KÜMELEME ÇÖZÜMLEMESİNDE UYGUN KÜMELEME ÖLÇÜTLERİNİN KARŞILAŞTIRILMASI

K. Setenay DİNÇER

Anadolu Üniversitesi
Sağlık Bilimleri Enstitüsü
Lisansüstü Öğretim Yönetmeliği Uyarınca
Tıbbi Biyoloji Anabilim Dalı
Biyoistatistik Bilim Dalında
DOKTORA TEZİ
Olarak Hazırlanmıştır.

Danışman : Prof. Dr. Kazım ÖZDAMAR

Haziran 1992

(97)

KABUL VE ONAY SAYFASI

K. Setenay DİNÇER'in DOKTORA tezi olarak hazırladığı "Kümeleme Çözümlemesinde Uygun Kümeleme Ölçütlerinin Karşılaştırılması" başlıklı bu çalışma, jürimizce Lisansüstü Öğretim Yönetmeliği'nin ilgili maddeleri uyarınca değerlendirilerek kabul edilmiştir.

05/06/1992

ÜYE: Prof. Dr. İsmail KAN (imza)

ÜYE: Prof. Dr. Ayşe BAŞARAN (imza)

ÜYE: Prof. Dr. Kazım ÖZDAMAR (imza)

Anadolu Üniversitesi Sağlık Bilimleri Enstitüsü Yönetim Kurulu'nun 08.06.1992 gün ve 185/463 sayılı kararıyla onaylanmıştır.

ASLI GİBİDİR
0 [illegible] 1992
İsmet YILMAZ
Enstitü Sekreteri

(imza)
Prof. Dr. Nurettin BAŞARAN
Enstitü Müdürü

ÖZET

Bu araştırmada, Kümeleme ölçütleri, g_{1a}, g_{1b}, g_2, g_{3a}, g_{3b}, Wilk's lamda, Hotelling-Lawley iz istatistiği ve Kofenetik korelasyon katsayısı (r_{cs}), çokdeğişkenli normal dağılımdan rasgele çekilmiş R10, R20, R30, R40 ve R50 grupları ile koşullu olarak çekilmiş K20, K30, K40, K50 ve K60 gruplarında test edildi. Bu gruplar 6 değişken ve farklı sayıda birim içeriyorlardı. Altı değişkene göre değerlerin Öklid uzaklıkları kullanılarak benzerlik matrisleri bulundu ve küme sayısı 2 ile 5 arasında olacak şekilde aşamalı olmayan kümeleme yöntemi K-Ortalamalar yöntemi ile kümelendi ve Aşamalı kümeleme yöntemi olan Tek Bağlantı Kümeleme yöntemi ile bağlantılar belirlendi. Rasgele ve koşullu olarak belirlenen gruplarda küme istatistikleri hesaplandı ve ağaç grafikleri helde edildi.

Oluşan kümelerin kümeleme kriterleri, g_{1a}, g_{1b}, g_2, g_{3a}, g_{3b}, r_{cs}, her küme sayısına göre hesaplandı. Çokdeğişkenli varyans çözümlemesi yapıldı ve kümenin türdeş bir kümemi değilmi olduğunu belirlemeye yarayan Wilk's lamda ve Hotelling-Lawley iz istatistikleri, F istatistikleri ve olasılıkları hesaplandı. Öteyandan 6 değişkenli her küme için doğru sınıflandırma olasılıkları Ayırma çözümlemesi ile hesaplandı.

Kümeleme ölçütlerinin hesaplanmasında SPSS/PC+, SYSTAT, BMDP and MINITAB veri çözümlemesi paket programlarından yararlanıldı.

g_{1a} g_{1b}, g_2, g_{3a}, g_{3b}, ölçütleri rasgele grupların kümelenmesinde daha fazla küme sayısının uygun olduğunu gösterirken, Wilk's lamda ve Hotelling-Lawley iz

istatistikleri uygun küme sayısını 3 olarak belirtmiştir.

Gruplarda bir ya da daha fazla bağımsız birimin kümelenmesi arttığı zaman Tek bağlantı kümeleme yöntemi ve rcs ölçütü kümelenmenin uygun olmadığını belirtmektedir. Karşıt olarak gruplarda 3 ya da daha fazla birim içeren küme sayısı varsa, TBK ve r_{cs} Wilk's lamda ve Hotelling-Lawley iz istatistikleri ile aynı küme sayısını göstermiştir.

İki farklı parametreli çokdeğişkenli normal dağılımdan türetilmiş olan koşullu gruplarda, g_{1a}, g_{3a} ve Tek bağlantı kümelemesi yöntemleri üç kümeyi uygun kümelenme olarak verirken, Wilk's lamda ve Hotelling-Lawley iz istatistikleri iki kümeyi önemli olarak vermiştir. Bu son iki istatistik, iki farklı dağılımdan çekilmiş birimleri, enuygun kümelenme sayısı olarak 2 kümeye ayırmıştır. Bu birimler ayırma fonksiyonlarına göre doğru sınıflandırma olarak iki sınıfta gruplandırılmıştır.

Koşullu gruplarda enuygun kümelenme ölçütü Wilk's lamda ve Hotelling-Lawley iz istatistikleri idi. Diğer kümelenme ölçütleri, g1a, g1b, g3a, g3b aynı özelliklere sahip verilerde farklı sonuçlar oluşturma eğilimleri göstermiştir.

İki Öklid uzaklığı ele alındığında, Wilk's lamda, Hotelling-Lawley iz istatistikleri ve Tek bağlantı kümelemesi aynı küme sayısını vermiştir.

Rasgele ve koşullu gruplarda, kümeleme ölçütleri gruplarda benzerlik ve uyşma göstermemiştir. Rasgele grupların birim sayıları artarken, Wilk's lamda ve Hotelling-Lawley iz istatistikleri dişındaki kümeleme ölçütleri uygunluk göstermeyen bir biçimde küme sayıslarında düzensin artış ve azalma göstermiştir.

Hem rasgele hemde koşullu gruplarda, g2 ölçütü koşullu kümelenmeye karşı önemli kümelenme , kümelenme duyarlılığı ve etkinliği göstermemiştir.

Biz, Wilk's lamda ve Hotelling iz istatistiğinin en etkin kümelenme ölçütü olduğunu belirledik. Tıp, Biyoloji ve diğer bilim dallarında, verilerden ayrıntılı bilgi edinmek için karmaşık veri matrisleri Kümeleme Çözümlemesi ile analiz edilmelidir.

Veri matrisleri K-Ortalamalar yöntemiyle 2 ve daha fazla küme sayısına göre, birim ve değişkenlerin aşamalı kümelenmesine ilişkin bilgi edinmek amacıyla da Tek Bağlantı Kümelemesi ile çözümlenmelidir.

SUMMARY

In this study, sensitivity and effectiveness of clustering criteria, g_{1a}, g_{1b}, g_2, g_{3a}, g_{3b}, Wilk's Lamda statistic, Hotelling-Lawley Trace statistic and Cofenetic correlation coefficient (r_{cs}) were tested in the randomly drawing groups R10, R20, R30, R40, R50 and conditionally drawing groups K20, K30, K40, K50 and K60 from multivariate normal distrubitions. Those groups were involved different number units and six variables. Using Euclidean distances of values obtained by six variables, similarity matrices were obtained and clustered with K-Means method which is a nonhierarchical clustering method according two and five cluster numbers and Units linked with Single Linkage Clustering method which is a hierarchical clustering method. Cluster statistics were calculated and Dendrograms were prepared according to randomly and conditionally selected groups.

Clustering criteria of obtaining clusters which were g_{1a}, g_{1b}, g_2, g_{3a}, g_{3b}, r_{cs} were determined and for each cluster number. Multivariate analysis of variance were performed and Wilk's Lamda and Hotelling-Lawley Trace Statistics that were the statistics obtaing whether is a homogenous cluster or heteregenous cluster, its F statistics and its probabilities calculated accordingly. On the other hand, for each cluster with six variables, correct classification probabilities were calculated with Discriminant analysis

SPSS/PC+, SYSTAT, BMDP and MINITAB data analysis packages were used for clustering criteria calculations.

While the g1a, g1b, g3a and g3b criteria gave more cluster number as suitable of cluster number for clustering the random groups, The Wilk's lamda and Hotelling-Lawley trace criteria gave 3 cluster for suitable cluster number.

Single linkage clustering method and rcs criterion were pointed out that units were not suitable in cluster when the clustering of one or more independent units were increasing in the groups. Alternatively if they were cluster with 3 or more units in those groups, simple linkage method and r_{cs} criterion gave the same cluster number as well as Wilk's lamda and Hotelling-Lawley Trace statistics.

In the conditional groups, which were drawn two multivariate normal distributions different parameters, Wilk's lamda and Hotelling-Lawley trace statistics gave 2 cluster as significant while the g_{1a}, g_{3a} and Single Linkage clustering method gave 3 cluster as suitable clustering. The latest two statistics clustered the units drawing two different distrubitions as most convenient clustering number into two clusters. These units classified into two groups as correct clasification according to discriminant fonctions.

