneklem varyans-kovaryans matrisi ve $S = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n-2}$ ise ortak örneklem varyans-kovaryans matrisi olmak üzere, T^2 istatistiği

$$T^2 = \frac{n_1 n_2}{n}(\overline{X}_1 - \overline{X}_2)' S^{-1} (\overline{X}_1 - \overline{X}_2) \qquad (3.19)$$

olarak hesaplanmaktadır (Krzanowski 1993).

Burada $\overline{X}_1$; 1. grupdaki örneklem ortalama vektörü

$\overline{X}_2$; 2. gruptaki örneklem ortalama vektörü

Varyans-kovaryans matrisleri bilinmiyor fakat ortak 2 tane grup için $\underline{\mu}_i$ $(i = 1,2)$ $i.$ gruptaki birimlerin seçildikleri yığının ortalama vektörü olmak üzere, hipotezler aşağıdaki şekilde oluşturulmaktadır.

$H_0 : \underline{\mu}_1 = \underline{\mu}_2$

$H_1 : \underline{\mu}_1 \neq \underline{\mu}_2$

Bu hipotezin test edilmesinde T^2 istatistiğine dayalı

$$F_{p,n-p-1} = \frac{n-p-1}{p}\frac{T^2}{n-2} \tag{3.20}$$

istatistiği kullanılır ve eğer $F_{p,n-p-1}$ hesaplanan değeri $F_{p,n-p-1;\alpha}$ tablo değerinden büyük ise $H_0 : \underline{\mu}_1 = \underline{\mu}_2$ hipotezi reddedilir ve iki grubun ortalama vektörlerinin birbirinden farklı olduğu sonucuna varılır (Seber 1984).

$k = 2$ durumunda, Wilks lambda ölçüt değerini hesaplamak için; A1, A5, A7 ve A8 bir kümeye, A2, A3, A4, A6, A9 ve A10'da diğer kümenin elemanları olmak üzere; Çizelge 3.3'de yer alan karesel öklid uzaklıklarından hesaplanan T, B ve W matrisleri sırasıyla ;

$$T = \begin{bmatrix} 172.5 & 0 & -73.5 \\ 0 & 91.6 & 14.4 \\ -73.5 & 14.4 & 110.1 \end{bmatrix}$$

$$B = \begin{bmatrix} 70.4167 & -17.3333 & -80.1667 \\ -17.3333 & 4.2667 & 19.7333 \\ -80.1667 & 19.7333 & 91.2667 \end{bmatrix}$$

$$W = \begin{bmatrix} 102.0833 & 17.3333 & 6.6667 \\ 17.3333 & 87.3333 & -5.3333 \\ 6.6667 & -5.3333 & 18.8333 \end{bmatrix}$$

biçiminde elde edilir. İki küme için Box homojenlik testi yapılmış ve oluşturulan kümelerin ortak varyans-kovaryans matrisli kitlelerden geldikleri anlaşılmıştır ($BoxM = 19.863$, $\chi^2 = 1.771$, p değeri $= 0.105$). Buradan Wilks Lambda ölçütünün değeri

$$\Delta = \frac{|W|}{|T|} = \frac{154230}{1209100} = 0.1276$$

olarak hesaplanır. Wilks Lambda istatistiği'nin anlamlılığı değerlendirildiğinde oluşturulan iki kümenin ortalama vektörlerinin birbirinden önemli derecede farklı olduğu anlaşılmaktadır (L test istatistiği $= 19.6381$, $\chi^2_{3;0.05} = 7.8147$).

$k = 3$ durumunda, Wilk's Lambda istatistiği'nin değerini hesaplamak için A1, A5, A7 ve A8 birinci kümenin, A2, A4, A6 ve A10 ikinci kümenin ve A3 ile A9'da üçüncü kümenin elemanları olmak üzere T,B ve W matrisleri

$$T = \begin{bmatrix} 172.5 & 0 & -73.5 \\ 0 & 91.6 & 14.4 \\ -73.5 & 14.4 & 110.1 \end{bmatrix}$$

$$B = \begin{bmatrix} 78.75 & 3.5 & -81 \\ 3.5 & 56.35 & 17.65 \\ -81 & 17.65 & 91.35 \end{bmatrix}$$

$$W = \begin{bmatrix} 93.75 & -3.5 & 7.5 \\ -3.5 & 35.25 & -3.25 \\ 7.5 & -3.25 & 18.75 \end{bmatrix}$$

biçiminde elde edilir. Üç küme için Box homojenlik testi yapılmış ve oluşturulan kümelerin ortak varyans-kovaryans matrisli yığınlardan gelmedikleri anlaşılmıştır. ($BoxM = 19.846$, $\chi^2 = 0.051$, p değeri $= 0.000$)

Wilks Lambda istatistiği ile amaçlanan varsayımların sağlandığı Wilks değerinde 0,001 değerinin altına düşen ilk küme sayısı uygun küme sayısı olarak belirlenmiştir (Dinçer ve Özdamar 1992).

4. İLLERİN SUÇ İSTATİSTİKLERİ BAKIMINDAN KÜMELENDİRİLMESİ

Çeşitli nedenlerle meydana gelen ve sosyoloji, psikoloji, kriminoloji, hukuk gibi farklı bilim dallarınca değişik şekillerde yorumlanan suç, topluma zarar verdiği veya tehlikeli olduğu kanunlarla kabul edilen eylemler (Dönmezer 1994) ya da genel tanımıyla toplum menfaatlerini ihlal eden fiiller olarak tanımlanmaktadır (Seyhan 2002). Ortaya çıkış sebebi, gelişmesi, niteliği gibi özelliklerinden dolayı birçok bilim dalına konu olan suç olayının ana kaynağı insandır. İnsanlar belirli bir mekânda yaşadıklarından ve suçlarında bir mekânda işlenmesinden, suç olayları coğrafyanın bir inceleme konusu olmuştur. Suç araştırmaları, suç ile suçun işlendiği yer arasında anlamlı ilişkiler bulunduğunu ortaya koymuştur (Appiahene-Gyamfi 2002, Karakaş 2004). Suçların farklı dağılış özellikleri göstermesi, bazı yerlerde artarken bazı yerlerde azalması, olayın mekânsal açıdan ele alınmasını gerektirmektedir. Suçların mekânsal dağılımı incelendiğinde gelişigüzel olmadığı açıkca görülebilir. Suçların nerede, niçin ve ne zaman meydana geldiğini anlamak, hangi suç tipleri hangi bölgelerde meydana gelmiş, demoğrafik değerler ile suçların nedenleri arasındaki ilişki, mekânlar ile suç tipleri arasındaki ilişki nedir gibi soruların yanıtlarını aramak suçla mücadele yöntemlerini geliştirebilir.

Bu bölümde Türkiye'deki 81 il, 11 farklı suç türüne göre kümeleme analizi ile incelenecektir. Analizde kullanılan değişkenler; nüfusa oranlı şekilde x_1: öldürme, x_2: cinsel suçlar, x_3; kişiyi hürriyetinden yoksun kılma, x_4; hırsızlık, x_5; gasp, x_6; dolandırıcılık, x_7; uyuşturucu ve uyarıcı madde imal ve ticareti, x_8; sahtecilik, x_9; zimmet, x_{10}; kaçakçılık, x_{11}: orman suçları olarak belirlenmiştir (Anonim 2006).

Kümeleme Analizi tekniği olarak Bölüm 2'de tanıtılan hiyerarşik kümeleme yöntemlerinden tek bağlantı tekniği, tam bağlantı tekniği, Ward yöntemi, hiyerarşik olmayan kümeleme yöntemlerinden ise $k-$ortalama tekniği kullanılacaktır. Uzaklık ölçüsü olarak Öklid uzaklığı kullanılacaktır. Analizlerde MATLAB 7.0 programı kullanılmıştır.

Ayrıca tezin konusu olmamakla beraber çok değişkenli istatistiksel analiz yöntemlerinden temel bileşenler analiz yöntemi ile de illerin durumu incelenecektir. İllerin farklı suç türlerine göre kümeleme analizi ile değerlendirilmesinden sonra illerin temel bileşen skorlarının grafiksel gösterimine de yer verilecektir.

Temel Bileşenler Analizi orijinal p değişkenin varyans-kovaryans yapısını daha az sayıda ve bu değişkenlerin doğrusal bileşenleri olan yeni değişkenlerle ifade eden bir yöntemidir. Başka bir deyişle, aralarında korelasyon bulunan p sayıda değişkenin açıkladığı yapıyı, aralarında korelasyon bulunmayan ve sayıca orijinal değişken sayısından daha az sayıda (l<p) orijinal değişkenlerin doğrusal bileşenleri olan değişkenlerle ifade etme yöntemidir. Temel Bileşenler Analizinin temel amacı boyut indirgemek olmakla beraber, ilişkili değişken setlerinden birimlerin temel bileşen değerlerini hesaplamak ve birimleri bu değerlere göre sıralamak için de kullanılmaktadır.

