Dumlupınar Üniversitesi Sosyal Bilimler Dergisi
Sayı: 3 Kasım 1999

KÜMELEME ANALİZİNDE GEÇERLİLİK PROBLEMİ VE KÜMELEME SONUÇLARININ DEĞERLENDİRİLMESİ

Dr. Zeki ÇAKMAK *

ÖZET

Bu çalışmada, kümeleme analizi genel olarak incelendikten sonra, kümeleme sonuçlarının geçerlilik problemi ele alınmış ve geçerlilik tekniklerinden bazıları gözden geçirilmiştir. Uygulama bölümünde aşamalı kümeleme yöntemleri yardımıyla oluşturulabilecek küme sayıları belirlenmiş ve eğitim yapıları birbirine benzeyen iller farklı küme sayıları için aşamalı olmayan kümeleme tekniklerinden K-ortalamalar yöntemiyle kümelendirilmiştir.

Elde edilen kümelerin geçerliliğini test etmek amacıyla kümeleme sonuçlarını diskriminant analizi uygulanmış ve iller yeniden sınıflandırılmıştır. Sonuç olarak, oldukça yüksek doğru sınıflandırma oranları bulunmuş ve K-ortalamalar yöntemiyle elde edilen kümelerin anlamlı olduğu sonucuna varılmıştır.

1. GİRİŞ

Bilimin temel amaçlarından birisi de karmaşık durumları daha makul bir duruma indirgeyerek sınıflandırmaktır. Yeomans (1979) nesnelerin sınıflandırılmasını tüm bilimsel çabaların temelini oluşturan bir çalışma olarak tanımlamıştır. Sınıflandırma problemi, araştırmacının bireyler üzerinde, bireyin çeşitli özellikleri bakımın-

* DPÜ İİBF İşletme Bölümü

dan ölçüm yapması ve elde edilen ölçümlere dayanılarak bireyin belirli bir guruba sınıflandırılmak istemesiyle ortaya çıkmaktadır (Anderson, 1951). Sınıflandırma insan zihninin temel işlevlerinden birisidir. İnsanoğlu herhangi bir bilimsel temele dayanmadan özellikleri kesin olarak birbirinden ayrılan canlı ve cansız varlıkları şekil, büyüklük, tat, koku vb. özelliklerine göre kolaylıkla sınıflandırabilmektedir. Ancak, nesneleri birden çok özelliğe göre sınıflandırmak oldukça güçleşmektedir. Bu durumda sınıflandırma problemini çözmek için genellikle diskriminant analizi; çok boyutlu ölçekleme, kümeleme analizi vb. çok değişkenli istatistik analiz tekniklerine başvurmak bir zorunluluk haline gelmektedir.

2. KÜMELEME ANALİZİ

Kümeleme analizi ile, p adet özelliğe (değişkene) sahip N sayıda bireyin benzerliklerine göre türdeş yapının sağladığı ayrık kümelerde toplanması amaçlanmaktadır (Duran ve Adel, 1974; s.43). Kümeleme analizi birbirine benzer olan bireylerin aynı guruplara toplanmasını amaçlaması bakımından diskriminant analizi ile benzerlikler gösterir, Ancak kümeleme analizinde guruplar, diskriminant analizinde olduğu gibi analiz öncesi değil, bireyler arasındaki benzerlikler belirlendikten sonra oluşturulmaktadır (Çakmak, 1993; s.12). Bu nedenle diskriminant analizinde mevcut veri yapısından elde edilen fonksiyonlar gelecek için tahminlerde kullanılırken, kümeleme analizi sadece mevcut veri yapısına ilişkin sonuçlar verdiğinden gelecekte kullanılma imkanı yoktur. Kümeleme analizi verileri değişkenlere göre de gruplamayı sağladığından, bu uygulanış şekliyle faktör analiziyle benzerlik göstermekte olup bilgileri özetleyici (veri indirgeme) özelliği vardır.

Kümeleme yöntemleri konusundaki çalışmalar 1963 yılında Sokal ve Sneath tarafından "Principles of Numerical Taxonomy" isimli kitabın yayınlanmasından sonra hız kazanarak bu konudaki yayınların sayısı izleyen sekiz yıl içinde ikiye katlanmıştır (Aldenderfer ve Blashfield, 1984). Kümeleme yöntemleri konusunda 1987 yılına kadar yapılan önemli çalışmalar (Milligan ve Cooper, 1987) metedolojik bir şekilde incelenmiştir.

Gerek fen bilimlerinde gerek sosyal bilimlerde oldukça yaygın kullanım alanı bulan kümeleme yöntemlerinin artmasında kuşkusuz bilgisayar teknolojisinin gelişmesi ve bilimsel bir süreç olarak sınıflandırmanın bütün bilim dallarında kullanılma gereğinin büyük rolü olmuştur.

Kümeleme analizinde diğer çok değişkenli istatistik analizlerde olduğu gibi verilerin normalliği varsayımı fazla önemli olmayıp uzaklık değerlerinin normalliği yeterli görülmektedir (Tatlıdil,1992; s.252). Analizde kullanılan değişkenler arasındaki ilişkilerin doğrusal olma zorunluluğu yoktur ve analiz seçilen kümeleme yöntemine göre sınıflayıcı (nominal), sıralayıcı (ordinal), aralıklı (internal), oransal (ratio) veya kategorik ölçekle ölçülen verilere uygulanabilmektedir (Kurtuluş, 1996; s.496, Hartigan 1975; s.9-10), verilen bir veri setinin nitel veya nicel değer alan verilerden oluşabileceğini, bu veri setinin farklı ölçeklerde ölçülen aynı tipteki değişkenlerden (heterojen), ve aynı ölçekte ölçülen değişkenlerden (homojen) oluşabileceğini vurgulayarak, bu tür veri setleriyle uygulamanın nasıl olacağını örneklerle açıklanmıştır (Hartigen, 1975, Chapters 14-16).

Kullanıcının amacına ve kullanım alanına göre kümeleme analizinin amaçları şu başlıklarla özetlenebilir (Everit, 1974; s.3):

- Doğru tiplerin belirlenmesi,
- Model uyumu,
- Gruplara dayalı tahmin,
- Hipotez testleri ve türetimi,
- Veri araştırması ve veri indirgenmesi dir

Kümeleme analizinde genel olarak N adet nesnenin (gözlemin) herbirinde p adet ölçümün yapıldığı N x p boyutlu X veri matrisi aşağıdaki gibidir.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}$$

Matriste yer alan x_{ij} elemanı j'ninci değişkenin i'ninci birey yada nesne için aldığı değerine karşılık gelmektedir.

Kümeleme analizinde genel amaç birbirine benzer olan bireylerin aynı gruplarda toplanmasıdır. Kümelere ayırma işlemi söz konusu iki gözlemin benzerlik (yakınlık) veya uzaklık ölçülerine göre yapılmaktadır. Bu bakımdan bireyler arasındaki benzerliği ölçmede hangi ölçünün kullanılacağı kümeleme analizinin en önemli sorunlarından birisini teşkil eder. Uygulamada sıkça kullanılan başlıca benzerlik ölçüleri (Aldenderfer ve Blashfield, 1984; s.22): (1) korelasyon katsayıları, (2) uzaklık ölçüleri, (3) birliktelik katsayları ve (4) olasılığa dayalı benzerlik katsayılarıdır. Korelasyon katsayısı ve uzaklık ölçüleri özellikle sosyal bilimlerde oldukça yaygın olarak kullanılmaktadır.