Most convenient clustering criteria were Wilk's lamda and Hotelling-Lawley trece statistics in the conditional groups. The other clustering criteria, g1a, g1b, g3a, g3b indicated different inclinations in order to produce different results in the data having same properties.

When Euclidean linkage distance of 2 is taken Wilk's lamda, Hotelling-Lawley trace statistics and Single Linkage Clustering method gave the same clustering number.

In the random and conditional groups, the clustering criteria did not indicate aggreement and association at results of groups.

While the units number of random groups increased, the clustering criteria exeluding the Wilk's lamda and Hotelling-Lawley trace statistics were observed disorderly the increasing and decreasing on the clustering number without agreement.

In both random and conditional groups, g2 criterion did not indicate significant clustering, sensitivity and effectiveness of clustering against to conditional clustering

We obtained that most effective clustering criteria were the Wilk's lamda and Hotelling-Lawley Trace statistics. In medicine, Biology and other sciences, complex data matrices should be analysed with Cluster Analysis to have detailed information from the data.

The data matrices should be analysed with K-Means method accordingly 2 or more cluster numbers and Single Linkage method for hierarchical clustering information of data units and variables.

TEŞEKKÜR

Bu çalışmanın gerçekleşmesinde yol gösterip, ilgi ve yardımlarını esirgemeyen değerli hocam Prof.Dr.Kazım ÖZDAMAR'a, çalışma arkadaşlarım Arş.Gör. Mevlüt TÜRE'ye, Arş.Gör. Canan BAYDEMİR'e, Arş.Gör. Fezan ŞAHİN'e ve çalışmalarım sırasında gösterdikleri maddi ve manevi destekten dolayı aileme teşekkürü borç bilirim.

İÇİNDEKİLER

Sayfa

İÇİNDEKİLER (Devam)

Sayfa

İÇİNDEKİLER (Devam)

İÇİNDEKİLER (Devam)

ŞEKİLLER DİZİNİ

ŞEKİLLER DİZİNİ (Devam)

ÇİZELGELER DİZİNİ

ÇİZELGELER DİZİNİ (Devam)

ÇİZELGELER DİZİNİ (Devam)

1. GİRİŞ

Günümüzde yapılan araştırmalarda birimlerin değişik özelliklerine ilişkin gözlemler yapılarak elde edilen veri matrislerinden objektif kararlara ulaşılmak istenmektedir. Bilgisayarlı veri analizlerinin yaygınlaşması ile orantılı olarak gözlemler çok sayıda birimi ve bu birimlerin çok sayıda özelliğini kapsamaktadır. Veri matrisinde değişkenler arasında ikili ya da çoklu düzeyde ilişkiler saptanmaktadır. Bu değişkenlerden bazılarının bağımlı, bazılarının bağımsız değişken olarak tanımlanması mümkün iken, veri matrisindeki bazı değişkenlerin diğer değişkenlerle olan ilişkileri hakkında karar vermek mümkün olamamaktadır. İncelenen konu ile ilgili bilgi birikiminin kısıtlı olması bu belirsizlikleri artırmaktadır.

Son yıllarda yapılan araştırmalarda incelenen konudaki gözlemlere dayalı olarak değişkenler arası neden-sonuç ilişkilerinin ortaya konmasında bağımlı, bağımsız değişken tanımlamalarının yapılabildiği ya da yapılamadığı veri yapılarını incelemede çokdeğişkenli çözümleme yöntemlerinden yaygın olarak yararlanılmaktadır (Paykel, 1971; Scott and Symons, 1971; Vogt, et al., 1987; Leach, 1988; Benjamini and Igbaria, 1991).

Veri matrisinde, değişkenler arasındaki birlikte değişim yapıları hakkında kesin ya da kesine yakın doğrulukta bilgi birikimi varsa bu veri matrisini ilişki analizleri ile analiz ederek kararlara ulaşmak olasıdır. Ayrıca değişkenler arasındaki neden-sonuç ilişkilerini çeşitli modeller aracılığı ile ortaya koymak da olasıdır. Fakat veri matrisindeki değişkenlerin bağımlı, bağımsız değişken ayırımı yapılamadığı zaman, tek değişkenli çözümleme yöntemleri ile ya da model kurarak işleme yoluna gitmek verilerin genel yapısı hakkında yeterli bilgi vermemekte ya da eksik bilgi vermektedir. Bu durumda veri yapıları önemli düzeyde karmaşık olmakta ve bu verileri tek değişkenli yöntemlerle incelemek ve bilimsel kararlara ulaşmak güçlük yaratmaktadır. Veri matrisinde birim sayısı ve değişken sayısı arttıkça, değişkenler arası neden-sonuç yapısının çözümlemesi için çokdeğişkenli çözümleme yöntemlerinden yararlanılması karar vermeyi kolaylaştırmaktadır (Armitage, 1983; Kurtuluş, 1985; Özdamar, 1988; Krzanowski and Lai, 1988).

Veri analizlerinin temel amacı; Klinik, laboratuvar ve saha çalışmaları ile bilimsel araştırmalarda, uygulamacı ve bilim adamlarına kullandıkları veriler hakkında daha yararlı bilgi edinebilmelerini sağlamaktadır.

Biyoloji ve tıp alanında yapılan araştırmalardan elde edilen veri matrisleri önemli karmaşıklıklar taşımakta ve Tıp, biyoloji ve diğer bilim dallarında veri matrislerinin çözümünde çokdeğişkenli çözümleme yöntemlerinden yaygın olarak yararlanılma gereği duyulmaktadır (Overall and Klett, 1972; Johnson and Wichern, 1988; Dawson-Saunders and Trapp, 1990).

Bu yöntemlerden en yaygın olarak kullanılanları, Çokdeğişkenli Varyans Çözümlemesi (Multivariate Analysis of Variance, MANOVA), Ayırma Çözümlemesi (Discriminant

Analysis), Faktör Çözümlemesi, Ana Bileşenler Çözümlemesi (Principal Component Analysis, PCA), Çokboyutlu Ölçekleme Çözümlemesi (Multidimensional Scaling), Uyum çözümlemesi (Corresponding Analysis) ve Kümeleme Çözümlemesi (Cluster Analysis) olarak sayılabilir (Tyron and Bailey, 1970; Hawkins, et al., 1982; Chatfield and Collins, 1980; Funkhouser, 1983; Lindsey, et al., 1987; Johnson and Wichern, 1988).

Bu yöntemlerden, Kümeleme Çözümlemesi, Tıp ve Biyoloji alanında hastalıkların, klinik ve laboratuvar bulgularının belirli özelliklere göre sınıflandırılmasını, yeni bir yaklaşımla ele alınmasını, hastalıkların toplumdaki görülme, yayılma ve bulaşkanlık kalıplarının ortaya çıkarılmasını sağlama gibi önemli kullanım alanı olan bir yöntemdir. Belirsizlik koşullarında; yeni hastalık, obje ve birim sınıflaması ölçütlerinin oluşturulmasında, yeni olgu kalıplarının belirlenmesi vb. konularda kümeleme çözümlemesi araştırmacılara yardımcı bir istatistiksel yöntemdir (Hawkins, et al., 1982; Johnson and Wichern, 1988; Özdamar, 1988; Bağcı vd., 1991).

Veri matrisindeki değişkenler, birimler ya da hem değişkenler hem de birimler hakkında ayrıntılı olarak bilgi edinilmek istenirse, bunların benzerlikleri en fazla olanları biraraya toplanarak sınıflandırmaları yapılır. Bu sınıflandırma sonunda aynı küme içindeki birimler ya da değişkenler benzer özellik gösterirken, diğer kümelerdeki değişkenler ya da birimlerle farklı özelliklere sahiptirler. Buradaki en önemli nokta, verilerin temelinde yatan doğal sınıflamayı ortaya çıkarmaktır (Anderberg, 1973; Sözer vd., 1986).

n birim üzerinde yapılan bir araştırmada; bir hastalık tanısındaki belirli semptomlarla ölçülen değerler benzerlik göstermediğinde veri matrisinden bu olgularla ilgili yeni yapıların oluşturulması ve yeni bir hastalık tanımlaması yapılması gerekebilir. Bu bireylerin ve değişkenlerin, küme ya da alt kümelere ayrılarak bireyler arası ya da değişkenler arası benzerlik ya da uzaklıklarından yararlanarak birimlerin uyabileceği yeni hastalıklar hakkında daha ayrıntılı bilgilerin elde edilmesi mümkün olabilir. Burada, birimlerin ya da değişkenlerin benzer özellik taşıyanlarını, benzerlik düzeylerine göre biraraya getirerek kümelere ayırmak amaçlanmaktadır (Armitage, 1983; Johnson and Wichern, 1988).

Kümeleme çözümlemesinin, hem birimleri hem değişkenleri kümelere ayırma ve sınıflandırma olanağı vermesi, bu yöntemin son yıllarda tıpta; psikiyatri, patoloji, mikrobiyoloji ve klinik bilim dallarında klinik tanı ve tedavide, klinik ve laboratuvar bulgularının sınıflandırılmasında yararlanılmasına olanak vermiş ve bu alanlarda alınacak kararların doğruluğunda önemli katkısı olmuştur (Overall and Klett, 1972; Özdamar, 1983; Mornon et al., 1989; Gaboriaud, 1990; Dawson-Saunders and Trapp, 1990).