Çizelge 4.1'de Tek Bağlantı tekniği, Tam Bağlantı tekniği, Ward yöntemi ve k – ortalama tekniği ile Silhouette indeks değerleri yer almaktadır.

Çizelge 4.1 Silhouette indeks değerleri

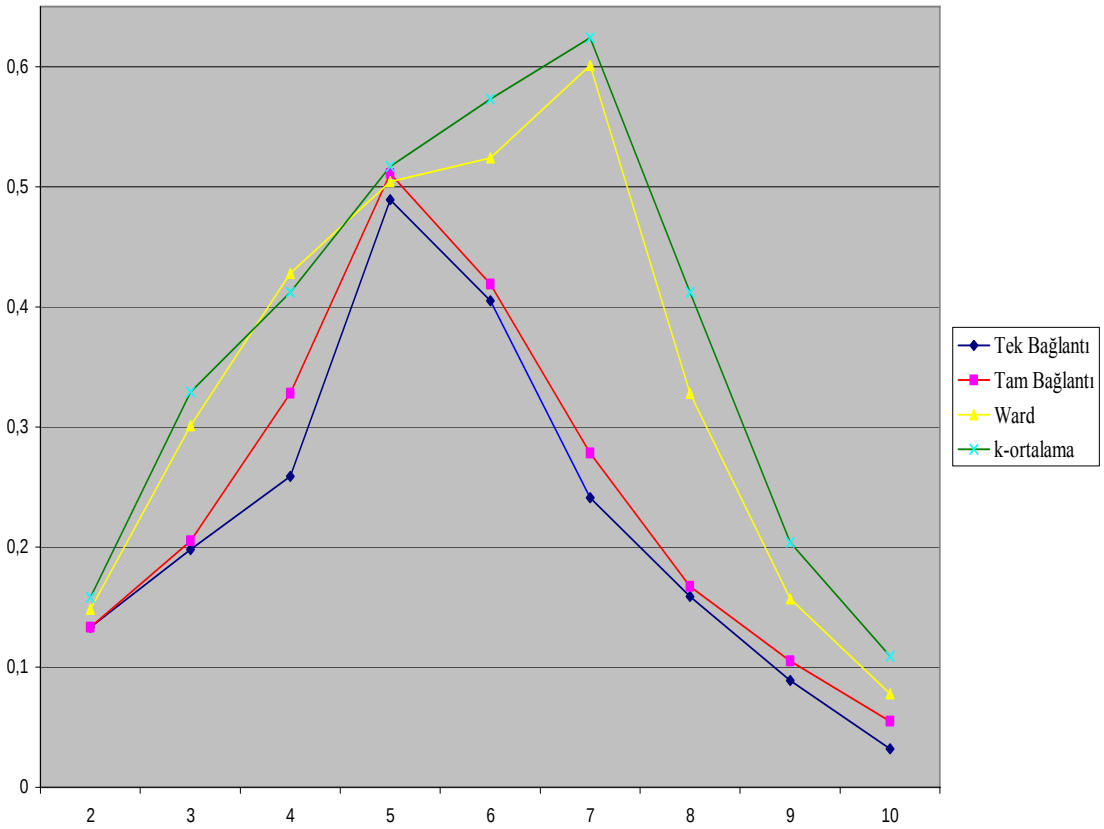
k	**Tek Bağlantı**	**Tam Bağlantı**	**Ward**	k – **ortalama**
2	0.133	0.133	0.148	0.158
3	0.198	0.205	0.301	0.329
4	0.259	0.328	0.428	0.412
5	**0.489**	**0.511**	0.504	0.517
6	0.405	0.419	0.524	0.573
7	0.241	0.278	**0.601**	**0.624**
8	0.159	0.167	0.328	0.412
9	0.089	0.105	0.157	0.204
10	0.032	0.055	0.078	0.109

Çizelge 4.1 incelendiğinde, Tek Bağlantı Yönteminde Silhouette indeksi en büyük değeri , k=5 küme sayısında 0,489 olarak almıştır ve dikkat edildiğinde küme sayısı , k=6 olduğunda Silhouette değeri azalmış ve küme sayısı arttıkça , Silhouette değeri ciddi azalış göstererek 0,489 değerinden 0,032 değerine kadar düşüş göstermiştir.

Tam Bağlantı Yönteminde de küme sayısı k=5 de Silhouette değeri 0,511 le en büyük değerini almış ve tek bağlantı yöntemine benzer şekilde küme sayısı arttıkça, Silhouette değeri azalış göstermiştir. Ward yönteminde küme sayısı k=7 olduğunda Silhouette indeksi 1 'e en yakın değer olan 0,601 değerini almış ve küme sayısı k=8 olduğunda 0,328 e düşmüş ve küme sayısının artmasıyla indeks değerleri paralel olarak azalmıştır. *k*-Ortalama yönteminde Silhouette indeks değeri 0,624 maksimum değerini Ward yönteminde olduğu gibi k=7 küme sayısında almıştır.

Şekil 4.1'de Silhouette indeks değerlerinin grafiği yer almaktadır. Bu grafiğe göre Tek Bağlantı ve Tam Bağlantı yöntemlerine göre uygun küme sayısı 5 iken, Ward ve k – ortalama yöntemlerine göre uygun küme sayısı 7 olarak belirlenmektedir.

Çizelge 4.2'de tek bağlantı yöntemine göre Silhouette indeksine göre belirlenen $k = 5$ küme için illerin kümelere göre dağılımı yer almaktadır.



Şekil 4.1 Kümeleme Yöntemlerinin Silhouette İndeksine göre gösterimi

Çizelge 4.2 Tek Bağlantı Yöntemine Göre Uygun Küme Sayısı $k = 5$ için İllerin kümelere göre dağılımı

Küme 1	Ardahan, Osmaniye
Küme 2	Adana, Adıyaman, Afyon, Ağrı, Aksaray, Amasya, Antalya, Ankara, Artvin, Aydın, Balıkesir, Bartın, Batman, Bayburt, Bilecik, Bingöl, Bitlis, Bolu, Burdur, Bursa, Çanakkale, Çankırı, Çorum, Denizli, Diyarbakır, Düzce, Edirne, Elazığ, Erzincan, Erzurum, Eskişehir, Gaziantep, Giresun, Gümüşhane, Hatay, Iğdır, Isparta, İzmir, Kahramanmaraş, Karabük, Karaman, Kars, Kastamonu, Kayseri, Kırıkkale, Kırklareli, Kırşehir, Kilis, Kocaeli, Konya, Kütahya, Malatya, Manisa, Mardin, Mersin, Muğla, Muş, Nevşehir, Niğde, Ordu, Rize, Sakarya, Samsun, Siirt, Sinop, Sivas, Şanlıurfa, Şırnak, Tekirdağ, Tokat, Trabzon, Tunceli, Uşak, Yalova, Yozgat, Zonguldak
Küme 3	İstanbul
Küme 4	Hakkari
Küme 5	Van

Çizelge 4.2 incelendiğinde, tek bağlantı yöntemi ile İstanbul, Hakkari ve Van'ın tek başına birer küme oluşturdukları, Ardahan ve Osmaniye'nin başka bir küme ve geri kalan diğer tüm illerin bir küme oluşturdukları gözlenmektedir. Van ilinde kaçakçılık suçundan hüküm giyen, Hakkâri ilinde ise uyuşturucu ve uyarıcı madde imal ve ticareti suçundan hüküm giyenlerin nüfuslarına oranla diğer illerden fazla olduğundan, bu iller tek başına birer küme oluşturmuştur. Bu iki ilin sınır ili olması ve işlenen suçların uyuşturucu ve kaçakçılık olması dikkat çekicidir.

Çizelge 4.3'de tam bağlantı yöntemine göre Silhouette indeksine göre belirlenen $k = 5$ küme için illerin kümelere göre dağılımı yer almaktadır.

Çizelge 4.3 Tam Bağlantı Yöntemine Göre Uygun Küme Sayısı $k = 5$ için İllerin kümelere göre dağılımı

Küme 1	Adıyaman, Ağrı, Aksaray, Bartın, Batman, Bayburt, Bilecik, Bingöl, Bitlis, Çankırı, Erzincan, Erzurum, Giresun, Gümüşhane, Hakkâri, Isparta, Kırıkkale, Kırşehir, Kilis, Kütahya, Mardin, Muş, Nevşehir, Niğde, Ordu, Rize, Siirt, Sinop, Sivas, Şanlıurfa, Şırnak, Trabzon, Yozgat
Küme 2	Adana, Afyon, Amasya, Artvin, Aydın, Balıkesir, Bolu, Burdur, Çanakkale, Çorum, Diyarbakır, Düzce, Edirne, Elazığ, Eskişehir, Hatay, Iğdır, İzmir, Kahramanmaraş, Karabük, Karaman, Kars, Kastamonu, Kırklareli, Kocaeli, Konya, Malatya, Manisa, Mersin, Muğla, Sakarya, Samsun, Tekirdağ, Tokat, Tunceli, Uşak, Yalova, Zonguldak
Küme 3	Ardahan, Osmaniye
Küme 4	Ankara, Antalya, Bursa, Denizli, Gaziantep, İstanbul, Kayseri
Küme 5	Van

Çizelge 4.3 incelendiğinde tam bağlantı yöntemi ile Van'ın tek başına bir küme oluşturduğu, Ardahan ve Osmaniye'nin başka bir küme ve geri kalan illerin de kendi aralarında üç küme oluşturdukları gözlenmektedir. Dördüncü kümede büyük illerimizden olan Ankara, Antalya, Bursa, Denizli, Gaziantep, İstanbul ve Kayseri'nin kendi arasında bir küme oluşturduğu gözlenmektedir.