Korelasyon katsayısı genellikle iki değişken arasındaki ilişkiyi belirten bir katsayıdır. Ancak, sayısal sınıflandırmada iki gözlem vektörü arasındaki ilişkiyi belirlemek için de kullanılmaktadır. x_i ve x_j gözlemleri arasındaki ilişkinin ölçüsü olan Karl Pearson'un Çarpım Moment Korelasyon Katsayısı (Aldenderfer ve Blashfield, 1984; s.22);

$$r_{jk} = \frac{\sum (x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k})}{\sqrt{\sum (x_{ij} - \overline{x_j})^2 \sum (x_{ik} - \overline{x_k})^2}}$$

şeklinde olup , x_{ij}, j gözlem için i. değişkenin değerini, $\overline{x_j}$ ise j. gözlem için bütün değişken değerlerinin ortalamasıdır. Bu yöntem değişkenlerin oransal veya aralıklı ölçekte ölçülmesi durumunda kullanılır. Katsayı değerleri [-1,1] aralığında olup, 0 değeri j. ve k. gözlemler arasında benzerlik olmadığını gösterir. r_{jk}=1 olması

j. ve k. gözlemler arasında tam bir ilişkiyi göstermesine rağmen x_j ve x_k gözlemlerinin aynı olduğunu göstermez.

Bireyleri kümelendirmede benzerlik (similarity) yada benzemezliğin (dissimilarity) ölçüsü olan sayısal değere uzaklık ölçüsü (distance measure) adı verilmektedir.

x_i ve x_j gözlem vektörleri arasındaki $d(x_i,x_j)=d_{ij}$ uzaklık değerini ifade etmek amacıyla geliştirilmiş ve çoğu istatistik paket programlarında yer alan pek çok uzaklık ölçüsü bulunmaktadır. Söz konusu ölçülerin üst sınırı olmamakla birlikte ölçeğe bağımlıdırlar ve aşağıdaki koşulları sağlamaları gerekir (Everit, 1974; s.56).

$d(x_i,x_j) \geq 0$; eğer i = j ise $d(x_i,x_j) = 0$
$d(x_i,x_j) = d(x_j,x_i)$ simetri özelliği
$d(x_i,x_j) \leq d(x_i,x_k) + d(x_k,x_j)$ üçgen eşitsizlik özelliği

Bu özellikleri taşıyan N x N boyutlu simetrik D uzaklıklar matrisi

$$D = \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1N} \\ & 0 & d_{23} & \cdots & d_{2N} \\ & & O & & M \\ & & & O & d_{N-1,N} \\ & & & & 0 \end{bmatrix}$$

şeklindedir.

X(N x p) veri matrisi, x_{ik}, i. gözlem için k. değişkenin değeri olmak üzere i. ve j. gözlemler arasındaki uzaklıkları veren ve en çok bilinen uzaklık ölçüleri aşağıdaki gibidir.

Öklit uzaklığı : $$d_{ij} = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right]^{1/2}$$

City Block Uzaklığı : $$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

Minkowski Uzaklığı : $$d_{ij} = \left\{ \sum_{k=1}^{p} |x_{ik} - x_{jk}|^r \right\}^{1/r}$$

Yukarıda kullanılan uzaklık ölçüleri kullanışlı olmakla birlikte değişkenlerin ölçü birimlerinden kolaylıkla etkilenirler. Örneğin, belirli bir ölçü biriminde iki birey birbirlerine en uzak olurken, ölçü birimleri değiştiğinde birbirlerine daha yakın hale gelerek bireyler arasındaki uzaklıkların sırası değişebilmektedir. Bu nedenle uzaklık hesaplamasından önce değişkenlerin standartlaştırılması yoluna gidilmelidir (Aldenderfer ve Blashfield, 1984; s.26).

Değişkenlerin ölçü birimlerinin farklı olması durumunda ölçek dezavantajlarına alternatif olarak geliştirilen ve öklit uzaklığının standartlaştırılmış şekli olan Karl Pearson uzaklığı;

$$d_{ij}^2 = \sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{S_k^2}$$

Burada S_k^2, k. değişkenin varyansıdır (Mardia, Kent and Bibby, 1979). Bu yöntemde değişkenler standartlaştırıldığı için bireyler arasındaki nisbi uzaklıklar korunmaktadır. Ancak, bu durumda da kümeler arasındaki belirgin farklılıklar azalmaktadır (Everitt, 1974; s.48-49). Başka bir ifade ile standartlaştırma işlemi gruplar arasındaki farklılıkları maksimum yapacak yerde minimum yaparlar (Manly, 1994; s.134).

Öklit ve City Block uzaklık ölçüleri değişkenlerin birbirinden bağımsız olduğunu varsayar. Değişkenler arasındaki korelasyonları da dikkate alan ve Mahalanobis D^2 uzaklığı olarak bilinen metrik aşağıdaki gibi tanımlanır.

$$d_{ij} = D^2 = (x_i - x_j) S^{-1} (x_i - x_j)$$

Formülde yer alan S^{-1} gruplar içi varyans-kovaryans matrisini, x_i ve x_j ise i. ve j. gözlem vektörlerinin değerlerini göstermektedir.

Yukarıda sözü edilen uzaklık ölçülerinin uygulanabilmesi için verilerin en azından aralıklı veya oransal ölçekte ölçülmüş olması gerekir.

Kümeleme analizi için hazırlanan veri setinden benzerlik veya uzaklık ölçülerinden birisinin kullanılması suretiyle benzerlik matrisi veya uzaklık matrisinin oluşturulmasından sonra, kümeleme yöntemlerinden birisi ile bireylerin kümelere (gruplara) atanması yapılır.

Milligan ve Cooper (1987) literatürde mevcut yüzden fazla kümeleme yöntemini aşamalı, aşamalı olmayan, yığma veya klik ve boyutsal gösterim yöntemleri olmak üzere dört kategoriye ayırmaktadır.Ancak, Anderberg (1973) tarafından yapılan ve genel kabul gören kümeleme teknikleri ise aşamalı ve aşamalı olmayan kümeleme teknikleri olarak sınıflandırılmaktadır.

Aşamalı kümeleme tekniklerinde herhangi bir uzaklık ölçüsü yardımıyla oluşturulan uzaklık matrisinden yararlanılarak birbirine en yakın yada en benzer birimden başlanılarak gözlemler birbirine bağlanarak tüm bireyler bir kümede toplanacak şekilde ağaç diyagramı oluşturulur.

N adet birimin kümelenmesi için aşamalı yöntemlerde kullanılan algoritmaların genel adımları aşağıdaki gibidir (Tatlıdil, 1992; Everitt, 1993):

- Birimler arasındaki uzaklıkları N x N boyutlu simetrik bir matrisin ($D=\{d_{ij}\}$) gösterdiği N küme ile işleme başlanır.

- Birbirine en yakın (D matrisindeki en küçük değer) iki küme birleştirilir.
- Küme sayısı bir indirgenerek yinelenmiş uzaklar matrisi bulunur.
- Tek küme oluşuncaya kadar 2. ve 3. adımlar N-1 kere tekrarlanır.

İki kümenin birleştirilmesi kriterinin belirlenmesine ilişkin farklı aşamalı teknikler bulunmaktadır. Bu tekniklerden en yaygın kullanımı olanlar ve çoğu istatistik paket programlarında da bulunabilen tekli bağlantı (en yakın komşuluk) tam bağlantı (en uzak komşuluk), ortalama bağlantı ve birim sayısı fazla olduğundan (N>50) iyi sonuçlar veren Ward Kümeleme teknikleridir.

Aşamalı olmayan kümeleme teknikleri değişkenlerden ziyade N adet bireyin k sayısındaki kümeye dağıtılması problemi ile ilgilenilir. Küme sayısı konusunda ön bilgi var ise veya gözlem sayısı fazla ise aşamalı olmayan kümeleme tekniklerinin kullanılması tercih edilmektedir. Çünkü, bu yöntemlerde uzaklıklar matrisinin kullanılması zorunluluğu yoktur ve doğrudan ham verilerle (X matrisi ile) çalışılır (Aldenderfer and Blashfield, 1984; s.47).