Kümeleme çözümlemesi, birimlerin ve değişkenlerin ayrılabilecekleri küme sayısını belirleme ve benzerlikleri saptama açısından bir çok yöntem içermektedir. Kümeleme Çözümlemesi en yaygın kullanılan biçimiyle Aşamalı ve Aşamalı Olmayan Kümeleme yöntemleri olarak iki farklı yaklaşımla veri matrisini kümelere ayırmaktır. Yöntem farklılığı birimlerin ve değişkenlerin girebilecekleri kümelerin farklılaşmasına neden olmaktadır. Bu yöntemlerde kendi içinde küme oluşturma yaklaşımları bakımından birçok ayrıma sahiptir.

Bu güne kadar aşamalı yöntemler üzerinde yapılan yöntem karşılaştırmalarında Tek Bağlantı Kümelemesi ve Ortalama Bağlantı Kümelemesi yöntemlerinin diğerlerine üstün olduğu belirlenmiştir. (Anderberg, 1973; Sneath and Sokal, 1973; Everitt, 1979; Özdamar, 1988; Bağcı vd., 1991).

Çokdeğişkenli bir yapıda birimlerin oluşturabilecekleri en uygun küme ya da alt küme sayısını saptamak için birçok ölçüt geliştirilmiştir. Bu ölçütler, kümeiçi (within) ve kümelerarası (between) varyans ya da kovaryans matrislerinin çeşitli biçimlerde ele alınarak hesaplanmasına, çokdeğişkenli varyans çözümlemesinde grupların farklılığını test etmeye yarayan istatistiklere ve kümelere giren birimlerin benzerliklerinin kümeleme öncesi ve sonrası benzerliklerini belirten ölçülerin ilişkilerine dayanmaktadır. Bunlardan ensık kullanılanları g_1, g_2, g_3, Wilk's lamda istatistikleri, Hotelling-Lawlay iz (trace) ve kofenetik korelasyon katsayısı olarak sayılabilir. Bu ölçütlerinde bir veri yapısında yapılan kümelemede farklı sonuçlar verdikleri görülmektedir. Fakat kümeleme ölçütleri üzerinde birbirlerine göre belirgin bir üstünlük olduğu ileri sürülmemiştir. Farklı ölçütler kullanılarak birimlerin ve değişkenlerin kümelenmesi yapılmıştır (Jambu and Lebeaux, 1986; Johnson and Wichern, 1988; Bağcı vd., 1991).

Veri matrisindeki bilgilere dayanarak birimlerin kaç kümeye ayrılabileceği sorusu kümeleme çözümlemesinde önemli bir sorudur. Yine değişkenlerin bir ya da bir kaçının birlikte açıkladığı yapıların ne kadar olduğu henüz net olarak çözümlenmiş değildir. Veri matrisinin ifade ettiği doğal sınıflama hakkında kesin bilgi yoksa kümeleme ölçütleri aracılığı ile en uygun sayıda küme ve alt kümeyi oluşturmak mümkün görülmektedir. Her ölçüt veri matrisinden farklı küme oluşturmaktadır. Ayrıca kümeleme yöntemleri ortaya koydukları yaklaşımların farklılığına bağlı olarak birim ya da değişkenlerin girdikleri kümelerin farklılaşmasına neden olmaktadır. Fakat ileri sürülen kümeleme ölçütleri kaç kümenin en uygun sınıflandırma olacağını ortaya koymuş değildir. (Krzanowski and Lai, 1988; Özdamar, 1988).

Tıbbi ve biyolojik uygulamalarda çok değişkenli veri setlerini ele alırken ortak yaklaşım, bir kaç grup içinden n bireyin bir bölümünü aramaktır, bu tip bir yaklaşım aynı grup içindeki bireylerin mümkün olduğunca birbirlerine benzemeleri fakat diğer gruplardaki bireylerden mümkün olduğunca ayrı olmalarını sağlamaktır. Burada istenilen, bulunan grupların, çalışılan populasyondaki doğal gruplar ile uyuşmasını ve grupların bu özellikleri ele alınarak doğal gruplar hakkında sonuç çıkarmayı sağlamaktır. Bu tür yaklaşımlar, tıp, psikoloji ve taksonomide (taxonomy) önemli yer tutmaktadır (Sneath and Sokal, 1973; Mardia, et al., 1979; Krzanowski and Lai, 1988).

Çok sayıdaki çokdeğişkenli gözlemleri, az sayıda türdeş gruplara ayırarak sınıflayan kümeleme çözümlemesi, p sayıda değişkenden, n birimi m sayıda türdeş sınıflara ayırır. Bu sınıflama yapılırken grup içi türdeşlik (minimum varyans) ve gruplar arası türdeş olmama (maksimum varyans) baz alınarak bir çok ölçüt tanımlanmıştır (Everitt,1979; Klastorin, 1983).

Grup içi türdeşliği enbüyüklüyen bir kümeleme ölçütünü ortaya koymak ve yapısal özellikleri tam bilinmeyen bir konuda elde edilen veri matrislerini uygun alt setlere bölmekte bu ölçütleri kullanmak nesnel bir yaklaşım olacaktır.

Bu araştırma, veri matrisini en uygun sayıda kümeye ya da alt kümeye ayırmada kullanılan;

- Gruplar içi türdeşliği enbüyüklemek (varyansı enküçüklemek), gruplar arası türdeşliği enküçüklemek (varyansı enbüyüklemek) için geliştirilen g_1, g_2 ve g_3 ölçümlerinin geçerliliğini test etmek,

- Belirlenen m sayıda kümenin elemanlarının p değişken yönünden türdeşliğini ve farklı gruplarda türdeş olmayışlarını çokdeğişkenli varyans çözümlemesi kullanarak Wilk's Lamda test istatistiği aracılığı ile oluşturulan küme sayısına bağlı olarak kümelenmenin sayısını test etmek,

- Çokdeğişkenli varyans çözümlemesi yöntemini kullanarak Hotelling-Lawley İz (Trace) ölçütü ile grup içi türdeşliği ve gruplar arası türdeş olmayışı test ederek, belirlenen küme sayısına göre değişimini incelemek,

- Küme sayısı belirlendikten sonra bu kümelerdeki birimlerin gerçek belirli ayırma fonksiyonlarına göre yeniden sınıflandırıldıklarında gerçek sınıflama ile yanlış sınıflama oranlarını ayırma çözümlemesi ile test etmek,

- Kümeleme öncesi benzerlik ölçüsü S_{ij} ile kümeleme sonrası bağlanma katsayıları C_{ij} arasındaki ilişkiyi ölçen kofenetik korelasyon katsayısını belirlemek ve kümelere göre geçerliliğini ve etkinliğini test etmek ve

- Farklı kümeleme ölçütlerinden elde edilen test istatistiklerini karşılaştırarak kümeleme çözümlemesinde uygun küme sayısının belirlenmesinde en etkin ölçütün belirlenmesi amacıyla yapılmıştır.

2. BENZERLİK KAVRAMI VE BENZERLİK ÖLÇÜLERİ

Birimleri belirli özelliklerine göre sınıflandırma ya da gruplandırma, bize birimler hakkında daha düzenli bilgiler vermektedir. Sınıflandırma, çoğu özelliği yönünden benzerlik gösteren birimleri bir ad altında toplamak olarak tanımlanabilir. Örneğin, hayvanlar; kediler, köpekler, atlar vd. olarak isimlendirilir. Hastalıklar, birimlerde meydana getirdiği belirtiler, klinik ve laboratuvar bulgularına göre tifo, paratifo, bronşit vb. gibi sınıflara ayırılırlar (Green, et al., 1967; Sethi, 1971; Ling, 1973; Scott and Knott, 1974; Vogt, et al., 1987).

Çevredeki incelenen birimleri birbirleriyle benzerlikleri yönünden belirli gruplar içinde toplayarak sınıflandırma yapmayı, birimlerin ortak özelliklerini ortaya koyma ve bu sınıflar ile ilgili genel tanımlamalar yapmayı sağlamıştır. İki değişkene göre incelenen ve farklı özellik taşıdıkları açıkça belli olan birimler, değişken sayısı arttırıldığında ortak yönlerinin de arttığı gözlenebilir ve çoğu benzer olan özelliklerinden dolayı bir sınıfta yer alabilirler. İşte bu nedenden dolayı birimleri sınıflandırırken bu birimlerin p adet değişkeni ölçülerek, tartılarak ya da nitel özellikler skor değerlerine göre sayısallaştırılarak veri matrisleri oluşturmak ve çokdeğişkenli bilgilere göre n bireyi sınıflara ayırmak, model sınıflar belirlemek bakımından en uygun yaklaşımdır. Böylece birimlerin, benzerlik gösterenlerini bir sınıfta toplamak ve bu benzerlikten yararlanarak grubun ortak özelliklerini tanımlamak ve incelemek değişkenler arasındaki neden-sonuç ilişkilerini çözümleyerek açıklamak daha kolay olacaktır (Overall and Klett, 1972; Hawkins, et al., 1982; Armitage, 1983).