Çizelge 4.4'de Ward yöntemine göre Silhouette indeksine göre belirlenen $k = 7$ küme sayısı için kümeleme sonuçları yer almaktadır.

Çizelge 4.4 Ward Yöntemine Göre Uygun Küme Sayısı $k = 7$ için İllerin kümelere göre dağılımı

Küme 1	Adıyaman, Aksaray, Bartın, Batman, Bayburt, Bilecik, Bingöl, Bitlis, Çankırı, Erzincan, Erzurum, Giresun, Gümüşhane, Isparta, Kırıkkale, Kırşehir, Kütahya, Mardin, Muş, Nevşehir, Niğde, Ordu, Rize, Sinop, Sivas, Şanlıurfa, Trabzon, Yozgat
Küme 2	Afyon, Amasya, Artvin, Burdur, Çorum, Diyarbakır, Düzce, Elazığ, Iğdır, Karabük, Karaman, Kars, Kastamonu, Kırklareli, Malatya, Tokat, Tunceli
Küme 3	Ağrı, Hakkâri, Kilis, Siirt, Şırnak
Küme 4	Ardahan, Osmaniye
Küme 5	Adana, Aydın, Balıkesir, Bolu, Çanakkale, Edirne, Eskişehir, Hatay, Kahramanmaraş, Kocaeli, Manisa, Mersin, Muğla, Sakarya, Samsun, Tekirdağ, Yalova, Zonguldak
Küme 6	Van
Küme 7	Ankara, Antalya, Bursa, Denizli, Gaziantep, İstanbul, İzmir, Kayseri, Konya, Uşak

Çizelge 4.4 incelendiğinde, Ward yöntemi ile Van'ın tek başına bir küme oluşturduğu, Ardahan ve Osmaniye'nin başka bir küme ve geri kalan illerin de kendi aralarında beş küme oluşturdukları gözlenmektedir. Son kümede büyük illerimizden olan Ankara, Antalya, Bursa, Denizli, Gaziantep, İstanbul, İzmir, Kayseri ve Konya'nın kendi arasında bir küme oluşturduğu gözlenmektedir.

Çizelge 4.5'de $k-$ ortalama yöntemine göre Silhouette indeksine göre belirlenen $k = 7$ küme için illerin kümelere göre dağılımı yer almaktadır.

Çizelge 4.5 k – ortalama Yöntemine Göre Uygun Küme Sayısı $k = 7$ için İllerin kümelere göre dağılımı yer almaktadır.

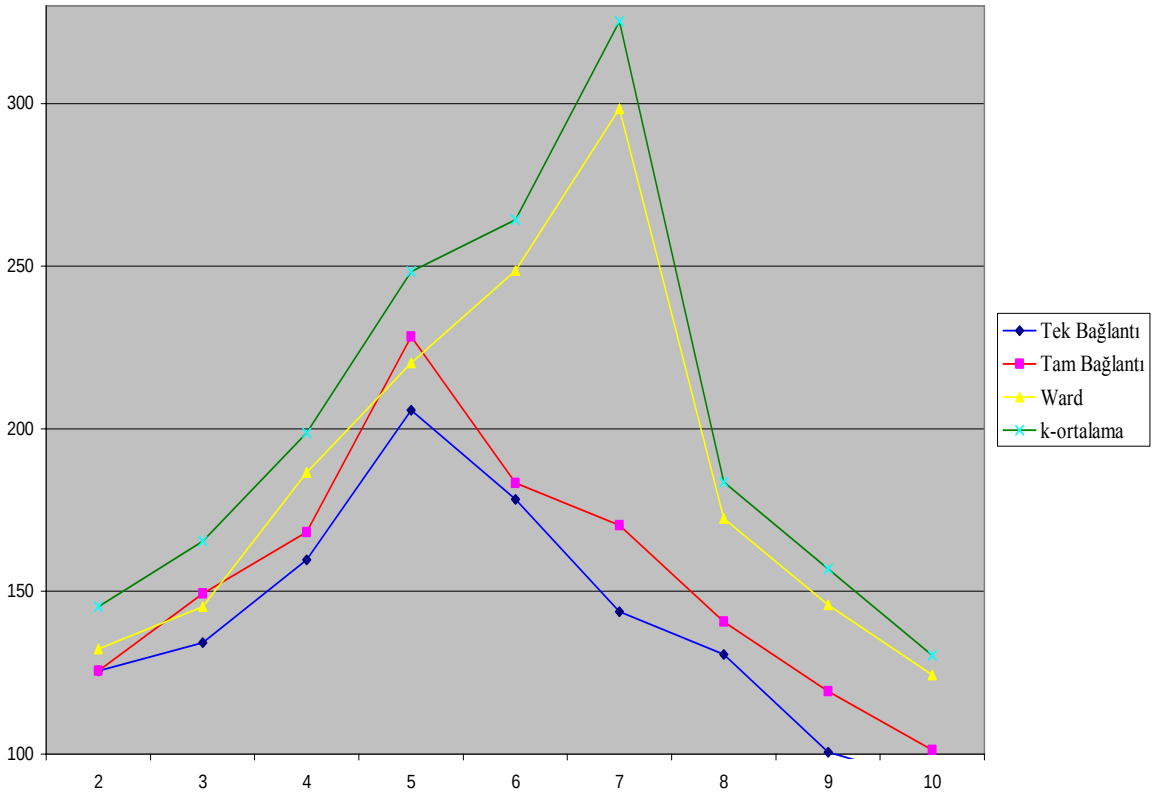
Küme 1	Ardahan, Osmaniye
Küme 2	Ankara, Antalya, Bursa, Denizli, Gaziantep, İstanbul, Kayseri, Konya
Küme 3	Hakkâri, Kilis
Küme 4	Adana, Aydın, Balıkesir, Bolu, Çanakkale, Edirne, Eskişehir, İzmir, Kocaeli, Manisa, Mersin, Muğla, Samsun, Tekirdağ, Uşak, Yalova, Zonguldak
Küme 5	Van
Küme 6	Afyon, Amasya, Artvin, Burdur, Çorum, Diyarbakır, Düzce, Elazığ, Hatay, Iğdır, Kahramanmaraş, Karabük, Karaman, Kars, Kastamonu, Kırklareli, Malatya, Sakarya, Tokat, Tunceli
Küme 7	Adıyaman, Ağrı, Aksaray, Bartın, Batman, Bayburt, Bilecik, Bingöl, Bitlis, Çankırı, Erzincan, Erzurum, Giresun, Gümüşhane, Isparta, Kırıkkale, Kırşehir, Kütahya, Mardin, Muş, Nevşehir, Niğde, Ordu, Rize, Siirt, Sinop, Sivas, Şanlıurfa, Şırnak, Trabzon, Yozgat

Çizelge 4.6'da tek bağlantı tekniği, tam bağlantı tekniği, Ward yöntemi ve k – ortalama tekniği ile Calinski ve Harabazs indeks değerleri, Şekil 4.2'de ise indeks değerlerinin grafiği yer almaktadır. Bu grafiğe göre tek bağlantı ve tam bağlantı yöntemlerine göre uygun küme sayısı 5 iken, Ward ve k – ortalama yöntemlerine göre uygun küme sayısı 7 olarak belirlenmektedir.

Çizelge 4.6 Calinski ve Harabazs indeks değerleri

k	**Tek Bağlantı**	**Tam Bağlantı**	**Ward**	k – **ortalama**
2	125.54	125.54	132.29	145.26
3	134.28	149.34	145.37	165.39
4	159.64	168.26	186.61	198.67
5	**205.64**	**228.34**	220.28	248.37
6	178.27	183.27	248.64	264.31
7	143.83	170.34	**298.35**	**325.19**
8	130.57	140.64	172.38	183.62
9	100.58	119.35	145.82	156.84
10	91.38	101.25	124.36	130.31

Çizelge 4.6 incelendiğinde, Tek Bağlantı Yönteminde , Calinski ve Harabasz indeks değerlerinin maksimumunu k=5 küme sayısında 205,64 değeriyle almıştır, indeks değerleri küme sayısı arttıkça artış ve 205,64 değeriyle maksimuma ulaşmış ve sonra tekrar azalış göstermiştir . Tam Bağlantı yönteminde k=5 küme sayısında Calinski ve Harabasz indeksi 228,34 değerini almıştır. Ward yönteminde ve *k*-Ortalama yönteminde k=7 küme olmuştur çünkü Ward yönteminde Calinski ve Harabasz indeksi 298,35 olarak diğer indeks değerlerinin en büyüğüdür. *k*-Ortalama yönteminde Calinski ve Harabasz indeksi maksimum değerini k=7 kümede 325,19 olarak almıştır.