Aşamalı olamayan kümeleme tekniklerinden en yaygın olarak kullanılanı Mac Queen tarafından geliştirilen K-ortalamalar tekniğidir. Bu teknikte bireyler kümeler içi kareler toplamı minimum olacak şekilde k adet kümeye bölünmektedir. Bu tekniğin algoritması aşağıdaki gibidir (Tatlıdil, 1992; s.258-259):

İlk k gözlem her birisi bir gözlemli küme olarak alınır.
Kalan N-1 gözlemin her biri en yakın kümeye atanmakta ve atanma işleminden sonra küme ortalamaları yeniden hesaplanmaktadır.

Tüm birimlerin kümelere atanması tamamlandıktan sonra 2. adımda hesaplanan ortalamalara göre yeniden atanmaları yapılmaktadır.

Bu atama işlemleri kümeler arasındaki gözlem geçişi duruncaya kadar tekrarlanır.

Aşamalı kümeleme tekniklerinin uygulanması ile küme sayısı hakkında bilgi sahibi olunacağından aşamalı olmayan K-ortalamalar yöntemini uygulamadan önce aşamalı kümeleme tekniğinin uygulanması yerinde olacaktır.

Kümeleme işlemini organize etmek için yedi adımdan oluşan bir yapı kullanılmaktadır. Yapılacak uygulamaya göre değişebilecek bu adımlar şöyle sıralanmaktadır (Milligan ve Cooper, 1987; s.33):

- Kümelendirilecek birimler seçilmelidir. Örnek elemanları kümenin genel yapısını temsil edecek şekilde seçilmiş olmalıdır.
- Kümeleme analizinde kullanılacak değişkenler seçilmelidir. Değişkenler, bireylerin kümelenmesine izin verecek yeterli bilgiyi içermelidir.
- Araştırmacı verilerin standartlaştırılıp standartlaştırılmayacağına karar vermelidir.
- Analizde kullanılacak uzaklık veya benzerlik ölçütü belirlenmelidir.

- Araştırmanın amacına uygun kümeleme yöntemi seçilmelidir. Çünkü farklı yöntemlerle farklı sonuçlara ulaşılabilmektedir.
- Küme sayısı belirlenmelidir.
- Kümeleme analizindeki son ve en önemli adım olup yorum, test ve uygulanabilirliktir. Yorum araştırmacının uygulama alanı hakkında özel bilgi sahibi olmasını gerektirir. Test yapılan kümelendirmenin anlamlı olup olmadığı probleminin belirlenmesini içermektedir. Uygulanabilirlik ise elde edilen sonuçların diğer örneklere uygulanıp uygulanamayacağının belirlenmesidir.

İzleyen kesimde 6. ve 7. adımlar ayrıntılı olarak incelenmeye çalışılacaktır.

3. KÜMELEME SONUÇLARININ GEÇERLİLİĞİNİN İNCELENMESİ

Kümeleme analizinde geçerli ve anlamlı sonuçlara ulaşabilmek için iki koşulun mutlaka sağlanması gerekir. Bunlardan ilki önemli değişkenlerin seçilmesi ikincisi ise küme sayısının isabetli seçilmesidir. Önemli değişkenlerin seçilmesi için verilere adımsal regresyon analizi ve temel bileşenler analizi gibi yöntemler uygulanabilir.

3.1 Küme Sayısının Belirlenmesi

Kümeleme yöntemlerinin çoğu analizi sonunda oluşacak küme sayısını belirlenmesini gerektirmektedir. Özellikle aşamalı olmayan kümeleme yöntemlerinde analize başlamadan önce küme sayısının belirlenmesi gerekir. Aşamalı yöntemlerde analiz öncesi böyle bir belirleme gerekmemektedir. Aşamalı kümeleme yöntemlerinde elde edilen dendogram yardımıyla küme sayısına görsel olarak karar verilebilmektedir. Bu nedenle verilere aşamalı olmayan kümeleme yöntemlerini uygulamadan önce aşamalı yöntemler uygulanarak muhtemel küme sayıları belirlenebilir (Punj ve Stewart, 1983; s.145).

Küme sayısına karar vermek için geliştirilen ölçütlerden bazıları şöyledir (Tatlıdil, 1992; s.260-264): N gözlem sayısını göstermek üzere küme sayısı $k \cong (N/2)^{1/2}$ ifadesi ile hesaplanabilir. Ancak gözlem sayısının artması küme sayısını da anlamsız olarak arttırmaktadır. Marriot (1971)'un önerdiği yöntemde ise, W, grup içi kareler toplamını göstermek üzere $M=k^2|W|$ ilişkisini sağlayan en küçük M değerini veren küçük k değeri gerçek küme sayısı olarak değerlendirilmektedir. Calinsky ve Harabasz (1974) tarafından geliştirilen yöntemde B gruplar arası, W gruplar içi kareler toplamı matrislerini göstermek üzere

$$C = [iz(B)/(k-1)]/[iz(W)/(N-k)]$$

eşitliğini maksimum yapan k değeri küme sayısı olarak alınmaktadır. Milligan ve Cooper (1985) küme sayısının belirlenmesi konusunda önerilen diğer yöntemleride inceleyerek söz konusu yöntemlerin hangi durumlarda daha etkin sonuç verdiklerini ortaya koymuşlardır. Küme sayısının belirlenmesi konusunun geçerlilik işlemleri ile birlikte ele alınması daha uygun olacaktır. Çünkü küme

sayısını belirleme yöntemlerinin birisi ile önerilen küme sayısı başka bir ölçüt ile geçerli olmayabilir.

3.2 Geçerlilik Yöntemleri

Kümeleme sonuçlarının geçerliliği sorunu kümeleme analizinin en zor ve sıkıcı kısmıdır. Veri setinin analizi ve kümeleme çözümünün elde edilmesinden sonra ulaşılan sonuçların anlamlılığı ve güvenilirliği konusunda bir garanti yoktur. Verilerde herhangi bir tabi gruplanma olmasa dahi bi kümeleme sonucuna ulaşılacaktır. Bu bakımdan elde edilen kümeleme çözümünün tesadüfi çözümden farklı olup olmadığının bazı test veya testler uygulanarak belirlenmesi gerekir. Kümeleme çözümlerinin kalitesini test etmek amcıyla geliştirilen bazı yöntemler aşağıda verilmiştir.

1. Cophenetic Korelasyonu

Sadece aşamalı yığma kümeleme yöntemleri kullanıldığında geçerli olan bu yöntem ilk olarak Sokal ve Rohlf (1962) tarafından önerilmiş ve geçerliliği Sneath ve Sokal (1973) tarafından kanıtlanmıştır (Aldenderfer ve Blashfield, 1984; s.62). Söz konusu korelasyon ağaç veya dendogramların bireyler arasındaki benzerlik veya benzemezliklerinin derecelerini başka bir ifade ile dendogramların performansını belirlemek için kullanılır.

Cophenetic korelasyonu orjinal benzerlik matrisindeki değerlerle dendogram kullanılarak elde edilen benzerlik matrisinin (implied similarity matrix) değerleri arasındaki korelasyondur. Birbiri ile ilişkilendirilen matrislerdeki değerlerin normal dağılım varsayımına uymaması ve iki matrisin taşıdıkları bilgilerin farklı olması durumunda, söz konusu korelasyonun kümeleme sonuçlarının kalitesi hakkında yanlış anlamaya yol açan bir gösterge olduğu yapılan Monte Carlo çalışmalarıyla gösterilmiştir (Aldenderfer ve Blashfield, 1984 ; s.63-64).