Türdeş olmayan populasyonlardaki birimleri göresel olarak daha alt sayıda model sınıflara ayırma, model kalıpları belirleme ve verilerin temelinde yatan neden-sonuç ilişkilerini bilimsel olarak ortaya koymada kümeleme çözümlemesi etkin bir yöntemdir. Kümeleme çözümlemesinde, birey ya da değişkenler arasındaki benzerliklerden giderek bir birimin tüm ilişki çiftleri arasındaki çok değişkenli benzerliğin nicel bir ölçütünü hesaplayarak, bu ölçüt aracılığı ile n bireyin ayrılabileceği türdeş alt grupları tanımlamak ve benzerlik düzeylerine göre birimleri biraraya getirerek kümelere ayırma amaçlanmıştır (Ward, 1963; Marriott, 1974; Everitt, 1979; Murtagh, 1983; Özdamar, 1988; Krzanowski and Lai, 1988).

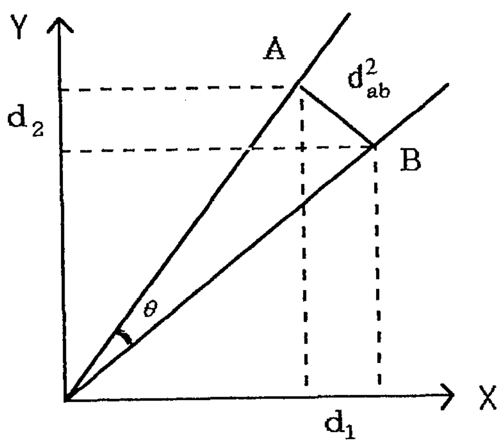
Birimler, ölçülen p değişken yönünden benzer, hemen hemen benzer ve farklı olmak üzere sınıflandırılabilir. Bu sınıflandırmayı bilimsel olarak yapmak için birimler arasındaki benzeşimleri benzerlik ya da uzaklık ölçüsü adı verilen nesnel ölçülerle değerlendirmek gerekir. Bu ölçüler; uzaklık türü ölçüler (distance like measures), ilişki türü ölçüler (correlation like measures), açısal uzaklık türü ölçüler, vektör çarpımları türü ölçüler ve diğer ölçüler olarak gruplandırılabilir (Overall and Klett, 1972; Özdamar, 1988; Johnson and Wichern, 1988; Lipkus, et al., 1988).

Birimlerin ya da değişkenlerin kümelenmesi için benzerlik matrisinden yararlanılır. S benzerlik matrisi, S_{jk} benzerlik ölçülerini içeren üçgen matris biçiminde, n(n-1)/2 elemana sahip bir matristir. S benzerlik matrisinin gösterimi Şekil 2.1'deki gibidir.

$$S = \begin{bmatrix} S_{21} & & & \\ S_{31} & S_{32} & & \\ \vdots & \vdots & \ddots & \\ S_{n1} & S_{n2} & \cdots & S_{n(n-1)} \end{bmatrix}$$

Şekil 2.1. S benzerlik matrisi gösterimi

İki boyutlu bir uzayda iki bireyin birbirlerine olan uzaklıkları Şekil 2.2'deki gibi gösterilebilir.



Şekil 2.2. Uzayda iki noktanın birbirine uzaklığının gösterimi

$d_{ab}^2=d_1^2+d_2^2$ olarak belirlenir. Çokdeğişkenli bir yapıda n bireyin p değişken için çok boyutlu bir uzayda yerleri vektör olarak belirtilir. İki birimin p boyutlu bir uzayda birbirine uzaklıkları Pisagor teoreminin genellemesine göre belirlenebilir. Bu uzaklıklar birimler arası benzerlik (similarity) ya da farklılık (dissimilarity) olarak isimlendirilir (Diggle, et al., 1976; Hawkins, et al., 1982; Johnson and Wichern, 1988).

A ve B gibi iki birim alalım ve bu birimlerin iki değişken için ölçümleri x ve y olsun. Bu ölçümlere göre uzaklık fonksiyonu d(A,B) olarak yazılabilir. Bu uzaklık fonksiyonun özellikleri aşağıdaki gibidir.

1. Simetri $d(A,B)=d(B,A)$
2. Pozitiflik $d(A,B)\geqslant 0$
3. Özdeşlik $d(A,A)=0$

Bazı uzaklık fonksiyonları yukarıda sayılan özeliklere ek olarak aşağıdaki özelliklere de sahiptir.

4. Belirlilik $d(A,B)=0$
   (Sadece ve sadece A=B ise doğrudur)
5. Üçgen eşitliği $d(A,B)\leq d(A,R)+d(R,B)$

Bazı amaçlar için sadece 1. ve 3. özellikleri sağlayan uzaklık fonksiyonlarından yararlanmak yeterlidir, bazı durumlarda ise 1. ve 5. özellikleri sağlayan fonksiyonlar gözönüne alınabilir. 1. özellik, özellikle Sosyolojik problemlerde herzaman geçerli olmayabilir. d(A,B) değerinin, A ve B arasındaki uzaklık ya da farklılık arttıkça artması beklenir. Bu yüzden d(A,B), 4. ve 5. özellikler sağlanamadığı durumlarda bile bir uzaklık fonksiyonu olarak tanımlanabilir (Mardia, et al. 1979).

Benzerlik ölçüleri içinde en çok kullanılan ve yukarıda sayılan özelliklerin büyük bir bölümüne uygun olan benzerlik ölçüleri; uzaklık türü ölçüler, ilişki türü ölçüler ve diğer benzerlik türü ölçülerdir (Anderberg, 1973; Green and Tull, 1973; Jambu and Lebeaux, 1986; Vogt, et al., 1987).

2.1. Uzaklık Türü Ölçüler

2.1.1. Öklid Uzaklığı

Birimler arasındaki uzaklığı, değişken ölçü birimlerinden etkilenmeksizin belirten bir ölçüdür. Bu nedenle kümeleme çalışmalarında en sık kullanılan benzerlik ölçüsüdür. Öklid ve Karesel Öklid olarak iki kullanılış biçimi vardır (Green and Tull, 1973; Marriott, 1974; Kendall, 1980; Jambu and Lebeaux, 1986; Vogt, et al., 1987; Özdamar, 1988).

(nxp) boyutlu veri matrisinde her satır bir sıra vektörü olarak alınırsa X_j ile X_k arasındaki uzaklık, d_{jk} ya da diğer bir adıyla n birimin oluşturduğu tüm mümkün n(n-1)/2 çift arasındaki uzaklıklar S benzerlik matrisini oluştururlar. Bu matrisin elemanları S_{jk}

$$S_{jk} = d_{jk} = \left[\sum_{i=1}^{p} \left(X_{ij} - X_{ik} \right)^2 \right]^{1/2} \tag{1}$$

olarak hesaplanır ve burada S_{jk}, i. değişken için j ve k birimleri arasındaki toplam uzaklığı belirtir. Bu uzaklık ölçüsü Öklid uzaklığı olarak adlandırılır. Öklid uzaklığı simetri ve pozitiflik özelliklerini taşır (Day and Heeler, 1971; Bock, 1975; Vogt, et al., 1987; Özdamar, 1988; Johnson and Wichern, 1988).

Diğer Öklid uzaklığı, öklid uzaklığının karesi olan Kare Öklid uzaklığıdır ve

$$S_{jk}^2=d_{jk}^2=\sum_{i=1}^{p}\left(X_{ij}-X_{ik}\right)^2 \qquad (2)$$

şeklinde hesaplanmaktadır (Anderberg, 1973; Kendall, 1980).

2.1.2. Minkowski uzaklığı

Birimler arasındaki uzaklığın

$$S_{jk}=d_{jk}=\left[\sum_{i=1}^{p}\left|X_{ij}-X_{ik}\right|^L\right]^{1/L} \qquad (3)$$

olarak hesaplandığı bir uzaklık ölçüsüdür. Öklid uzaklığının L üssü olarak bir genellemesidir. Sık kullanılan bir uzaklık ölçüsü değildir. L=2 alındığında Minkowski uzaklığı, Öklid uzaklığı olur (Anderberg, 1973; Johnson and Wichern, 1988).

2.1.3. Mahalanobis uzaklığı

Bu ölçüt de öklid uzaklığının genel bir çeşidi olarak adlandırılır ve

$$d_{ij}=\sqrt{(X_i-X_j)^T B\,(X_i-X_j)} \tag{4}$$

olarak hesaplanır. Burada T transpoze işlemini ifade eder ve B pozitif p*p boyutlu bir kare matristir ve $(X_i-X_j)^T B (X_i-X_j)=0$ olduğunda simetrik matris özelliğini taşır (Maronna and Jacovkis, 1974; Flury and Riedwyl, 1985; Vogt, et al., 1987; Johnson and Wichern, 1988).

2.1.4. Vektör çarpım uzaklığı

Vektör çarpım ölçütü, p boyutlu bir uzayda noktalar arasındaki veri vektörleri ve göresel uzunlukları arasındaki açısal farkın benzerlik ölçüsü olarak alındığı bir ölçüdür (Overall and Klett, 1972; Marriott, 1974).