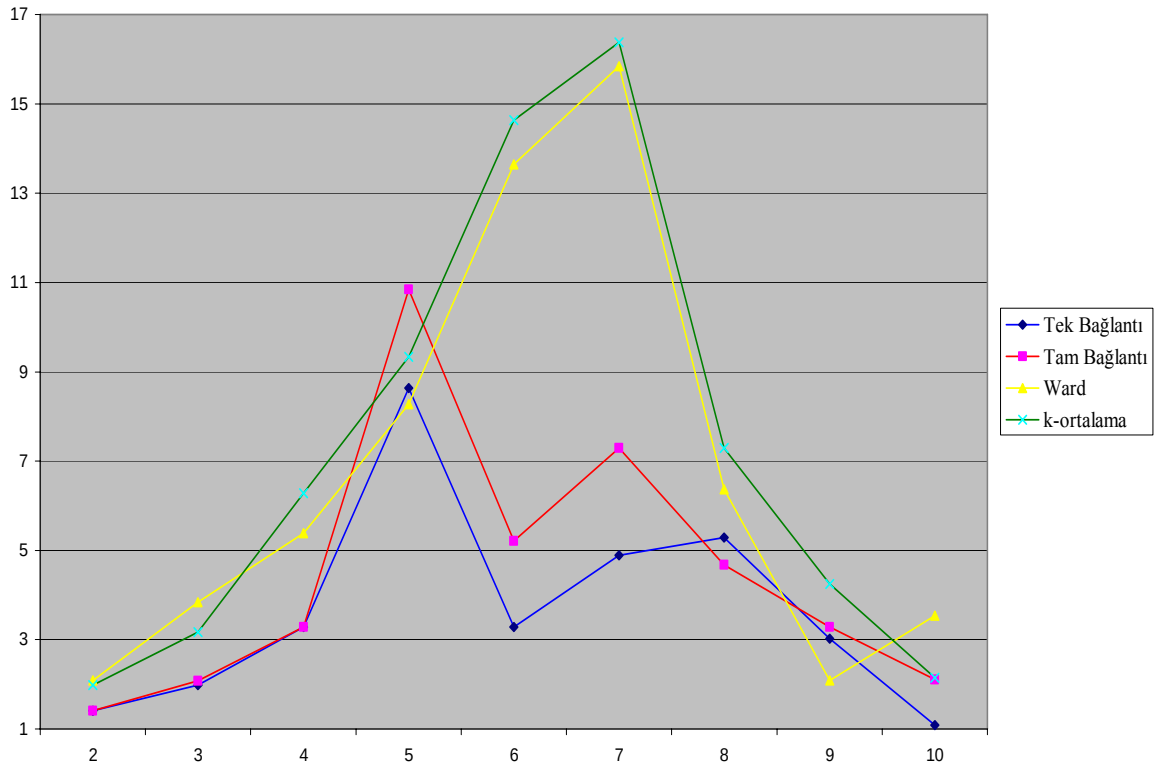
Şekil 4.2 Kümeleme Yöntemlerinin Calinski ve Harabazs İndeksine Göre Gösterimi

Çizelge 4.7'de tek bağlantı tekniği, tam bağlantı tekniği, Ward yöntemi ve k – ortalama tekniği ile Krzanowski ve Lai indeks değerleri, Şekil 4.3'de ise indeks değerlerinin grafiği yer almaktadır. Bu grafiğe göre tek bağlantı ve tam bağlantı yöntemlerine göre uygun küme sayısı 5 iken, Ward ve k – ortalama yöntemlerine göre uygun küme sayısı 7 olarak belirlenmektedir.

Çizelge 4.7 Krzanowski ve Lai indeks değerleri

k	**Tek Bağlantı**	**Tam Bağlantı**	**Ward**	***k* – ortalama**
2	1.41	1.41	2.09	1.98
3	1.98	2.08	3.84	3.17
4	3.29	3.29	5.39	6.28
5	**8.64**	**10.84**	8.28	9.34
6	3.29	5.21	13.64	14.64
7	4.89	7.29	**15.84**	**16.38**
8	5.29	4.68	6.37	7.29
9	3.02	3.29	2.09	4.25
10	1.09	2.11	3.54	2.15

Krzanowski ve Lai indeks değerleri Çizelge 4.7'de incelendiğinde, Tek Bağlantı yönteminde küme sayısı 5 olduğunda 8,64 olmuştur , küme sayısı arttıkça indeks değeri azalmaktadır. Tam Bağlantı Yönteminde indeks değeri k=5 de 10,84 değerini almıştır . Ward yöntemi ve *k*-Ortalama yönteminde indeks değerleri incelendiğinde 7 kümede maksimum değeri almıştır .



Şekil 4.3 Kümeleme Yöntemlerinin Krzanowski ve Lai İndeksine Göre Gösterimi

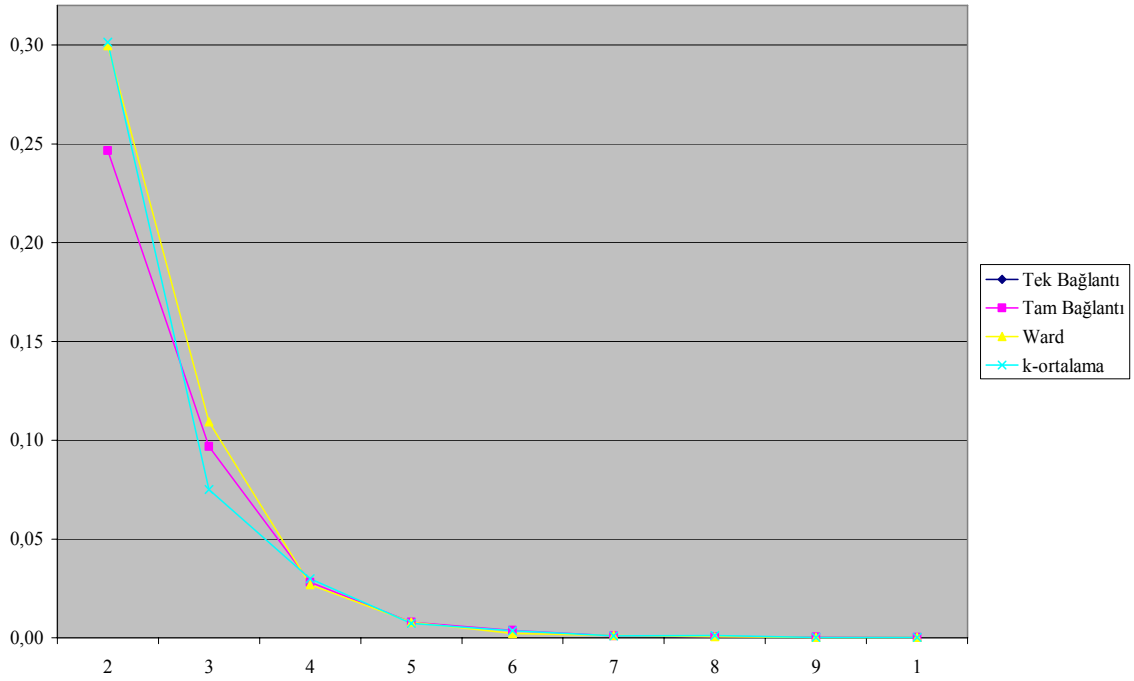
Çizelge 4.8'de tek bağlantı tekniği, tam bağlantı tekniği, Ward yöntemi ve k – ortalama tekniği ile Wilk's Lambda değerleri, Şekil 4.4'de ise indeks değerlerinin grafiği yer almaktadır. Bu grafiğe göre tek bağlantı yöntemine göre uygun küme sayısı 10, tam bağlantı yöntemine göre 7, Ward ve k – ortalama yöntemlerine göre uygun küme sayısı 8 olarak belirlenmektedir.