2. İstatistik Önem Testleri

Kümeleme sonuçlarının geçerliliğini başka bir ifade ile kümelerin anlamlılığını test etmenin diğer bir yolu da kümeleri oluşturmada kullanılan değişkenler için çok değişkenli varyans analizi veya diskriminant analizinin uygulanmasıdır. Aşamalı ve aşamalı olmayan kümeleme çözümleri için uygulanabilen varyans analizi yüksek seviyede anlamlı sonuçlar vermesi nedeniyle oldukça çekici bir yöntem olarak görülmektedir (Aldenderfer ve Blashfield, 1984 ; s.64). T toplam kareler toplamı ve çapraz çarpım matrisini, B gruplar arası kareler toplamı ve çapraz çarpım matrisini W ise gruplar içi kareler toplamı ve çapraz çarpım matrisini göstermek üzere varyans analizinin bir denklemi olan T = B + W ilişkisine dayalı olarak kümeleme sonuçlarının kalitesine belirlemek amacıyla bazı yöntemler geliştirilmiştir.

Milligan ve Majahan (1980) kümeleme sonuçlarının kalitesini test eden yöntemleri özetleyerek bu yöntemler arasında Arnold (1979) tarafından teklif edilen yöntemin diğer yöntemlere göre daha iyi sonuç verdiğini vurgulamıştır. Arnold, ilk olarak Friedman ve Rubin (1967) tarafından önerilen istatistiği, kümeleme çözümü-

nün istatistik önem testi olarak kullanılabileceğini önermiştir. Söz konusu test istatistiği

$$C = \log (\max \frac{|T|}{|W|})$$

şeklinde verilmektedir (Punj ve Stewart, 1983; s.145).

Gruplar içi kareler toplamı ve çapraz çarpım matrisi (W)'nin determinantının minimasyonu |T|/|W| oranının maksimizasyonu anlamına gelmektedir. Tek değişkenli durum için (p=1) C istatistiği maksimum değerine ulaşmaktadır (Lee, 1979; s.709). Bir dizi ardışık bölümleme yöntemi ile |T|/|W| oranının maksimizasyonu MICKA, CLUSTAN ve CLUS gibi istatistik paket programlarıyla gerçekleştirilebilmektedir (Punj ve Stewart, 1983; s.140). Bu test kriterinde amaç, elde edilen kümelerin optimal şekilde bölünüp bölünmediğinin araştırılmasıdır. |T|/|W| oranını maksimize edilmesi gruplar içinde aynı türden gözlemlerin toplanması, gruplar arasında ise heterojenliğin arttırılması anlamına gelmektedir.

Öklit uzaklığı veya iz(W)'yi minimize etmek suretiyle optimum bölümlemenin araştırıldığı yöntemler, ham veriler için farklı, standartlaştırılmış veriler için farklı sonuçlar verebilirken |T|/|W| oranının maksimizasyonu verilerin standartlaştırılması işleminden etkilenmemektedir. Ancak, kümelerin aynı yapı ve boyuta sahip olması varsayımı bu yöntemin zayıf tarafını teşkil etmektedir (Everit, 1974; s.76-77).

Kümeleme sonuçlarının kalitesini test etmenin diğer bir yolu da, çok değişkenli varyans analizi MANOVA'nın uzantısı niteliğinde olan çok değişkenli bir istatistik analiz tekniği olan Diskriminant Analizi'nin kümeleme sonuçlarına uygulanmasıdır. Bilindiği gibi iki veya daha fazla grubun tek bir değişken açısından karşılaştırılmasında tek değişkenli varyans analizi (ANOVA) yöntemi, iki veya daha fazla değişken açısından karşılaştırılmasında ise MANOVA yöntemi kullanılmaktadır. Verilere çok değişkenli varyans analizi uygulanması sonucunda, gruplar arası fark yoktur (grup ortalama vektörleri birbirine eşittir) anlamını taşıyan H_0 hipotezi reddedildikten sonra gruplar arası farkın olduğu sonucuna varılır.

Diskriminant analizinin iki ana amacı vardır. Bunlardan birincisi, tahmin değişkenlerinin doğrusal bileşenlerinden oluşan diskriminant fonksiyonları aracılığıyla gruplar arası farklılığa etki eden tahmin değişkenlerinin hangileri olduğunu ortaya çıkarmaktır. İkincisi ise, p tane değişken üzerinde ölçümü yapılmış N tane gözlemin k tane kümeden meydana geldiğinin bilinmesi durumunda, söz konusu gözlemlerin diskriminant fonksiyonları aracılığıyla analiz öncesi belirlenmiş olan k tane kümeye en az hata ile atanmasıdır.

Kümeleme analizi sonuçlarının ne ölçüde geçerli olduğu ikinci amaca yönelik, karar amaçlı analiz olarak da adlandırılan sınıflandırma yöntemlerinden birisi ile araştırılabilir. Sınıflandırma, daha önce belirlenen gruplardan birisine ait olduğu bilinen ancak hangi hangi gruba ait olduğu kesin olarak bilinmeyen bir bireyin en az hata ile ait olduğu guruba ait olması olarak tanımlanmakta olup iyi bir sınıflandırma, yanlış sınıflandırmanın kötü etkilerini minimize eder (Eisenbeis and Avery, 1972; s.12).

Diskriminant fonksiyonları X_1, X_2, ... , X_p tahmin değişkenlerinin doğrusal bileşenleri olarak

$$Y = v_1X_1 + v_2X_2 + ... + v_pX_p$$

şeklinde ifade edilir. Fisher tarafında geliştirilen bu fonksiyonda v_i katsayıları, gruplar arası varyansın guruplar içi varyansa oranını maksimum yapacak şekilde hesaplanırlar. Bunun için diskriminant kriteri olarak adlandırılan

$$\lambda = \frac{v'Bv}{v'Wv}$$

fonksiyonunun v'ye göre kısmi türevi alınıp sıfıra eşitlenir ve gerekli işlemler yapılırsa $(B-\lambda W)v = 0$ veya $(W^{-1}B-\lambda I)v = 0$ homojen denklem sistemine ulaşılır (Tatsuoka, 1971; s.160). Bu denklem sisteminin sıfırdan farklı çözümüne ulaşabilmek için λ'nın $|W^{-1}B-\lambda I| = 0$ eşitliğini sağlaması gerekir. Bu determinant denkleminin çözümü öz değer olarak adlandırılan λ'nın köklerini ve bulunan her λ değeri için $(W^{-1}B-\lambda I)v = 0$ denkleminin çözümü aranan v vektörlerini verir. Bu durumda i.ci özdeğer için bulunan vektör v_i ile gösterilirse i.ci diskriminant fonksiyonu

$$Y_i = v_i'X$$

olarak elde edilmiş olur.

Elde edilen diskriminant fonksiyonları yardımıyla kümeleme analizi sonucunda gruplara sınıflandırılan bireyleri grup üyeliği olasılıklarının belirlenmesi yoluyla tekrar sınıflandırmak ve böylece kümeleme sonuçlarını test etmek mümkündür. Grup üyeliği olasılıkları Bayes kuralına göre

$$P(G_g|X_i) = \frac{P(G_g)P(X_i \mid G_g)}{\sum_{j=1}^{k} P(G_j)P(X_i \mid G_j)} \qquad g = 1,2,.....,k$$

formülü ile tahmin edilir.

Burada:

$P(G_g)$: herhangi bir bireyin g. gruptan olması olasılığı olup "önceki olasılıklar" olarak tanımlanır.

$P(X_i|G_g)$: g. gruba ait olduğu bilinen bir i.ci bireyin X_i değer bileşimine sahip olma olasılığıdır. Koşullu olasılık yada uzaklık olasılıkları adı verilen olasılıklar Mahalonobis uzaklıklarıyla ilişkili olarak

$$P(X_i|G_g) = (2\pi)^{-p/2} \exp\{-\frac{1}{2}\chi^2_{ig}\}\, dX_1, dX_2, ..., dX_p$$

şeklinde tanımlanır (Tatsuoka, 1971; s.228). Bu ifade yukarıdaki formülde yerine konup gerekli kısaltmalar yapılırsa; "sonraki olasılıklar" olarak tanımlanan ve X_i gözlem vektörü verilmişken bireyin g. grubun üyesi olma olasılığı,

$$P(X_i|G_g) = \frac{P(G_g)\exp(-\frac{1}{2}\chi^2_{ig})}{\sum_{j=1}^{k} P(G_j)\exp(-\frac{1}{2}\chi^2_{ij})} \quad g = 1, 2, .., k \quad j = 1, 2,..,N$$

şeklinde tanımlanır (1) ve birey en büyük $P(G_g|X_i)$ değerinin elde edildiği gruba sınıflandırılır.