X ve Y değişkenlerine ait sıra vektörleri

$X^T = (x_1, x_2, \ldots, x_n)$

$Y^T = (y_1, y_2, \ldots, y_n)$

şeklinde gösterilir. Bu iki vektörün içsel çarpımı

$$X^T Y = \sum_{i=1}^{m} x_i y_i \tag{5}$$

olarak ifade edilir ve $X^T Y$'nin karekökü vektör uzunluğu olarak alınır.

Bir vektörün kendisinin transpozu ile çarpımı X^TX, X kareler toplamı olarak adlandırılır ve $||X||$ şeklinde gösterilir. X ve Y vektörlerinin çarpımları kareler toplamı cinsinden ele alınırsa,

$$X^TY = \|X\| \, \|Y\| \operatorname{Cos}\alpha \tag{6}$$

eşitliği elde edilir. $\operatorname{Cos}\alpha$ iki nokta arasındaki uzaklığın ölçüsüdür ve

$$\operatorname{Cos}\alpha = \frac{X^TY}{\|X\| \, \|Y\|} \tag{7}$$

olarak hesaplanır. $\operatorname{Cos}\alpha$ değeri, X ve Y arasındaki benzerliğin bir ölçüsü olarak alınabilir ve birimlerin kümelenmesi yerine değişkenlerin kümelenmesi gerektiğinde kullanılan bir ölçüdür (Overall and Klett, 1972; Anderberg, 1973; Özdamar, 1988).

2.2. İlişki Türü Ölçüler

2.2.1. Pearson ilişki katsayısı

Değişkenler arası kümeleme sözkonusu olduğunda kullanılmakta olan bir benzerlik ölçüsüdür. Pearson İlişki katsayısı

$$r_{jk} = \frac{\sum_{i=1}^{p} \left(X_{ij}-\overline{X}_j\right)\left(X_{ik}-\overline{X}_k\right)}{S_jS_k} \tag{8}$$

olarak hesaplanır. Burada S_j j. birimin, S_k ise k. birimin standart sapmalarını belirtmektedir. Birimler arası benzerlik ölçüsü olarak kullanılmak istendiğinde birimlerin değişken vektörlerinden (sıra vektörleri) yararlanılarak Pearson ilişki katsayısı hesaplanabilir (Anderberg, 1973; Özdamar, 1988).

2.2.2. Gamma katsayısı

Bu benzerlik ölçüsü daha çok niteliksel verilere uygulanabilen bir ölçüdür. Sıralı ölçekle elde edilen verilerin 2x2 tablosu biçiminde gösterildiği durumlarda gözlerdeki değerler kullanılarak

$$Q = (ad - bc)/(ad + bc) \tag{9}$$

şeklinde hesaplanır (Wilkinson, 1984).

2.3. Diğer Benzerlik Ölçüleri

Yukarıda sayılan benzerlik ölçülerinden başka, birimlerin ve değişkenlerin benzerliklerini belirlemek için kullanılan çok sayıda benzerlik ölçüsü ileri sürülmüştür. Bunlardan en yaygın kullanılanlardan biri, sütun toplamları yüzde değerleri, PCT (Percent Column Totals) ölçütüdür. Bir uzaklık indeksi olan bu ölçüt, toplam içinde değerlerin karşılaştırılarak, iki profil içinde farklılaşan değerlerin yüzdesini gösterir.

Bir başka ölçüt olan Jaccard benzerlik katsayısı, mikrobiyolojik, taksonomik bulgularda karakterlere ilişkin ikili cevaplardaki benzerlik oranını göstermektedir.

Özellikle nominal ve ordinal ölçekle elde edilmiş ikili şekilde sınıflandırabilen ve olasılık kurallarına dayalı olan basit (simple) ve çiftli (double) eşleştirilmiş (matching) benzerlik ölçüleri taksonomik problemlerde sınıflandırma amacı ile kullanılmaktadır (Anderberg, 1973; Bursalıoğlu, 1985; SYSTAT 5.0 tutorial, 1990).

3. KÜMELEME YÖNTEMLERİ

Veri matrisinden birimlerin ya da değişkenlerin kümelenmesinde bir çok yöntemden yararlanılmaktadır. Bu yöntemlere kümeleme yöntemleri adı verilir.

Kümeleme yöntemleri, veri matrisinden elde edilen benzerlik matrisini çözümleme biçimine göre 5 gruba ayrılır.

Bunlar;

1. Aşamalı Kümeleme Yöntemleri
   (Hierarchical Clustering Methods)
2. Aşamalı Olmayan Kümeleme Yöntemleri
   (Nonhierchical Clustering Methods)
3. Ardışık Parçalama Yöntemleri
   (Iterative Partioning Methods)
4. Optimizasyon Yöntemleri
   (Optimizing Procedures)
5. Diğer Kümeleme Yöntemleri

olarak adlandırılırlar.

Bu yöntemlerden yaygın olarak kullanılanları, Aşamalı ve Aşamalı Olmayan Kümeleme Yöntemleridir (Anderberg, 1973; Marriott, 1974; Mardia, 1979; Hawkins, et.al., 1982; Murtagh, 1983; Johnson and Wichern, 1988).

3.1. Aşamalı Kümeleme Yöntemleri

Aşamalı Kümeleme Yöntemleri, birimlerin kümelenmesinde n birimin n(n-1)/2 ya da değişkenlerin kümelenmesinde p değişkenin p(p-1)/2 tüm olası çiftlerinin aralarındaki benzerlik düzeylerine göre, birimleri ya da değişkenleri birbirlerine aşamalı biçimde bağlamayı amaçlayan yöntemlerdir. Bu yöntemlerde benzerlik ölçülerine göre saptanan birimler ya da değişkenler arası uzaklıklar en yakın birimler birbirleriyle aşamalı olarak birleştirilerek bir ağaç grafiği (dendrogram) oluşturmak üzere birbirlerine bağlanırlar (Anderberg, 1973; Sneath and Sokal, 1973; Mardia, et al., 1979; Murtagh, 1983; Wilkinson, 1984; Özdamar, 1988; Corsten and Denis, 1990).

Ağaç grafiği, Şekil 7.1.'de de görüldüğü gibi X ekseninin birimlerin numaralarını, Y ekseninin ise bağlantı uzaklıklarını simgelediği bir grafik türüdür. Ağaç grafiğinde dallanma, n-1 sayıda dal ile başlar ve 2'li, 3'lü vd. birim bağlantıları ile bir gövdeye bağlanma biçiminde gelişir. Bu bağlantılar çeşitli uzaklık ölçütlerinde birimlerin oluşturdukları alt kümeleri verir (Mardia et al., 1979; Johnson and Wichern, 1988; Bağcı vd., 1991).

Aşamalı kümeleme yöntemlerinde kümeleme işleminde genellikle iki algoritmadan yararlanılır.

Bunlar

1. Birim kümelerine ait alt kümelerinin ard arda toplanması yığılması (Agglomerative Procedures)
2. Birim kümelerinin ard arda bölünmesi (Divisive Procedures)

olarak isimlendirilen algoritmalardır (Anderberg, 1973; Sneath and Sokal, 1973; Mardia, et al., 1979; Chatfield and Collins, 1980; Norusis, 1984; Hartigan and Engelman, 1988).

En yaygın kullanılan algoritma yığılımlı (Agglomerative) yöntemdir. Bu yöntem; oluşturulan kümelerin bir çekirdek etrafında küre veya elips şeklinde kümeleri birleştirir ve benzer birimleri bağlantı uzaklıkları göstererek bağlar ve küme çekirdeği varsayımı kurmadan (n-1) kümeden 1 küme oluşuncaya kadar aşama aşama birbirleriyle bağlanmalarını sağlar. İkinci algoritma birimler ard arda bölünerek (Divisive) kümeleri oluşturur. Gruplardan birbirini izleyenleri bölünerek, başlangıçta uzaklıklarına göre n birimden bir grup oluşturulur ve sonuçta bir birim kalana kadar biribirine bağlanması sağlanır.

Küme oluşturan birimlerin birbirlerine bağlanmasında benzerlik matrisinin farklı şekilde ele alınmasına göre bir çok aşamalı kümeleme yöntemi geliştirilmiştir.

Bunlardan en yaygın olarak kullanılanları;

1. Tek Bağlantı Kümeleme Yöntemi (Single Linkage Cluster Analysis)

2. Tam Bağlantı Kümeleme Yöntemi
   (Complete Linkage Cluster Analysis)
3. Ortalama Bağlantı Kümeleme Yöntemi
   (Average Linkage Cluster Analysis)
4. Küresel Ortalama Bağlantı Kümeleme Yöntemi
   (Centroid Linkage Cluster Analysis)
5. Ortanca Bağlantı Kümeleme Yöntemi
   (Median Linkage Cluster Analysis)
6. En Küçük Varyans Kümeleme Yöntemi
   (Ward Method)
7. Esnek Kümeleme Yöntemi
   (Lance and Williams Flexible Clustering Method)

olarak sayılabilir (Anderberg, 1973; Sneath and Sokal, 1973; Everitt, 1979; Özdamar 1988; Johnson and Wichern, 1988).

Bu yöntemler aşağıda kısaca açıklanmıştır.