Çizelge 4.8 Wilk's Lambda değerleri

k	**Tek Bağlantı**	**Tam Bağlantı**	**Ward**	k – **ortalama**
2	0.24646000	0.24646000	0.29957000	0.30132000
3	0.09743300	0.09691500	0.10937000	0.07499800
4	0.07009800	0.02815300	0.02700000	0.02991600
5	0.02100200	0.00797130	0.00801090	0.00725900
6	0.00921150	0.00369700	0.00210920	0.00355420
7	0.00654660	**0.00094475**	0.00109200	0.00925520
8	0.00438190	0.00059350	**0.00054740**	**0.00012597**
9	0.00327030	0.00032765	0.00016923	0.00011360
10	**0.00029082**	0.00011253	0.00010647	0.00010374

Çizelge 4.8 incelendiğinde, Wilk's Lambda değerlerini, durdurma değeri olan 0,001 değerinin altına düşen değerine karşılık gelen küme sayısı uygun küme sayısı olarak belirleneceğinden Tek Bağlantı Yöntemine göre 0,001 değerinin altına düşen ilk değer k=10 küme sayısı olan 0,000029082 değeridir. Tam Bağlantı Yönteminde 0,00094475 değerini k=7 de almıştır.

Ward Yöntemi ve k-ortalama Yönteminde k=8 de 0,001 değerinin altına düşmüştür.



Şekil 4.4 Kümeleme Yöntemlerinin Wilk's Lambda İstatistiğine Göre Gösterimi

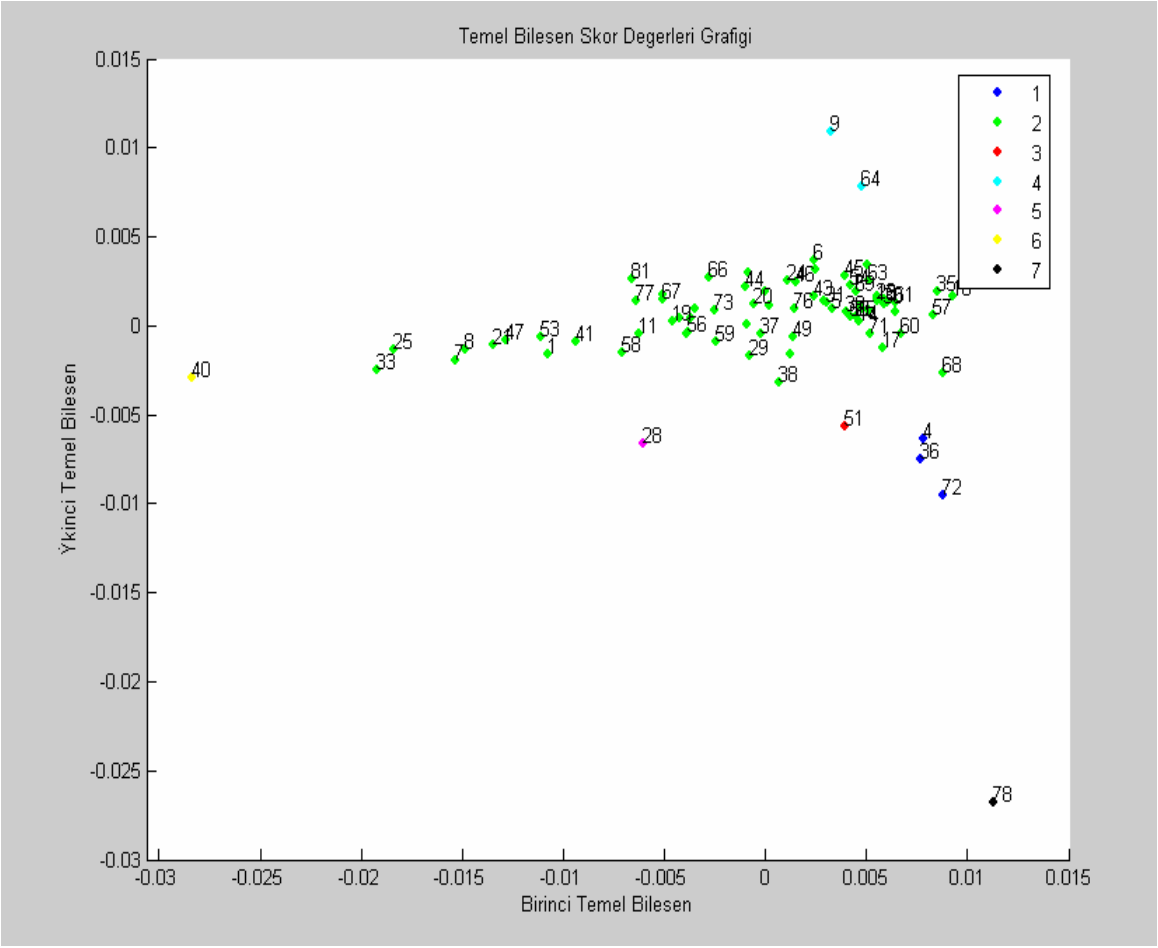
Çizelge 4.9'da $k-$ortalama yöntemine göre Wilk's Lambda istatistiği'ne göre belirlenen küme sayısı $k=8$ için kümeleme sonuçları yer almaktadır.

Çizelge 4.9 $k-$ortalama Yöntemine Göre Uygun Küme Sayısı $k=8$ için İllerin kümelere göre dağılımı

Küme 1	Afyon, Amasya, Artvin, Burdur, Çorum, Düzce, Elazığ, Hatay, Iğdır, Kahramanmaraş, Karabük, Karaman, Kars, Kastamonu, Malatya, Sakarya, Tokat
Küme 2	Van
Küme 3	Adıyaman, Ağrı, Bartın, Batman, Bayburt, Bilecik, Bingöl, Bitlis, Çankırı, Erzincan, Erzurum, Giresun, Gümüşhane, Isparta, Kırşehir, Kütahya, Mardin, Muş, Nevşehir, Niğde, Ordu, Rize, Siirt, Sivas, Şanlıurfa, Şırnak, Trabzon
Küme 4	Ardahan, Osmaniye
Küme 5	Hakkâri, Kilis
Küme 6	Ankara, Antalya, Bursa, Denizli, Gaziantep, İstanbul, Kayseri, Konya
Küme 7	Diyarbakır, Kırklareli, Tunceli, Aksaray, Kırıkkale, Sinop, Yozgat
Küme 8	Adana, Aydın, Balıkesir, Bolu, Çanakkale, Edirne, Eskişehir, İzmir, Kocaeli, Manisa, Mersin, Muğla, Samsun, Tekirdağ, Uşak, Yalova, Zonguldak

Kümeleme analizi sonuçlarından Tek Bağlantı ve Tam Bağlantı yöntemlerinin birbirine benzerlik gösterdiği ve her iki yöntemle de küme geçerliliği indekslerinden en iyi kümelemenin 5 küme olduğu tespit edilmiştir. Ayrıca Ward ve $k-$ortalama yöntemlerinin birbirine benzerlik gösterdiği ve her iki yöntemle de küme geçerliliği indekslerinden en iyi kümelemenin 7 küme olduğu tespit edilmiştir.

Ayrıca uygulama çalışması için Temel bileşen değerleri sonuçlarının kümeleme analizi sonuçları ile benzerlik gösterdiği gözlenmektedir.



Şekil 4.5 Temel Bileşen Skor Değerlerinin Grafiği

Şekil 4.5 incelendiğinde, Van (78 nolu birim), İstanbul (40 nolu birim), Edirne (28 nolu birim) ve Kilis'in (51 nolu birim) diğer birimlerden ayrıldığını yani tek başlarına küme oluşturduğunu, Ardahan (9 nolu birim) ve Osmaniye'nin (64 nolu birim) birbiri ile benzeştiği, Ağrı (4 nolu birim), Hakkari (36 nolu birim) ve Şırnak'ında (72 nolu birim) birbiri ile benzeştiği gözlenmektedir.

5. SONUÇLAR ve TARTIŞMA

Kümeleme Analizi, çok boyutlu uzayda birbirine yakın olan gözlemlerden meydana gelen grupları veya kümeleri bulmayı amaçlamaktadır. Diğer bir ifade ile analiz, örneklem verilerini gözlemlerin benzerliklerine göre en uygun kümelere ayırmaktadır. Kümeleme Analizi, kümelerin sayısına veya küme yapılarına ilişkin herhangi bir varsayımda bulunmaz. Diğer çok değişkenli istatistiksel analiz yöntemlerinde önemli bir yer tutan normallik varsayımı, bu analizde prensipte kalmakta ve uzaklık değerlerinin normalliği yeterli görülmektedir (Tatlıdil 1996). Bunun yanı sıra Kümeleme Analizi araştırmanın amacına bağlı olarak değişkenleri veya bireyleri gruplama imkânı sağlar.

Kümeleme analizi veri nesnelerini yalnızca nesneleri tanımlayan ve ilişkilerini ortaya koyan verilerden çıkarılacak bilgiler ışığında gruplar. Amaç aynı grup içerisindeki nesnelerin birbirine benzer veya ilişkili olması; farklı gruptakilerin ise birbirinden farklı olması ya da ilişkilerinin bulunmamasıdır. Aynı gruptakilerin birbirine benzeme oranı ya da farklı gruptakilerin ise birbirinden farklı olma oranları kümelemenin ne kadar iyi olduğunun ya da kümelerin birbirlerinden ne kadar kesinlikle ayrıldıklarının göstergesidir.