Analize dahil olan tüm bireylerin grup üyeliği olasılıklarının belirlenmesi yoluyla sınıflandırılması yapıldıktan sonra ,sınıflandırma sonuçları "sınıflandırma matrisi" adı verilen bir matris ile özetlenir. Matrisin ana köşegeni üzerindeki değerler doğru sınıflandırılan bireylerin sayısını göstermektedir. Doğru olarak sınıflandırılan bireylerin toplamının toplam birey sayısına oranı doğru sınıflandırma yüzdesini verir. Elde edilen sınıflandırmanın başarılı olup olmadığına tau istatistiğinin hesaplanması ile karar verilebilir. Söz konusu istataistik,

$$\text{tau} = \frac{n_d - \sum_{i=1}^{k} q_i n_i}{N - \sum_{i=1}^{k} q_i n_i}$$

şeklinde tanımlanır (Klecka, 1980; s.51). Burada;
n_d = doğru sınıflandırılan birey sayısını,
q_i = i. gruba ilişkin önceki olasılığı,
n_i = i. gruptaki birey sayısını,
N = gruplardaki toplam birey sayısını,

$\sum_{i=1}^{k} q_i n_i$ = tesadüfi sınıflandırma ile beklenen doğru sınıflandırma sayısını göstermektedir.

Eğer, 0<tau<1 ise sınıflandırma yöntemi tesadüfi sınıflandırmadan beklenen sonuca göre başarılı olduğuna karar verilir ve tesadüfi sınıflandırmaya göre yüzde kaç daha az hata yapıldığı şeklinde yorumlanır.

3. Çapraz Geçerlilik Yöntemleri

Diğer çok değişkenli istatistik analizlerde olduğu gibi kümeleme çözümlerinin güvenilirliği ve geçerliliğinin istatistiksel anlamlılığı gibi açıklanması gerekir. Kümeleme sonuçlarının geçerliliği konusunda çeşitli çapraz geçerlilik yöntemleri önerilmiştir. Oldukça sık kullanılan yöntemlerden birisi, örneğin iki kısma ayrılarak her iki kısımada aynı kümeleme yöntemini uygulamaktır. Daha sonra her iki kısmın tanımlayıcı istatistiklerinin (discriptive statistics) karşılaştırılması yoluyla söz konu-

su kümelerin benzerliği araştırılabilir. Ancak bu yaklaşımla güvenilirliğin objektif bir ölçüsü elde edilememektedir (Punj and Stewart, 1983 ; s.145).

Çapraz geçerlilik için diskriminant analizi de kullanılabilmektedir. Bu yaklaşımda da analiz edilecek örnek iki kısma ayrılarak birinci kısımdan kümeleme çözümü elde edildikten sonra diskriminant fonksiyonları elde edilmekte ve ikinci kısma uygulanmaktadır. Diskriminant fonksiyonlarıyla ikinci kısım için yapılan sınıflandırma ile kümeleme analizi ile yapılan sınıflandırma arasındaki uyumun derecesi kümeleme çözümünün sağlamlığı konusunda bir ölçü olacaktır. Ancak diskriminant fonksiyonunun katsayılarının grupları ayırmada etkisiz olmaları, başka bir ifade ile istatistiksel olarak anlamlı olmamaları ve mevcut örnek büyüklüğünün geçerlilik için yetersiz olabilmesi bu yöntemin zayıf tarafını teşkil etmektedir.

McIntyre ve Blashfield (1980) regresyon analizinde kullanılan çapraz geçerlilik mantığına dayalı olarak, çapraz geçerlilik yöntemine farklı bir yaklaşımda bulunmuşlardır. Oldukça basit ve bilgisayarda çalıştırılması kolay olan bu yöntemin adımlarını aşağıdaki gibi sıralamak mümkündür (Milligan and Cooper, 1987 ; s.347);

Yeterince büyük bir veri seti tesadüfi olarak ikiye bölünüp kümeleme amaçlı iki örnek elde edilir.

Birinci örneğe kümeleme analizi uygulanarak, kümelerin merkezleri hesaplanır. (Bu adımda küme sayısının belirlendiği farz olunmaktadır.)

İkinci örnekteki her birey (gözlem) ile birinci örnekten elde edilen küme merkezleri arasındaki öklit uzaklıkları hesaplanır.

İkinci örnekteki her birey, birinci örnekteki küme merkezlerinden hangisine daha yakın ise kümeye atanır. Bu atama işlemi ikinci örneğin, birinci örneğin özelliklerine dayalı olarak kümelenmesini ortaya çıkarmaktadır.

İkinci örnek kendi verileriyle doğrudan kümeleme analizine tabi tutularak kümeleme yapılır. Böylece ikinci örnekten karşılaştırma amaçlı iki kümeleme elde edilmiş olur.

Aynı veriler için 4. ve 5. adımlardaki atamaların uyuşma derecesi, kümeleme çözümünün tutarlılığının bir göstergesi olacaktır.

McIntyre ve Blashfield söz konusu tutarlılığın ölçüsü olarak kappa istatistiğinin kullanılabileceğini tespit etmişler ve kabul edilebilir bir tutarlılık düzeyine ulaşılması halinde son çözüm için veri setlerinin birleştirilebileceğini vurgulamışlardır.

4. Monte Carlo Yöntemi

Kümeleme sonuçlarının geçerliliğini test etmenin diğer bir yolu da Monte Carlo yöntemi olup yapılacak işlemleri üç adımda özetlemek mümkündür (Aldenderfer and Blashfield, 1984; s.66-74):

Tesadüfi sayı üretecini kullanarak yapay bir veri seti oluşturulur. Bu veri setinde kümeler yoktur fakat orjinal veri setinin tüm özelliklerine sahiptir.

Aynı kümeleme yöntemi hem orjinal verilere hem yapay verilere uygulanır. Veri setlerine uygulanan kümeleme analizi sonuçlarını karşılaştırabilmek için k-ortalamalar yöntemi tatbik edilir.

Bu aşamada orjinal ve yapay veri setlerinden elde edilen kümeleme sonuçlarına ilişkin istatistikler karşılaştırılır. Bu durumda geçerliliğin ölçüsü olarak varyans analizi (ANOVA) uygulanarak F-oranları hesaplanır. Her iki veri setindeki değişkenler için hesaplanan F-oranlarının yeterince büyük ise bu kümelerin homojenliğinin bir göstergesidir. Ayrıca, değişken sayısı ikiden fazla ise temel bileşenler analizi uygulanarak sistem iki boyutlu hale getirilir ve böylece her iki veri setindeki gözlemlerin serpilme diyagramları iki boyutta çizilerek karşılaştırılmaları mümkün olmaktadır.

4. UYGULAMA

Kümeleme analizi sonuçlarını değerlendirmek üzere araştırma kapsamına alınan 76 il, söz konusu illerin eğitim göstergeleri olarak anılan 6 değişken itibari ile birbirine benzer yapıda olan illerin belirlenmesine çalışılacaktır. İlleri eğitim yapısı benzerlik gösteren homojen alt gruplarda toplamak ve uygun küme sayısını belirleyebilmek için verilere kümeleme analizi yöntemleri uygulanmıştır. Analizde kullanılan veriler Devlet Planlama Teşkilatının "İllerin Sosyo-Ekonomik Gelişmişlik Sıralaması Araştırması (1996)" isimli çalışmasında (Dinçer, Özaslan ve Satılmış, 1996; s.110) elde edilmiş ve verilerin istatistik analizi STATISTICA paket programı aracılığı ile gerçekleştirilmiştir.