3.1.1. Tek Bağlantı Kümeleme Yöntemi (TBK)

Aşamalı kümeleme yöntemleri içinde en basit ve en çok kullanılan yöntem Tek Bağlantı Kümeleme yöntemidir. Bu yöntem, Minimum yöntem, Bağlantı çözümlemesi, En Yakın Komşu Kümeleme Çözümlemesi (Nearest Neighbor Cluster Analysis) gibi isimlerle de anılır (Mardia, et al., 1979; Punj and Stewart, 1983; Vogt, et al., 1987; özdamar, 1988).

Bu yöntemde, birbirine en yakın olan noktalar gruplanır. Öncelikle noktalar arası n(n-1)/2 uzaklıkları küçükten büyüğe doğru sıralanır. n kümeden S_{jk} benzerlik ölçüsü en küçük olan m ve q kümeleri birleştirilerek işleme başlanır. İkinci aşamada birleştirilecek t ve r kümeleri için benzerlik ölçüsü S_{tr} olur. Bu şekilde uzaklık yönünden en yakın olanlar en güçlü biçimde birbirlerine bağlanarak aşamalı olarak kümelenirler (Chatfield and Collins, 1980; Johnson and Wichern, 1988; Özdamar, 1988).

Burada S_{jk} uzaklık türü bir ölçü ise

$$S_{tr}=\min\left(S_{mr},S_{qr}\right) \tag{10}$$

olarak, ilişki türü bir ölçü ise

$$S_{tr}=\max\left(S_{mr},S_{qr}\right) \tag{11}$$

olarak hesaplanır (Anderberg, 1973).

3.1.2. Tam Bağlantı Kümeleme Yöntemi (TAMBK)

Tam bağlantı kümeleme yönteminde, birbirlerine benzerlik ölçüsü ile en uzak iki birim bir araya getirilir ve küme oluşturulur. Bu yöntem En Uzak Komşu (Furthest Neighbour), Maksimum yöntem gibi isimlerle de anılır.

t ve r birimlerinin bir küme oluşturması için benzerlik ölçüsü S_{tr}, uzaklık türü cinsinden ise

$$S_{tr}=\max\left(S_{mr},S_{qr}\right) \tag{12}$$

şeklinde, ilişki türü ölçü cinsinden ise

$$S_{tr} = \min\left(S_{mr}, S_{qr}\right) \tag{13}$$

şeklinde hesaplanır.

TAMBK, TBK yönteminin tersine en uzak komşu niteliğine sahip olan birimleri birleştirerek küme oluşturur. Bu yöntemde kümeleme bittiğinde, her küme için maksimum değer dikkate alınacağından, bağlayıcı etkisi olmayan birbirine yakın kümeler meydana getirme eğilimindedir (Anderberg, 1973; Johnson and Wichern, 1988; Özdamar, 1988).

3.1.3. Ortalama Bağlantı Kümeleme Yöntemi (OBK)

S benzerlik matrisi elemanı S_{jk}'lar dikkate alınarak, m kümesinden bir birim ve q kümesinden bir birim alınarak oluşturulan tr çiftinin benzerlik ölçüsü S_{tr}, m ve q kümelerine ait bu birimlerin benzerlik ölçülerinin toplanması ile

$$S_{tr}=S_{mr}+S_{qr} \tag{14}$$

şeklinde hesaplanır.

En çok benzer çiftin bulunmasında grupiçi benzerlik ortalaması m ve q kümelerine ilişkin benzerlik ölçülerinden ve birim sayılarından yararlanılarak

$$\overline{S}_{mq}=\frac{T_m+T_q+S_{mq}}{\left(n_m+n_q\right)\left(n_m+n_q-1\right)/2} \tag{15}$$

şeklinde hesaplanır. Burada T_m, m kümesi içindeki birimlerin tüm mümkün çift benzerlikleri toplamı; n_m, m kümesindeki birim sayısı; T_q, q kümesi içindeki birimlerin tüm mümkün çift benzerlikleri toplamı; n_q, q kümesindeki birim sayısıdır. S_{mq} minimum olduğunda m ve q kümeleri birleştirilir. Daha sonra bunların sıra ve sütunları S benzerlik matrisinden çıkarılarak, S matrisi yeniden oluşturulur. Bu işlem, tüm birimlerin kümelerden herhangi birine atanması sağlanıncaya kadar devam eder (Anderberg, 1973; Sneath and Sokal, 1973; Hawkins, et al., 1982; Özdamar, 1988; Johnson and Wichern, 1988).

3.1.4. Ortanca Bağlantı Kümeleme Yöntemi (MBK)

Bu yöntem, genellikle değişkenlerin değerlerinin sıralı ölçekle elde edildiği ya da ölçüm değerleri yerine skor değerleri ele alındığında ilgili kümelerin ortaya çıkarılmasında kullanılır.

S benzerlik matrisi elemanı S_{jk}'lar dikkate alınarak, m kümesinden bir birim ve q kümesinden bir birim alınarak oluşturulan tr çiftinin benzerlik ölçüsü S_{tr} m ve q kümelerine ait bu birimlerin benzerlik ölçülerinin toplamları ele alınarak hesaplanır. S_{jk} uzaklık türü bir benzerlik ölçüsü ise,

$$S_{tr}=\frac{1}{2}\left(S_{mr}+S_{qr}\right)-\frac{1}{4}\,S_{mq} \qquad (16)$$

olarak ve S_{jk} ilişki türü bir ölçü ise,

$$S_{tr} = \frac{1}{2}\left(S_{mr} + S_{qr}\right) + \frac{1}{4}\left(1 - S_{mq}\right) \qquad (17)$$

şeklinde hesaplanır.

3.1.5. Küresel Ortalama Bağlantı Kümelemesi (KOBK)

S benzerlik matrisi elemanları, S_{jk}'ların kare öklid uzaklığı olduğu durumlarda kullanılan bu yöntemde m ve q kümeleri ilk aşamada birleştirildikten sonra t kümesinin diğer bir r kümesi ile birleştirilmesinde S_{tr},

$$S_{tr} = \frac{n_m}{2n_m} S_{mr} + \frac{n_q}{n_m + n_q} S_{qr} - \frac{n_m n_q}{n_m + n_q} S_{mq} \qquad (18)$$

olarak hesaplanır (Anderberg, 1973).

Bu yöntemi diğer yöntemlerden ayıran en önemli özellik, birleştirilecek kümenin ortalamalarının yeni kümenin ortalamasını hesaplamak için ağırlık olarak alınmasıdır. Ayrıca bu ağırlıkların her kümedeki birim sayısı ile orantılı olmasının gerekli olmamasıdır (Ward, 1963; Anderberg, 1973; Hawkins, et al., 1982).

3.1.6. En Küçük Varyans Kümeleme Yöntemi (WARD yöntemi)

Olguların fonksiyonel ilişki içinde olduğu durumlarda, bilgi kaybının ortaya çıktığı veri yapılarına uygulanması için ileri sürülen bir yöntemdir (Vogt, et al., 1987; Garety, et al., 1988).

Bu yönteme göre her birimin başlangıçta farklı bir alt grupta yer aldığı kabul edilir. Her aşamada iki alt grup bir sonraki seviyeyi oluşturmak için birleştirilir. Bu durumda k(k-1) alt grup olduğu varsayılır. Eğer bir önceki seviyede k alt grup varsa bunlardan kayıp fonksiyonunun artışını enküçükleyen (minimizing) grup seçilir. İki ya da daha fazla birleştirmeden oluşan aşamalar minimum değer ortaya koyuyorlarsa onlar arasından gelişigüzel bir seçim yapılır. Bu işlemler, tüm birimler aynı grupta yer alıncaya kadar tekrarlanır. Gelişigüzel kayıp fonksiyonunun seçimi küme ortalamalarından tüm birimlerin farklarının kareler toplamına (hata kareler toplamı) bağlıdır. k kümesinde yer alan n_1 noktanın k kümesinin ortalamalar vektörüne olan öklid uzaklıkları toplamı, hata kareler toplamıdır ve W_k olarak ifade edilir. W_k,

$$W_k = \sum_{i=1}^{p} \sum_{j=1}^{n_k} \left(X_{ijk} - \overline{X}_{ik}\right)^2$$

$$= \sum_{i=1}^{p} \sum_{j=1}^{n_k} X_{ijk}^2 - n_k \sum_{i=1}^{p} \overline{X}_{ik}^2 \qquad (19)$$

şeklinde hesaplanır. Burada, W_k değeri k=1,2,...,n kümelerde hesaplanarak, toplam kümeiçi hata kareler toplamı,

$$W = \sum_{i=1}^{n} W_k \tag{20}$$

şeklinde hesaplanır (Anderberg, 1973; Vogt, et al., 1987; Bools, et al., 1990).

3.1.7. Lance ve Williams'ın Esnek Kümeleme Yöntemi

Bu yöntemde kümelerin birleştirilmesinde kritik S_{jk} değerinin saptanması amaçlanır. m ve q kümeleri birleştirildikten sonra m ile r ve q ile r kümeleri birleştirildiğinde

$$S_{tr} = \alpha_m S_{rm} + \alpha_q S_{rq} + \beta S_{mq} + \gamma \left| S_{rm} - S_{rq} \right| \tag{21}$$

olarak hesaplanır. Burada $\alpha_m, \alpha_q, \beta$ ve γ optimal kümelemeyi sağlayacak parametrelerdir. Bu parametrelerin tahminleri deneysel olarak belirlenir ve birimlerin kümelenmesinde diğer aşamalı yöntemlerden herhangi biri kullanılabilir (Hawkins, et al., 1982).