Kümeleme analizindeki en önemli problemlerden birisi en uygun küme sayısının bilinmemesidir. Herhangi bir kümeleme analizi tekniği uygulandıktan sonra yapılan kümelemenin kalitesini değerlendirmek için küme geçerliliği teknikleri kullanılmaktadır. Bu tez çalışmasında illerin farklı suç türlerine göre kümeleme analizi ile değerlendirilmesi yapılmış ve farklı kümeleme teknikleri ile oluşturulan kümelerin geçerlilikleri Silhouette, Calinski ve Harabazs, Krzanowski ve Lai indeksleri ve Wilk's Lambda İstatistiği ile değerlendirilmiştir. Bu indeksler küme içi değişim ve kümeler arası değişim ölçüleri arasındaki ilişkilere dayanmaktadır.

Elde edilen sonuçlardan tüm indekslerde çoğunlukla tek bağlantı ve tam bağlantı yöntemlerine göre uygun küme sayısı 5 iken, Ward ve $k-$ ortalama yöntemlerine göre uygun küme sayısı 7 olarak belirlenmektedir. Analizler sonucunda İstanbul, Hakkâri ve

Van illerinin çoğunlukla tek başına küme oluşturdukları, Ardahan ve Osmaniye'nin bir küme, büyük illerimizden olan Ankara, Antalya, Bursa, Denizli, Gaziantep, İzmir, Kayseri ve Konya'nın kendi arasında bir küme oluşturduğu ve geri kalan diğer tüm illerin bir küme oluşturdukları gözlenmektedir. Van ilinde kaçakçılık suçundan hüküm giyen, Hakkâri ilinde ise uyuşturucu ve uyarıcı madde imal ve ticareti suçundan hüküm giyenlerin nüfuslarına oranla diğer illerden fazla olduğundan ve İstanbul gibi en büyük kentimizin tek başına küme oluşturmaları beklenebilir. Van ve Hakkari'nin sınır ili olması ve işlenen suçların uyuşturucu ve kaçakçılık olması dikkat çekicidir. Ayrıca küme yapıları incelendiğinde (Ağrı, Hakkari, Kilis, Siirt, Şırnak) gibi doğu illerini bir kümeye, Ankara, Antalya, Bursa, Denizli,Gaziantep, İstanbul, İzmir, Kayseri, Konya, Uşak illerini aynı kümeye toplayan Ward yönteminin diğer kümeleme yöntemlerine göre daha anlamlı sonuç verdiği gözlenmiştir.

Farklı suç türlerine göre illerin kümelendirilmesi uygulamasından elde edilen sonuçlar incelendiğinde hemen hemen tüm durumlarda tek bağlantı ve tam bağlantı yöntemlerinin ve ayrıca Ward ve $k-$ortalama yöntemlerinin birbirine benzerlik gösterdiği tespit edilmiştir. Bununla birlikte Wilk's kriterine göre tüm kümeleme yöntemlerinde diğer kümelere göre daha fazla sayıda uygun küme sayısına ulaşılmıştır.

Ayrıca tezin konusu olmamakla beraber çok değişkenli istatistiksel analiz yöntemlerinden temel bileşenler analiz yöntemi ile de illerin durumu incelenmiştir. Temel bileşen değerleriyle illerin gruplandırılması ile kümeleme analizi sonuçları benzerlik gösterdiği gözlenmiştir.

KAYNAKLAR

Anderberg, M.R. 1973. Cluster Analysis for applications. Academic Press, New York. Page 553–555.

Anonim. 2006. Adalet İstatistikleri 'Suç Türü ve Suçun işlendiği ile göre hükümlüler,2006'.http://www.tuik.gov.tr/PreIstatistikTablo.do?istab_id=28. Erişim tarihi:15.04.2008

Appiahene-Gyamfi, J. 2002. 'An Analyses of the Broad Crime Trends and Patterns in Ghana', Journal of Criminal Justice, 30, 229–243.

Atasoy, S. 2001. ' Suç Önleme ve Denetlemede Coğrafi Bilgi Sistemlerinin Kullanımı: Suç Haritalama, Coğrafi Bilgi Sistemi Destekli Trafik Kaza Analizi'. Coğrafi Bilgi Sistemleri Bilişim Günleri, Fatih Üniversitesi, İstanbul.

Blashfield, R.K. and Aldenferder, M.S. 1978, "The Literature on Cluster Analysis", Multivariate Behavioral Research,13,p.271-295

Calinski, R.B. and Harabasz, J. 1974. A dendirite method for cluster analysis . Communications in Statistics 3, p.1-27.

Chatfield, C. and Collins, H. 1980, J.A.Introduction to Multivariate Analysis, Chapman and Hall, p.224

Cox, T.S., Lookhart, G.L.,Walker, D.E., Harrel,L.G., Albers, L.D, and Rodgers, D.M. 1985. Genetic Relationship Among Hard Red Winter Wheat Cultivars as Evaluated by Pedigree Analysis and Glidain Polyacrylamid Gel Electrophoretic Patterns.Crop Science, Vol.25, p.1058-1062.

Çakır, F. 1994. 'Karşılıklı Bağımlılığın Ölçülmesinde Kümeleme Analizi ve Bir Uygulama. Marmara Üniv., Sosyal Bilimler Enst. Ekonomometri A.B.D., Yüksek Lisans Tezi.

Dibb, S. 1998. Market Segmentation: Strategies for success, Marketing Intelligence&Planning, 16/7, pp. 394–406.

Dinçer, K.S. ve Özdamar, K. 1992. Kümeleme çözümlemesinde uygun kümeleme ölçütlerinin karşılaştırılması. Hacettepe Fen ve Mühendislik Bilimleri Dergisi. 14: 17–33.

Dönmezer, S. 1994. Kriminoloji. Beta Basım İstanbul.

Everitt, B. 1974. Cluster Analysis. Heinmann.London, p.122

Everitt, B.S. 1979.Unresolved problems in cluster analysis, Biometrics, 35, p. 169-181

Everitt, B., Landau, S., and Leese, M. 2001, Cluster Analysis, Oxford University Pres, London.

Franco, J., Crossa, J., Villasenor, J., and Eberhat, S.A. 1997. Plant Genetic Resources. Published in Crop Sciences, V.37 (3).p.972–980

Geler, D. 2005. Sosyo-ekonomik değişkenliklerine göre illerin kümelenmesi, Gazi Üniv. Fen Bilimleri Enst. İstatistik Anabilim Dalı, Yüksek Lisans Tezi. Ankara.

Green, E.P. 1989. Analysing Multivariate Data, Philadelphia, p.427

Han, J. and Kamber, M. 2001. Data Mining Concepts and Techniques, Morgan Kaufmann Publishers Inc n Publishers Inc.

Haris, R.J. 1975. A primer of Multivariate Statistics. Academic Pres New York . p.101-115

Harries, K. 1999. Mapping Crime: Principle and Practice, Web sitesi www.ojp.usdoj.gov./nij/pubs-sum/178919.htm Erişim tarihi: 06.05.2008

Hartigan, J.A. 1975. 'Clustering Algorithms' ,Wiley New York, http://statsoft.com/textbook/esc.html., 01.01.2008.

Hawkins , D.M. 1982. Topics in Applied Multivariate . Camridge University Pres. p.150-165

Hofman I. and Jarvis, R. 1998. Robust and Efficient Cluster Analysis Using a Shared Near Neighbours Approach. 14.Pattern Recognition Conference-Volume 1.p.243

Hubert, L. 1974, "Approximate Evaluation Techniques for the Single-Link and Complete-Link Hierarcihal Clustering Procedures", Journal of the American Statistical Association,69, 698-704.

Işık, M. 2006. Kümeleme Yöntemleri ile Veri Madenciliği Uygulamaları, Marmara Üniv. Bilgisayar Müh. Bilimleri, Bilgisayar ve Kontrol, Yüksek Lisans Tezi. İstanbul.

Johnson, A.R. and Wichern, D.W. 1988. Applied Multivariate Statistical Analysis. Prentice-Hall International Editions. New Jersey, p.554.

Kantardzic, M. 2003. Data Mining: Concepts, Models and Algorithms, IEEE Press and John Wiley, New York.

Karakaş, E. 2004. 'Elazığ Şehrinde Hırsızlık Suç Dağılışı ve Özellikleri'. Fırat Üniversitesi Sosyal Bilimler Dergisi. 14(1), 19–37.

Kaufman, L. and Rousseeuw, P.J. 1990. Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley and Sons.

Kinnear, C.T. and Taylor, R.J. 1979.Marketing Research, Mc Grow Hill Book Co. New York, p. 596–599.

Kosaki, T. and Juo, S.R.A. 1988. Soil Grouping Technique by Cluster Analysis.Department of Soil Crop Sciences , Texas A.& M.University. College Station.Texas.77843 U.S.A.