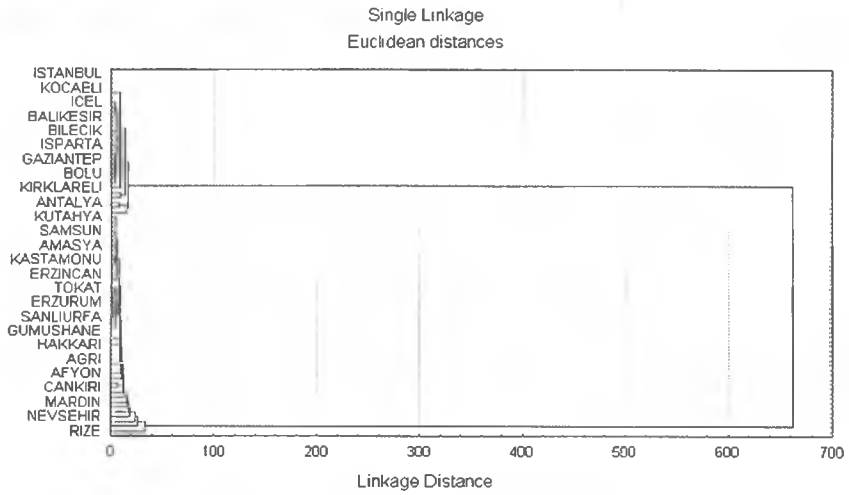
Analizde kullanılan değişkenler ve açıklamaları aşağıdaki gibidir.
VAR1: Okur-Yazar Nüfus Oranı
VAR2: Okur-Yazar Kadın Nüfus Oranı
VAR3: Üniversite Bitirenlerin Okul Bitirenlere Oranı
VAR4: İlkokullar Okullaşma Oranı
VAR5: Ortaokullar Okullaşma Oranı
VAR6: Liseler Okullaşma Oranı

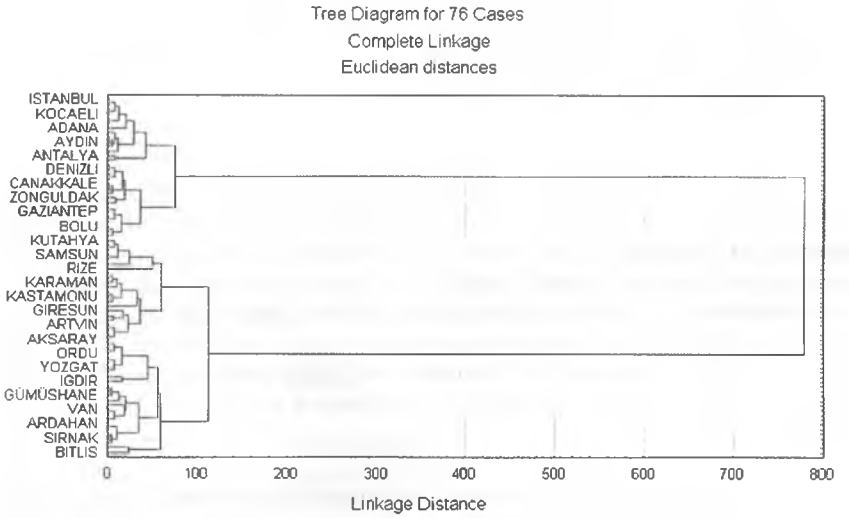
İllerin genel eğitim düzeyini göstermesi açısından önem taşıyan okur-yazar nüfus oranı, o ildeki 6 ve yukarı yaştaki okuma-yazma bilen nüfusun 6 ve yukarı yaş toplam nüfusa; okur-yazar kadın nüfus oranı ise, o ildeki 6 ve yukarı yaştaki okuma-yazma bilen kadın nüfusun 6 ve yukarı yaş toplam kadın nüfusa oranlanması ile elde edilmektedir. Yüksek okul yada fakülte bitirenlerin, okul bitiren nüfusa oranlanmasıyla üniversite bitirenlerin oranı bulunmaktadır. İlkokul, ortaokul ve liselerde okullaşma oranları ise ilgili eğitim kademelerindeki öğrenci sayılarının çağ nüfusuna oranlanması suretiyle elde edilmektedir.

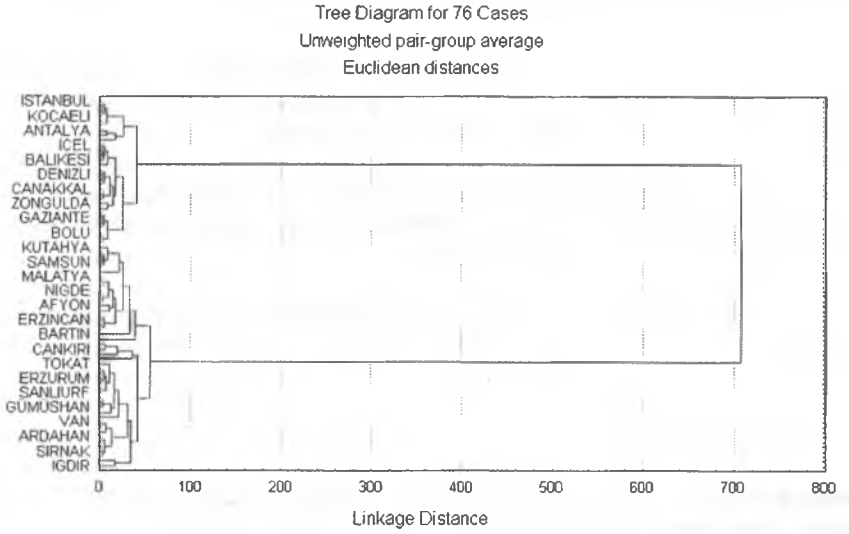
İllerin eğitim yapısı bakımından; kaç kümede toplanabileceği tekli bağlantı, tam bağlantı, ortalama bağlantı ve Ward yöntemleriyle elde edilen dendogramlar

yardımıyla küme sayısına karar verilebilir. Bu amaçla analizde, aşamalı olmayan yöntemlere basamak teşkil etmesi bakımından yukarıda sözü edilen aşamalı yöntemlere ilişkin dendogramlar elde edilmiştir. analizde çeşitli küme sayıları için K-ortalamalar yöntemi uygulanarak eğitim yapısı bakımından birbirine benzeyen iller ayrı ayrı kümelerde toplanmış ve bu yöntemle kümelendirilen iller diskriminant analizi ile tekrar sınıflandırılmış ve doğru sınıflandırma yüzdeleri elde edilmiştir. Analizde elde edilen bulgular aşağıdaki gibi özetlenebilir.

Aşamalı kümeleme yöntemlerinin sonuçları dendogramlarla gösterilmiştir. Söz konusu dendogramlar incelendiğinde genelde iki ana küme gözlenmekte olup bu kümelerin bazılarının da alt kümlere bölündüğü ve uygun küme sayısının iki ile altı arasında tespit edilebileceği gözlenmektedir.







K-ortalamalar yöntemi için küme sayısının analizinden önce verilmesi gerekmektedir. Ancak illerin kaç grupta toplanacağı bilinemediği için, aşamalı yöntemlerden elde edilen bilgilerden de yararlanarak K=2,3,4,5,6 küme sayısı için kümeleme yapılmıştır. Bu kümelere göre illerin eğitim yapısı bakımından oluşturduğu gruplar ve söz konusu kümeleme sonuçlarının değerlendirilmesi için diskriminant analizi yardımıyla yapılan sınıflandırma sonuçları aşağıda sunulmuştur.