3.2. Aşamalı Olmayan Kümeleme Yöntemleri

Aşamalı olmayan kümeleme yöntemleri, istenilen küme sayısının oluşturulabilmesi için bir parametre bulunmasını amaçlar. Bu yöntem, birimlerin rasyonel olarak seçilen sayıda küme oluşturmalarını sağlayan ve bu kümelerdeki birimlerin p sayıda değişkene ilişkin istatistiklerini elde etme ve kümelerin temel özelliklerini belirleme imkanı veren bir yöntemdir. Bu yaklaşımda, aşamalı olarak alt kümelerin (grupların) birbirlerine bağlanma düzeyleri ve bağlantı uzaklıkları yerine, tek seviyede olayı çözümlemek ve küme yapısını bulmak amaçlanmaktadır. Küme sayısı arttırıldığında ikinci bir çözüm bulunursa, oluşturulan bu kümeler ilk çözümde elde edilen kümenin bir alt kümesi olmayacak, başka bir küme olarak nitelendirilecektir. Böylece aşamalı yöntemde birbirleriyle bağlanarak ana kümeyi oluşturan birimler gerçekte başka özellikler gösteren kümeler olacaktır (Johnson and Wichern, 1988; Özdamar, 1988).

Aşamalı olmayan yöntemler içinde en yaygın kullanılanları K-Ortalamalar Yöntemi (MacQuen's K-Means Method), Yığma Kümeleme Yöntemi (Hill Climbing Method) olarak sayılabilir (Johnson and Wichern, 1988).

3.2.1. K-Ortalamalar Yöntemi

n birimin m kümeye ayrılmasında p boyutlu uzayda gösterimi olan en yakın ortalamaya ilişkin çekirdek noktaya atanmasını içerir. Ortalamalar, başlangıçta alınan m noktanın değerleridir (Anderberg, 1973; Jamison, 1988).

Bu yöntemde, n birimin m kümeye parçalanması olasılık kurallarına göre çekirdek noktalar belirleyerek ya da rastgele belirlenen m küme sayısı seçerek ya da verilerin incelenmesi ile deneysel olarak seçilen m sayıda küme belirleyerek yapılabilir. Birimler m sayıda kümeye ayrılması şu şekilde sağlanır. Deneysel ya da olasılık kurallarına göre belirlenmiş m nokta, oluşacak kümelerin merkezleri olarak kabul edilir. Bu çekirdekler etrafında, minimum varyans veren noktalar aranır ve veri yapısına göre ilk ele alınan noktaların minimum varyans kavramını doğrulayıp doğrulamadığı araştırılır. Eğer seçilen nokta hala geçerli küme çekirdeği özelliğini taşıyorsa işlem devam eder. Eğer çekirdek veri yapısı içinde minimum varyans kavramını gerçeklemiyorsa gerçek noktalardan en uygun olanı çekirdek ile yer değiştirir. Bu işlem en uygun kümeler içi varyansın minimum, kümeler arası varyansın enbüyük olduğu koşul gerçekleşinceye kadar devam eder ve birimlerin kümelere atanması optimum düzeyde sağlanır.

Birimleri kümelere atayan bu yöntemde, her birim Öklid uzaklığı ile kendisine en yakın olan küme merkezi seçilen nokta ile kümelenir (Anderberg, 1973; Wilkinson, 1984; Sözer vd., 1986; Özdamar, 1988).

Bu yöntem birimlerin incelenmesi ile k=2'den başlayarak küme sayılarını her defasında arttırarak deneysel olarak en uygun kümelemeyi bulma şeklinde uygulanabilir (Wilkinson, 1984; Kristeller and Robin, 1989).

3.2.2. Yığma Kümeleme Yöntemi

Bu yöntemde, birimler en yakın ortalamalı kümeye atanma yerine önceden belirlenen bir istatistiksel ölçüte göre bir kümeden diğerine hareket ederek en uygun durum sağlanıncaya kadar yeniden atama işlemleri sürdürülerek kümeler oluşturulur. Bu istatistiksel ölçüt, toplam küme içi Kovaryans matrisi W ve kümeler arası Kovaryans matrisi B ile ilgili olarak geliştirilmiştir. Toplam Varyans-Kovaryans matrisi

T'nin determinantı |T|'nın, |W|'ye oranının maksimum olduğu değerlerin logaritması olarak alınan bir C katsayısı hesaplanarak bu değere göre kümelerin parçalanması ileri sürülmüştür (Anderberg, 1973; Özdamar, 1988).

4.KÜMELEME ÖLÇÜTLERİ

Kümeleme çözümlemesinde grupiçi türdeşliği ve gruplararası türdeş olmamayı sağlamak amacıyla bir çok ölçüt tanımlanmıştır.

Birimlerin ayrılabilecekleri küme sayısı belirlendikten sonra, kümeler için belirlenen kümeleme ölçütleri, birimlerin kümeler içi türdeşliğini ve kümeler arası türdeş olmamasının sağlandığı durumda istenen kümelemeyi verir. Bu amaçla g_1, g_2, g_3, trW, |W|, r_{cs}, LR (Likelihood Ratio), Wilk's Lamda, Hotelling Lawley İz değeri gibi birçok kümeleme ölçütü tanımlanmıştır (Overall and Klett, 1972; Anderberg, 1973; Sneath and Sokal, 1973; Klastorin, 1983; Punj and Stewart, 1983; Johnson and Wichern, 1988).

Birimlerin küme içi ve kümeler arası ortalama uzaklıklarını ölçmek için farklı yöntemler tanımlanabilir. Küme içi türdeşliğin r. ölçümünü göstermek için w_r ve kümeler arası türdeş olmamanın r. ölçümünü belirtmek için b_r tanımlamasını kullanılabilir. Küme içi türdeşliği enküçüklemek, kümeler arası türdeş olmamayı enbüyüklemek istendiği zaman, belirleme ölçütü iki farklı yöntemle aşağıdaki gibi tanımlanabilir.

Bu ölçütler

1. $g_r = w_r - b_r$

2. $g_r = w_r / b_r$

şeklinde hesaplanır (Overall and Klett, 1972 ; Klastorin, 1983).

Bu ölçütlerin geçerliliğinin test edilmesinde yararlanılan z test istatistiğinin kullanılabilmesi için, her bir r. istatistiğin tahmin edilen değeri $E(g_r)$ ve varyansı $Var(g_r)$ hesaplanır. Hesaplamada en uygun yaklaşım, kümelerin sayısı m ve grup büyüklüğü n_k'nın sabit tutulacağı ve m ile n_k'nın bu değerlerine sahip bütün olası bölümlerini kabul etmektir. Bütün olası parçaları birer birer saymaktan kaçınmak için, hastalıkların yer-zaman kümelenmesinde (space-time clustering) geliştirilen bir yaklaşımdan yararlanılarak permütasyonel dağılımının parametrelerini doğrudan bulmak mümkündür. Mantel Yöntemi olarak bilinen bu yaklaşım her türlü simetrik ve simetrik olmayan durumlar için genellenmiştir. Bu yöntemden yararlanarak, birimler arası kümelenme istatistiği z'nin beklenen değeri E(z) ile olan farkını, z'nin varyansı Var(z) ile orantılayarak standart normal dağılım yaklaşımı ile kümelemenin önemliliği test edilebilir (Özdamar, 1981,1983; Klastorin, 1983).

4.1. g_r Ölçütleri

g_r ölçütleri 3 farklı yöntemle tanımlanmıştır.

4.1.1. g_1 Ölçütü

Küme içi türdeşliğinin ölçümünü gösteren w değeri, her bir küme içindeki birimler arasında başlangıçdaki ölçülmüş çiftlerin uzaklıklarını ifade eder ve bu değerlerin ağırlıksız ortalamalarını içerir. Bu değer,

$$w_1 = \frac{1}{m}\left\{\sum_{k=1}^{m}\frac{2}{n_k(n_k-1)}\left[\sum_{\substack{i,j\in G_k \\ (i<j)}} d_{ij}\right]\right\} \tag{22}$$

şeklinde hesaplanır. Burada m küme sayısı, n_k her küme içindeki birim sayısı, d_{ij} her birime ait uzaklık ölçüsüdür.

Aynı model içinden küme uzaklıkları değerini gösteren b_1, ağırlıksız ortalama olarak ifade edilir ve

$$b_1 = \frac{2}{m(m-1)}\left\{\sum_{k=1}^{m-1}\sum_{l=k+1}^{m}\frac{1}{n_k n_l}\left[\sum_{i\in G_k}\sum_{j\in G_k} d_{ij}\right]\right\} \tag{23}$$

şeklinde hesaplanır. Burada n_k ve n_l farklı kümelerdeki birim sayılarını gösterir.

Buradan g_1 ölçümü yukarıda da verildiği gibi iki farklı şekilde,

1. $g_{1a} = w_1 - b_1$ (24)
2. $g_{1b} = w_1 / b_1$ (25)

olarak hesaplanır (Overall and Klett, 1972; Klastorin, 1983).