Krzanowski, W.J. and Lai, Y.T. 1985. A criterion for determining the number of groups in a data set using sum of squares clustering.Biometrics 44, p.23-44.

Krzanowski, W.J. 1993. Principles of Multivariate Analysis .' A users perspective clarendon pres .Oxford. p.150-178

Lebeda, A. and Jendrulek, T. 1987. Cluster Analysis as a Method for Evaluation of Genetic Similarity in Specific Host-Parasite İnteraction. Theor Appl Genet 75: 194- 199.

Line, C.S. and Butler, G. 1990. Cluster Analysis for Analysing Two-way Classification Data .Published in Crop Sciences .V.82(2). p.344–348

Marriot, F.H.C. 1971, Practical Problems in a method of Cluster Analysis ,Biometrics, 27.p.501-514

Bolshakova, N. and Azuaje, F. Cluster validation techniques for genome expression data. Web sitesi. www.cs.tcd.ie/publications/tech-reports/reports.02/TCD-CS-2002-33.pdf. Erişim tarihi: 12.02.2008.

Öztürk, İ. 1999. (n<p) Boyutlu verilerde Farklı Kümeleme Yöntemlerinin karşılaştırılmalı olarak incelenmesi , Harran Üniv., Fen Bil.Enst.Zootekni Anabilimdalı, Doktora Tezi.Urfa

Romesburg, H. 1984. Cluster analysis for researchers, Malabar.

Rousseuw, P.J. 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.Journal of Computational and Applied Mathematics 20.p.53-65.

Ruiz, F.J.M. 1998. Strategic Group Analysis in Strategic Marketing: An application to Spanish Saving Banks, Marketing intelligence & Planning, 16/4.

Seyhan, K. 2002. 'Polislik ve Suçun Önlenmesi' Türkiyede Devlet Toplum ve Polis (Editör. Hasan Hüseyin Çevik, Turkut Göksu) Seçkin Yayınları.Ankara.

Seber, G.A.F. 1984. Multivariate Observation , John Wiley and Sons, Inc. New York.

Silahtaroğlu, G. 2004. Veri Madenciliğinde Kümeleme Analizi ve Öğretim Başarısının Değerlemesine ilişkin bir Uygulama, İstanbul Üniv. Sosyal Bilimler Enst. İşletme Anabilim Dalı, Sayısal Yöntemler Dalı, Doktora Tezi. İstanbul.

Sharma, S. 1996. Applied Multivariate Techniques, John Wiley and Sons Inc, New York.

Tatlıdil, H. 1996. Uygulamalı Çok Değişkenli Analiz, Ankara. S.329-343.

Terlemez, L. 2001. Kümeleme Analizi ile Avrupa Birliğine aday ülkelerin ekonomik durumlarının incelenmesi, Anadolu Üniv. Fen Bil. Enst. İstatistik Anabilimdalı, Yüksek Lisans Tezi. Eskişehir.

Yazgan, E. 2001. Kümeleme Analizi Yöntemlerinin Karşılaştırılmalı olarak incelenmesi ve Tarımsal araştırmalarda kullanılması, Çukurova Üniv., Fen Bil. Enst. Zootekni Anabilimdalı, Yüksek Lisans Tezi. Adana.

Yılmaz, O. 1996. 'Kümeleme Analizi ve Matemeatiksel Proğramlama Teknikleri'. Gazi Üniv. Fen Bil.Enst., Yüksek Lisans Tezi.Ankara .S.33-35.

Yılmaz, A. ve Günayergün, S. 2004. 'Türkiye'de Şehir Asayiş Suçları; Dağılış ve Başlıca Özellikleri'http:\\yayim.meb.gov.tr\dergiler.Erişim tarihi:15.01.2008

EK 1 SUÇ TÜRÜ VE SUÇUN İŞLENDİĞİ İLE GÖRE CEZA İNFAZ KURUMUNA GİREN HÜKÜMLÜLER (2006)