Küme Sayısı = 2

1. Küme: { İstanbul, Ankara, İzmir, Kocaeli, Bursa, Eskişehir, Antalya, Tekirdağ, Adana, İçel, Muğla, Aydın, Balıkesir, Kırklareli, Kayseri, Denizli, Bilecik, Edirne, Zonguldak, Çanakkale, Isparta, Manisa, Uşak, Konya, Gaziantep, Hatay, Sakarya, Bolu, Burdur, Kırıkkale}

2. Küme: { Kütahya, Nevşehir, Elazığ, Trabzon, Samsun, Kırşehir, Rize, Malatya, Amasya, Karaman, Afyon, Niğde, Kastamonu, Çorum, Giresun, Artvin, Erzincan, Sivas, Aksaray, K.Maraş, Bartın, Tokat, Çankırı, Sinop, Ordu, Erzurum, Diyarbakır, Yozgat, Ş.Urfa, Tunceli, Adıyaman, Kars, Gümüşhane, Bayburt, Batman, Mardin, Van, Siirt, Iğdır, Hakkari, Bitlis, Ardahan, Bingöl, Ağrı, Şırnak, Muş}

Bu illerin diskriminant analizi ile tekrar yapılan sınıflandırmada %100'lük bir doğru sınıflandırma elde edilmiştir. Ayrıca, Wilks Lambda olarak adlandırılan ve gruplar arası farklılıkların bir ölçüsü olan ve 0 ile 1 arasında değer alan istatistik, Λ=0.0002 olarak hesaplanmıştır. Bunun anlamı ise analizde kullanılan ayırıcı değişkenlerin gruplar arasındaki farklılıkları açıklamada etkili olduklarıdır.

Küme Sayısı = 3

1. Küme: { İstanbul, Ankara, İzmir, Kocaeli, Bursa, Eskişehir, Antalya, Tekirdağ, Adana, İçel, Muğla, Aydın, Balıkesir, Kırklareli, Kayseri, Denizli, Bilecik,

Edirne, Zonguldak, Çanakkale, Isparta, Manisa, Uşak, Konya, Gaziantep, Hatay, Sakarya, Bolu, Burdur, Kırıkkale}

2. Küme: { Kütahya, Nevşehir, Elazığ, Trabzon, Samsun, Kırşehir, Rize, Malatya, Amasya, Karaman, Afyon, Niğde, Kastamonu, Çorum, Giresun, Artvin, Erzincan, Sivas, Aksaray, K.Maraş, Bartın, Tokat, Çankırı}

3. Küme: {Sinop, Ordu, Erzurum, Diyarbakır, Yozgat, Ş.Urfa, Tunceli, Adıyaman, Kars, Gümüşhane, Bayburt, Batman, Mardin, Van, Siirt, Iğdır, Hakkari, Bitlis, Ardahan, Bingöl, Ağrı, Şırnak, Muş}

Diskriminant analizi ile yapılan tekrar sınıflandırılmada %97.37 doğru sınıflandırma elde edilmiştir. Daha Önce ikinci kümede yer alan Tokat ve Çankırı illeri analiz sonunda üçüncü kümede yer almışlardır.

Küme Sayısı = 4

1. Küme: {İstanbul, Ankara, İzmir, Kocaeli, Bursa, Eskişehir, Antalya, Tekirdağ, Adana, İçel, Muğla, Aydın, Balıkesir, Kırklareli, Kayseri, Denizli, Bilecik, Edirne, Zonguldak, Çanakkale, Isparta, Manisa, Uşak, Konya, Gaziantep, Hatay, Sakarya, Bolu, Burdur, Kırıkkale}

2. Küme: {Kütahya, Nevşehir, Elazığ, Trabzon, Samsun, Kırşehir, Rize, Malatya, Amasya, Karaman, Afyon, Niğde, Kastamonu, Çorum, Giresun}

3. Küme: {Artvin, Erzincan, Sivas, Aksaray, K.Maraş, Bartın, Tokat, Çankırı, Sinop, Ordu, Erzurum, Diyarbakır, Yozgat, Ş.Urfa, Tunceli}

4. Küme: {Adıyaman, Kars, Gümüşhane, Bayburt, Batman, Mardin, Van, Siirt, Iğdır, Hakkari, Bitlis, Ardahan, Bingöl, Ağrı, Şırnak, Muş}

Bu illerin diskriminant analizi ile tekrar sınıflandırılması sonucunda 1. ve 2. kümelerdeki illerin doğru sınıflandırma oranı %100 iken 3. kümede %73.3 ve 4. kümede %87.5 doğru sınıflandırma elde edilmiştir. Toplam doğru sınıflandırma oranı ise %92.1 olarak gerçekleşmiştir. Analiz öncesi üçüncü kümede yer alan Artvin, Erzincan, Sivas ve Aksaray illeri analiz sonunda ikinci kümede yer almışlardır. Ayrıca analiz öncesi 4. kümede yer alan Adıyaman ve Kars illeri analiz sonrası 3. kümede yer almışlardır.

Küme Sayısı = 5

1. Küme: { İstanbul, Ankara, İzmir, Kocaeli, Bursa, Eskişehir, Antalya, Tekirdağ, Adana, İçel, Muğla, Aydın, Balıkesir, Kırklareli, Kayseri, Denizli, Bilecik, Edirne, Zonguldak, Çanakkale, Isparta, Manisa, Uşak, Konya, Gaziantep, Hatay, Sakarya, Bolu, Burdur, Kırıkkale}

2. Küme: { Kütahya, Nevşehir, Elazığ, Trabzon, Samsun, Kırşehir, Malatya, Amasya, Karaman, Afyon}

3. Küme: {Rize, Niğde, Kastamonu, Çorum, Giresun, Artvin, Erzincan, Sivas, Aksaray, K. Maraş, Bartın, Tokat, Çankırı}

4. Küme: { Sinop, Ordu, Erzurum, Diyarbakır, Yozgat, Ş.Urfa, Tunceli, Adıyaman, Kars, Gümüşhane, Bayburt, Batman, Mardin, Iğdır, Bitlis}

5. Küme: { Van, Siirt, Hakkari, Ardahan, Bingöl, Ağrı, Şırnak, Muş}

Bu illerin tekrar sınıflandırılması sonucunda 1. , 4. ve 5. kümelerde yer alan illerin doğru sınıflandırma oranı %100 iken 2. kümede %80 ve 3. kümede %76.9 doğru sınıflandırma oranı elde edilmiştir. Analiz öncesi ikinci kümede yer alan Karaman ve Afyon illeri analiz sonunda üçüncü kümede yer almışlardır. Analiz öncesi 3. kümede yer alan Çankırı ve Tokat illeri 4. kümede, Bartın ili ise 2. kümede yer almışlardır. Küme sayısının 5 olduğu bu sınıflandırmadaki toplam doğru sınıflandırma oranı %93.4 olarak gerçekleşmiştir.

Küme Sayısı = 6

1. Küme: { İstanbul, Ankara, İzmir, Kocaeli, Bursa, Eskişehir, Antalya, Tekirdağ, Adana, İçel, Muğla, Aydın, Balıkesir, Kırklareli, Denizli}

2. Küme: { Kayseri, Bilecik, Edirne, Zonguldak, Çanakkale, Isparta, Manisa, Uşak, Konya, Gaziantep, Hatay, Sakarya, Bolu, Burdur, Kırıkkale}

3. Küme: { Kütahya, Nevşehir, Elazığ, Trabzon, Samsun, Kırşehir, Malatya, Amasya, Karaman, Afyon}

4. Küme: {Rize, Niğde, Kastamonu, Çorum, Giresun, Artvin, Erzincan, Sivas, Aksaray, K.Maraş, Bartın, Tokat, Çankırı}

5. Küme: { Sinop, Ordu, Erzurum, Diyarbakır, Yozgat, Ş.Urfa, Tunceli, Adıyaman, Kars, Gümüşhane, Bayburt, Batman, Mardin, Iğdır, Bitlis}

6. Küme: { Van, Siirt, Hakkari, Ardahan, Bingöl, Ağrı, Şırnak, Muş}

Küme sayısının 6 olarak alınması durumunda daha önce 1. kümede yer alan 30 il iki gruba ayrılmıştır. Diğer 4 kümenin elemanları ise hiç değişmemiştir. Diskriminant analizi ile yapılan tekrar sınıflandırmada ise 4. küme hariç tüm kümelerde %100'lük bir doğru sınıflandırma elde edilirken toplam %98.7'lik bir doğru sınıflandırma oranı elde edilmiştir. Analiz öncesi 4. kümede yer alan Tokat ili analiz sonrası 5. kümede yer almıştır.

Ayrıca diskriminant analizi ile yapılan sınıflandırılmanın tesadüfi sınıflandırmaya göre başarılı olup olmadığını belirlemek üzere hesaplanan *tau* istatistiği 0.984 olarak hesaplanmıştır. söz konusu istatistik 1'e yakın olduğu için sınıflandırma yöntemi başarılı olmuştur. Başka bir ifade ile tesadüfi sınıflandırmaya göre %98.4 daha az hata yapılmıştır.

5. SONUÇ

Çok değişkenli istatistik analiz tekniklerinden kümeleme analizinin ve kümeleme sonuçlarının kalitesini değerlendirme yöntemlerinin tanıtıldığı bu çalışmada ulaşılan sonuçları aşağıdaki gibi özetlemek mümkündür.

Yapılan çalışmada muhtemel küme sayısını belirlemek amacıyla aşamalı yöntemlerin dendogram sonuçlarından yararlanılmış olup bu yöntemlerle kümeleme işlemi yapılmamıştır. Küme sayısının belirlenmesi konusunda önerilen yöntemlerin farklı sonuçlar vermesi dolayısıyla, uygun küme yapısına karar verirken uygulayıcının konu hakkındaki ön bilgisi ve tecrübesi kümeleme sonuçlarının tutarlılığı bakımından büyük önem taşımaktadır.

Analizde kullanılan değişken sayısının çok fazla olmaması nedeniyle değişken sayısının azaltımına gidilmediği bu çalışmada aşamalı olmayan kümeleme tekniklerinden K-ortalamalar yöntemi ile kümeleme yapılmıştır. Bu yöntemde küme sayısının önceden verilmesi gerektiğinden yöntem farklı küme sayıları için (K=2,3,4,5,6) tekrarlanmıştır. Her tekrar sonucunda elde edilen kümeleme sonuçlarının illerin sosyo-ekonomik gelişmişlik sıraları (Dinçer ve diğerleri, 1996; s.51) ile karşılaştırıldığında kümelerin yorumlanabilir nitelikte oldukları görülmüştür. Diğer bir ifade ile kümeler illerin sosyo-ekonomik gelişmişlik sıralarına göre oluşmuştur. araştırma kapsamında bulunmayan Yalova, Karabük, Kilis, Osmaniye illeri daha önce bağlı bulundukları iller çerçevesinde değerlendirilmiştir.

K-ortalamalar yöntemine göre elde edilen kümeleme sonuçlarının geçerliliği diskriminant analizi ile test edilmiştir. K-ortalamalar yöntemine göre elde edilen kümeler diskriminant analizi için veri olarak kullanılmış ve bu verilerden elde edilen diskriminant fonksiyonları ile illerin grup üyeliği olasılıkları (sonraki olasılıklar)'nın belirlenmesi yoluyla yapılan sınıflandırmada ön olasılıklarda dikkate alınmıştır. Ayrıca, Wilks Lambda istatistiklerinin sıfıra yakın olması (sırasıyla 0.00023 , 0.00009, 0.00006, 0.00021 ve 0.000016) değişkenlerin gruplar arası farklılıkları açıklamada etkili oldukları, başka bir ifade ile grup merkezlerinin birbirlerinden kesinlikle ayrılmış olduğunun bir göstergesi olarak ortaya çıkmıştır. K-ortalamalar yöntemi ile elde edilen kümeleme sonuçlarının test edilmesi sonucunda illerin eğitim yapısı bakımından iki kümeye bölündüğü durumda %100, altı kümeye bölündüğü durumda %98.7 ve üç kümeye bölündüğü durumda ise %97.4 doğru sınıflandırma elde edilmiştir. Ancak, küme sayısının dört ve beş olması durumlarında da doğru sınıflandırma oranı %92'den daha az değildir. Bu nedenle K-ortalamalar yöntemi ile elde edilen kümeleme sonuçlarının tutarlı olduğunu söylemek mümkündür.

YARARLANILAN KAYNAKLAR

Aldenderfer, M. S. and R. K. Blashfield. (1984). **Cluster Analysis**, Beverly Hills: Sage Publications.

Anderberg, M. R. (1973). **Cluster Analysis for Researchers** New York: Academic Press.

Anderson, T. W. (1951). "**Classification by Multivariate Analysis** ", Psychometrika, 16.

Arnold, S. J. (1979). "**A Test for Clusters**", Journal of Marketing Research, 16.

Baker, F. B., and L. J. Hubert (1975) "**Measuring the Power of Hierarchical Cluster Analysis**" Jasa, 69.

Dıncer, B., M. Özaslan ve E. Satılmış (1996). **İllerin Sosyo-Ekonomik Gelişmişlik Sıralaması Araştırması** . Yayın No: DPT: 2466, Ankara.

Duran, B. S. and P. L. Odel (1974). **Cluster Analysis** (Lecture Notes in Economics and Mathematical Systems, Econometrics; Managing Editors: M. Beckmann and H. P. Kunzı). Springer Verlag: NewYork.

Everitt, B. (1974). **Cluster Analysis,** London: Heinemann Educational Books Ltd.

Everitt, B. (1993) **Cluster Analysis** for Applications Academic Press, New York.

Friedman, H. D., amd J. Rubin (1967) "**On Some Invariant Criteria for Grouping Data**" Jasa, 62.

Green, P. E. And D. S. Tull (1973). **Research for Marketing Decisions** (2nd Ed.) Prentice Hall, New Delhi.

Hartigan, John A. (1975), **Clustering Algorithms**, New York: John Willey & Sons.

Jaın, A. K. And R. C. Dubes, (1988) **Algorithms for Clustering Data Englewood Cliffs,** NJ: Prentice Hall.

Johnson, R. A. and D. W. Wichern (1988). **Applied Multivariate Statistical Analysis**, (2nd Ed.). Prentice Hall, Englewood Cliffs, New Jersey.

Lee, Kerry L. (1979), "**Multivariate Test for Clusters**", JASA, 74.

Mardia, K. W., J. T. Kent and J. M. Biby (1979). **Multivariate Analysis. Academic Press,** New York.

Mıllıgan, G. W., and V. Majahan (1980). "**A Note on Proceduresfor Testing the Quality of a Clustering a Set of Objects**", Decision Sciences, 11.

Milligan, G.W. and M. C. Cooper (1987), "**Methodology Review: Clustering Methods**". Applied Psychological Measurement, Vol. II No:4

Morrison, D. G. (1967). "**Measurement Problems in Cluster Analysis**" Management Science, 13.

Punj, G., and D. W. Stewart (1983). "**Cluster Analysis in Marketing Research: Review and Suggestions for Application**". Journal of Marketing ResearchVol. XX.

Tatlıdil, H. (1992), **Uygulamalı Çok Değişkenli İstatistiksel Analiz,** H.Ü. Fen Fakültesi İstatistik Bölümü, Ankara

Tatsuoka, M. M. (1971). **Multivariate Analysis:** Techniques for Educational and Psychological Research John Wiley and Sons, Inc., New York.

Yeomans, K. A. (1979). **Multivariate Classification: Data Reduction Using Component and Cluster Analysis,** Brimingham Wards, The University of Aton Management Centre, Working Paper Series No:145 (July 1979)