	Adam (X1) Öldürme	Cinsel suçlar(X2)	kisiyi hür.yoks bırak(x3)	Hırsızlık(x4)	Yağma (Gasp)(x5)	Dolandırıcılık (x6)	uyuşturucu (x7)	Sahtecilik (x8)	Zimmet (x9)	Kaçakçılık (x10)	Orman suçları (x11)
ADANA	0,00383	0,00192	0,00091	0,01276	0,00197	0,01724	0,00454	0,00373	0,00030	0,00136	0,00055
ADIYAMAN	0,00059	0,00030	0,00000	0,00221	0,00000	0,00325	0,00133	0,00428	0,00000	0,00074	0,00074
AFYON	0,00326	0,00157	0,00109	0,00616	0,00109	0,00845	0,00024	0,00121	0,00012	0,00036	0,00072
AĞRI	0,00158	0,00053	0,00018	0,00263	0,00035	0,00263	0,00315	0,00035	0,00000	0,00823	0,00000
AKSARAY	0,00399	0,00258	0,00023	0,00846	0,00164	0,00376	0,00164	0,00094	0,00047	0,00023	0,00000
AMASYA	0,00227	0,00028	0,00085	0,00681	0,00085	0,00596	0,00000	0,00170	0,00057	0,00000	0,00539
ANKARA	0,00210	0,00146	0,00023	0,01170	0,00230	0,02296	0,00388	0,00333	0,00025	0,00123	0,00014
ANTALYA	0,00439	0,00140	0,00048	0,01058	0,00270	0,02270	0,00382	0,00309	0,00019	0,00101	0,00097
ARDAHAN	0,00336	0,00000	0,00168	0,00839	0,00000	0,00503	0,00000	0,00084	0,00000	0,00000	0,02182
ARTVİN	0,00179	0,00179	0,00000	0,00776	0,00060	0,00895	0,00000	0,00418	0,00000	0,00060	0,00537
AYDIN	0,00461	0,00160	0,00130	0,01342	0,00210	0,01212	0,00230	0,00290	0,00000	0,00130	0,00070
BALIKESİR	0,00363	0,00182	0,00100	0,01081	0,00064	0,01126	0,00154	0,00245	0,00009	0,00073	0,00100
BARTIN	0,00313	0,00000	0,00000	0,00376	0,00063	0,00376	0,00251	0,00000	0,00000	0,00063	0,00313
BATMAN	0,00116	0,00000	0,00000	0,00656	0,00000	0,00386	0,00116	0,00019	0,00000	0,00135	0,00058
BAYBURT	0,00000	0,00000	0,00000	0,00234	0,00000	0,00468	0,00000	0,00000	0,00117	0,00000	0,00000
BİLECİK	0,00151	0,00000	0,00050	0,00654	0,00000	0,00352	0,00050	0,00101	0,00000	0,00050	0,00000
BİNGÖL	0,00041	0,00000	0,00000	0,00449	0,00000	0,00326	0,00489	0,00041	0,00000	0,00163	0,00000
BİTLİS	0,00097	0,00048	0,00000	0,00169	0,00000	0,00073	0,00024	0,00000	0,00000	0,00024	0,00048
BOLU	0,00228	0,00000	0,00038	0,00797	0,00038	0,01328	0,00076	0,00152	0,00000	0,00114	0,00152
BURDUR	0,00284	0,00122	0,00000	0,00650	0,00081	0,00934	0,00081	0,00081	0,00000	0,00081	0,00244
BURSA	0,00257	0,00137	0,00021	0,00924	0,00228	0,02191	0,00282	0,00315	0,00041	0,00083	0,00066
ÇANAKKALE	0,00235	0,00064	0,00021	0,00942	0,00043	0,01135	0,00128	0,00128	0,00021	0,00064	0,00214
ÇANKIRI	0,00293	0,00146	0,00037	0,00110	0,00000	0,00476	0,00037	0,00183	0,00000	0,00037	0,00110
ÇORUM	0,00143	0,00178	0,00018	0,00642	0,00018	0,00767	0,00107	0,00054	0,00000	0,00036	0,00446
DENİZLİ	0,00408	0,00102	0,00147	0,01042	0,00170	0,02696	0,00102	0,00374	0,00011	0,00136	0,00068
DİYARBAKIR	0,00221	0,00054	0,00027	0,00642	0,00114	0,00709	0,00489	0,00241	0,00013	0,00207	0,00067
DÜZCE	0,00425	0,00091	0,00030	0,00243	0,00061	0,00759	0,00395	0,00061	0,00000	0,00030	0,00759
EDİRNE	0,00706	0,00235	0,00078	0,01360	0,00235	0,01230	0,00366	0,00314	0,00026	0,00811	0,00078
ELAZIĞ	0,00235	0,00134	0,00017	0,00454	0,00235	0,01008	0,00319	0,00286	0,00000	0,00235	0,00050
ERZİNCAN	0,00095	0,00095	0,00064	0,00476	0,00095	0,00286	0,00032	0,00191	0,00000	0,00032	0,00000
ERZURUM	0,00282	0,00073	0,00063	0,00428	0,00125	0,00615	0,00104	0,00083	0,00000	0,00021	0,00094
ESKİŞEHİR	0,00333	0,00250	0,00028	0,00763	0,00139	0,01208	0,00222	0,00333	0,00014	0,00139	0,00097
GAZİANTEP	0,00315	0,00217	0,00014	0,01962	0,00308	0,02362	0,00624	0,00287	0,00014	0,00147	0,00098
GİRESUN	0,00116	0,00019	0,00097	0,00232	0,00039	0,00387	0,00077	0,00155	0,00019	0,00039	0,00097
GÜMÜŞHANE	0,00209	0,00052	0,00000	0,00261	0,00052	0,00104	0,00157	0,00000	0,00000	0,00052	0,00261
HAKKARİ	0,00073	0,00000	0,00000	0,00073	0,00000	0,00257	0,01578	0,00000	0,00037	0,00550	0,00000
HATAY	0,00215	0,00069	0,00023	0,00869	0,00185	0,00823	0,00338	0,00092	0,00015	0,00308	0,00454
IĞDIR	0,00222	0,00111	0,00000	0,00389	0,00000	0,00945	0,00000	0,00334	0,00056	0,00500	0,00000
ISPARTA	0,00110	0,00091	0,00110	0,00438	0,00091	0,00511	0,00110	0,00164	0,00000	0,00055	0,00037
İSTANBUL	0,00199	0,00096	0,00028	0,01430	0,00316	0,03599	0,00558	0,00397	0,00009	0,00065	0,00024
İZMİR	0,00267	0,00089	0,00032	0,00873	0,00159	0,01774	0,00345	0,00218	0,00016	0,00067	0,00040
K.MARAŞ	0,00271	0,00135	0,00058	0,00899	0,00145	0,00754	0,00261	0,00184	0,00019	0,00116	0,00541
KARABÜK	0,00000	0,00249	0,00000	0,00648	0,00100	0,00598	0,00100	0,00249	0,00000	0,00050	0,00249
KARAMAN	0,00357	0,00475	0,00040	0,00911	0,00119	0,00832	0,00040	0,00158	0,00000	0,00000	0,00238
KARS	0,00627	0,00104	0,00035	0,00627	0,00000	0,00453	0,00104	0,00070	0,00035	0,00174	0,00697
KASTAMONU	0,00341	0,00155	0,00031	0,00589	0,00000	0,00744	0,00031	0,00093	0,00031	0,00093	0,00465
KAYSERİ	0,00383	0,00173	0,00000	0,01122	0,00128	0,02071	0,00027	0,00173	0,00018	0,00119	0,00000
KIRIKKALE	0,00333	0,00103	0,00000	0,00718	0,00051	0,00205	0,00051	0,00180	0,00000	0,00026	0,00051
KIRKLARELİ	0,00428	0,00061	0,00031	0,00856	0,00122	0,00581	0,00489	0,00184	0,00000	0,00092	0,00031
KIRŞEHİR	0,00126	0,00084	0,00042	0,00544	0,00042	0,00461	0,00084	0,00042	0,00000	0,00084	0,00000
KİLİS	0,00612	0,00102	0,00000	0,01123	0,00000	0,00204	0,01123	0,00000	0,00000	0,00510	0,00000
KOCAELİ	0,00177	0,00125	0,00015	0,01000	0,00088	0,01133	0,00154	0,00162	0,00037	0,00118	0,00213
KONYA	0,00274	0,00083	0,00033	0,00672	0,00083	0,02077	0,00091	0,00199	0,00004	0,00083	0,00012
KÜTAHYA	0,00132	0,00146	0,00015	0,00526	0,00000	0,00468	0,00044	0,00088	0,00000	0,00015	0,00249
MALATYA	0,00422	0,00087	0,00000	0,00455	0,00076	0,01029	0,00238	0,00097	0,00022	0,00054	0,00043
MANİSA	0,00391	0,00188	0,00063	0,00993	0,00180	0,01095	0,00360	0,00196	0,00016	0,00094	0,00094
MARDİN	0,00103	0,00026	0,00000	0,00167	0,00051	0,00180	0,00154	0,00103	0,00013	0,00154	0,00180
MERSİN	0,00345	0,00221	0,00016	0,01239	0,00151	0,01336	0,00458	0,00517	0,00005	0,00205	0,00140
MUĞLA	0,00214	0,00189	0,00050	0,00869	0,00227	0,00995	0,00491	0,00202	0,00000	0,00113	0,00088
MUŞ	0,00123	0,00000	0,00041	0,00593	0,00020	0,00164	0,00102	0,00327	0,00000	0,00245	0,00061
NEVŞEHİR	0,00258	0,00226	0,00097	0,00290	0,00032	0,00290	0,00129	0,00129	0,00032	0,00000	0,00000
NİĞDE	0,00137	0,00082	0,00000	0,00355	0,00082	0,00410	0,00109	0,00219	0,00055	0,00027	0,00000
ORDU	0,00168	0,00079	0,00045	0,00359	0,00000	0,00449	0,00000	0,00067	0,00011	0,00022	0,00303
OSMANİYE	0,00122	0,00102	0,00102	0,00630	0,00041	0,00406	0,00102	0,00183	0,00000	0,00081	0,01687
RİZE	0,00249	0,00138	0,00028	0,00387	0,00055	0,00470	0,00304	0,00055	0,00028	0,00055	0,00083
SAKARYA	0,00439	0,00181	0,00065	0,00904	0,00103	0,01046	0,00426	0,00207	0,00000	0,00116	0,00865
SAMSUN	0,00413	0,00168	0,00034	0,01044	0,00059	0,01238	0,00312	0,00253	0,00008	0,00126	0,00581
SİİRT	0,00263	0,00000	0,00000	0,00301	0,00000	0,00113	0,00075	0,00075	0,00000	0,00526	0,00075
SİNOP	0,00538	0,00323	0,00054	0,00538	0,00054	0,00377	0,00161	0,00108	0,00054	0,00000	0,00161
SİVAS	0,00098	0,00084	0,00000	0,00520	0,00014	0,00422	0,00028	0,00056	0,00028	0,00042	0,00028
ŞANLIURFA	0,00200	0,00035	0,00018	0,00535	0,00024	0,00365	0,00118	0,00118	0,00006	0,00194	0,00006
ŞIRNAK	0,00124	0,00074	0,00025	0,00396	0,00000	0,00173	0,00099	0,00025	0,00000	0,01288	0,00025
TEKİRDAĞ	0,00325	0,00226	0,00056	0,00904	0,00169	0,01003	0,00099	0,00169	0,00014	0,00014	0,00042
TOKAT	0,00265	0,00115	0,00046	0,00449	0,00115	0,00403	0,00138	0,00115	0,00023	0,00069	0,00667
TRABZON	0,00170	0,00104	0,00019	0,00273	0,00019	0,00726	0,00000	0,00028	0,00009	0,00019	0,00066
TUNCELİ	0,00524	0,00000	0,00000	0,00654	0,00131	0,00654	0,00131	0,00262	0,00000	0,00000	0,00000
UŞAK	0,00423	0,00212	0,00000	0,00423	0,00000	0,01664	0,00151	0,00060	0,00000	0,00030	0,00333
VAN	0,00128	0,00039	0,00010	0,00257	0,00020	0,00148	0,00504	0,00069	0,00000	0,03170	0,00000
YALOVA	0,00378	0,00108	0,00000	0,00918	0,00162	0,01295	0,00108	0,00270	0,00000	0,00108	0,00486
YOZGAT	0,00467	0,00206	0,00055	0,00481	0,00055	0,00247	0,00124	0,00206	0,00027	0,00027	0,00082
ZONGULDAK	0,00376	0,00394	0,00036	0,01289	0,00036	0,01307	0,00143	0,00143	0,00054	0,00018	0,00519

ÖZGEÇMİŞ

Adı Soyadı : Azize Celile GÜNAY ATBAŞ
Doğum Yeri :Ankara
Doğum Tarihi :23/07/1979
Medeni Hali :Evli
Yabancı Dili : İngilizce

Eğitim Durumu (Kurum ve Yıl)

Lise : Tevfik İleri İmam Hatip Süper Lisesi-(1993-1997)
Lisans : Ankara Üniversitesi Fen Fakültesi, İstatistik Bölümü-(1997-2001)
Yüksek Lisans : Ankara Üniversitesi Fen Bilimleri Enstitüsü , İstatistik Anabilim Dalı (Eylül2005-Ağustos 2008)

Çalıştığı Kurum/Kurumlar ve Yıl

Şekerbank T.A.Ş – Bireysel Müşteri Temsilcisi (2001-2003)
Adli Sicil ve İstatistik Genel Müdürlüğü- Çözümleyici (2003-2008)
Adalet Bakanlığı Strateji Geliştirme Başkanlığı- Şube Müdürü (2008-