4.1.2. g_2 ölçütü

Küme içi türdeşlik ve kümeler arası türdeş olmamanın diğer bir tanımı da, her bir küme içi uzaklıkları ortalamasının bütün küme içi uzaklıkları ortalaması olan w_2'nin, $(n_k-1)/2$ şeklinde ağırlıklandırılması ile şu şekli alır

$$w_2 = \sum_{k=1}^{m} \frac{1}{n_k} \left[\sum_{\substack{i,j \in G_k \\ i<j}} d_{ij} \right] \tag{26}$$

w_2'nin alternatif bir tanımı, (nxp) boyutlu bir uzayda X veri matrisinden türetilmiştir. Toplam ayrılış (farklılaşma, scatter) matris

$$T = X'X \tag{27}$$

şeklindedir. Bu değer, kümeleme için bilinen kümeleme ölçütü varyans kriteri olur

$$T = W + B \tag{28}$$

şeklinde hesaplanır. Burada W küme içi ayrılış matrisi ve B kümeler arası ayrılış matrisini göstermektedir. W ve B değerleri,

$$W = \sum_{k=1}^{m} \left[\sum_{i \in G_k} (R_{ik} - C_k)' (R_{ik} - C_k) \right] \tag{29}$$

$$B = \sum_{k=1}^{m} n_k C_k' C_k \tag{30}$$

formülleri ile hesaplanır. Burada R_{ik}, X matrisinde i. satır vektörü ve C_k, G_k grup için ortalama vektördür (Arnold, 1979 ; Klastorin, 1983).

Bir çok yayında tanımlandığı gibi (Vogt, et al., 1987; John, et al., 1988), İz(W) (Trace (W)) olarak adlandırılan değer, toplam kümeler içi kareler toplamı olup w_2 değerine eşittir. b_2 değeri de iz(B) değerine eşitlendiğinde, g_2 ölçütü

$$g_2 = iz(W) - iz(B) \tag{31}$$

şeklinde hesaplanır. Buradaki değerler herhangi bir birimin seti için sabit olduğundan g_2 ölçütü, w_2 değerine indirgenebilir (Klastorin, 1983).

4.1.3. g_3 ölçütü

g_3 ölçütü, bir küme içindeki $\binom{n_k}{2}$ çiftlerinin sayısı ile ağırlıklı ortalama çift uzaklıklarının ortalaması olarak tanımlanır. Bu durumda,

$$w_3 = \frac{\sum_{k=1}^{m}\left[\sum_{\substack{i,j\in G_k\\(i<j)}} d_{ij}\right]}{\sum_{k=1}^{m}\binom{n_k}{2}} \tag{32}$$

ve

$$b_3 = \frac{\left\{\sum_{k=1}^{m-1}\sum_{l=k+1}^{m}\left[\sum_{\substack{i,j\in G_k\\(i<j)}} d_{ij}\right]\right\}}{\left[\binom{N}{2}-\sum_{k=1}^{m}\binom{n_k}{2}\right]} \tag{33}$$

buradan g_3 ölçütü iki farklı şekilde,

1. $g_{3a} = w_3 - b_3$ (34)
2. $g_{3b} = w_3 / b_3$ (35)

şeklinde hesaplanır (Klastorin, 1983).

4.2. Wilk's Lamda

Çokdeğişkenli varyans çözümlemesinde ortaya çıkan sonuçların değerlendirilmesinde kullanılan Wilk's Lamda değeri, grup ortalama vektörleri arasındaki farklılığın değerlendirilme ölçütü olarak kullanılmaktadır.

Çokdeğişkenli varyans çözümlemesinde (MANOVA) m grup için toplum ortalama vektörü modeli

$$X_{lj} = \mu + \tau_l + e_{lj} \qquad j = 1,2,\ldots,n_l \quad \text{ve} \quad l = 1,2,\ldots,m \tag{36}$$

şeklindedir. Burada toplum μ vektör ortalamasını, τ_l , l. grup etkisi ve e_{lj} hatayı göstermektedir.

Buradan birim vektörü

$$\underset{\text{Birim}}{x_{lj}} = \underset{\text{Örnek Ort.}}{\bar{x}} + \underset{\text{İşlem Etkisi}}{(\bar{x}_l - \bar{x})} + \underset{\text{Hata}}{(\bar{x}_{lj} - \bar{x}_l)} \tag{37}$$

olarak yazılabilir (Johnson and Wichern, 1988).

Çok değişkenli analizlerde çapraz çarpımlar (cross-product)

$$(\bar{x}_{lj} - \bar{x})(\bar{x}_{lj} - \bar{x})' \tag{38}$$

dan l ve j için

$$\sum \sum_{j=1}^{n_l} (x_{lj} - \overline{x})(x_{lj} - \overline{x})' = \sum_{l=1}^{m} n_l (\overline{x}_l - \overline{x})(\overline{x}_l - \overline{x})' + \sum_{l=1}^{m} \sum_{j=1}^{n_l} (x_{lj} - \overline{x}_l)(x_{lj} - \overline{x}_l)' \quad (39)$$

nel Kareler Toplamı — İşlemler Arası Kareler Toplamı — Gruplar İçi Kareler Toplamı (Hata Kareler Toplamı)

otal Sum Of Squares) — (Treatment Between Sum Of Squares) — (Residual Within Sum Of Squares)

Çapraz çarpımlar matrisinde hata kareler toplamı (Within sum of squares)

$$W = \sum_{l=1}^{m} \sum_{j=1}^{n_l} (x_{lj} - \overline{x}_l)(x_{lj} - \overline{x}_l)' \quad (40)$$

$$= n(n_1 - 1)S_1 + (n_2 - 1)S_2, \ldots, (n_m - 1)S_m$$

burada S_l l. örneğin kovaryans matrisidir.

Hipotez : $H_0: \tau_1 = \tau_2 = \ldots = \tau_m = 0$

"Gruplar arasında farklılık yoktur." şeklinde kurulur (Johnson and Wichern, 1988)

Bu etkileri gösteren değerler Çizelge 4.1'de MANOVA tablosunda gösterilir.

Çizelge 4.1 Çokdeğişkenli Varyans Analizi (MANOVA) Tablosu

Değişim Kaynağı	Serbestlik Derecesi	Matris Kareleri Toplamı Çapraz Çarpımlar
Gruplar	$m-1$	$B=\sum_{l=1}^{m} n_l(\bar{x}_l-\bar{x})(\bar{x}_l-\bar{x})'$
Hata	$\sum_{l=1}^{m} n_l - m$	$W=\sum_{l=1}^{m}\sum_{j=1}^{n_l}(x_{lj}-\bar{x}_l)(x_{lj}-\bar{x}_l)'$
Genel	$\sum_{l=1}^{m} n_l - 1$	$B+W=\sum_{l=1}^{m}\sum_{j=1}^{n_l}(x_{lj}-\bar{x})(x_{lj}-\bar{x})'$

Varyans analizinin çok değişkenli genellemesi tek değişkenli varyans analizinde kare toplamlarının parçalanmasına benzer şekilde Genel Kareler Toplamları ve Çapraz çarpımlar matrisinin parçalanmasını kapsar. Bu durumda test istatistiği iki determinantın genel varyans oranı olarak hesaplanır ve

$$\lambda=\frac{|W|}{|B+W|}=\frac{\left|\sum_{l=1}^{m}\sum_{j=1}^{n_l}(x_{lj}-\bar{x}_l)(x_{lj}-\bar{x}_l)'\right|}{\left|\sum_{l=1}^{m}\sum_{j=1}^{n_l}(x_{lj}-\bar{x})(x_{lj}-\bar{x})'\right|} \qquad (41)$$

olarak yazılır. Bu oran Wilk's Lamda olarak bilinmektedir. Aynı zamanda benzerlik oranı kriteri olarakta tanımlanmaktadır (Likelihood Ratio Criterion) (Overall and Klett, 1972; Bock, 1975; Johnson and Wichern, 1988; Kieser and Groeneveld, 1989).

Buradan $|T|=|W+B|$ olarak yazılır ve bu değer kısaca;

$$\lambda = \frac{|W|}{|T|} \tag{42}$$

değerine eşittir.

λ istatistiği, gruplar içi varyansın toplam varyansa oranı yeterince küçük ise tüm gruplar için çok değişkenli ortalamaların eşit olduğu hipotezi reddedilir. (Overall and Klett, 1972)

λ 'nın önemliliği F test istatistiği ile test edilir ve

$$z = \text{antilog}_e \frac{\log_e \lambda}{s} \tag{43}$$

değerinden;

$$F = \frac{1-z}{z} \frac{ms[p(k-1)-2]/2}{p(k-1)} \tag{44}$$

olarak hesaplanır. Burada, p değişken sayısı, k grup sayısıdır ve

$$s = \sqrt{\frac{p^2(k-1)^2-4}{p^2+(k-1)^2-5}} \tag{45}$$

$$m = n-1-1/2(k+p) \tag{46}$$

değerlerine eşittir. F test istatistiği p(k-1) ve ms-(p(k-1)-2)/2 serbestlik dereceli F dağılımı ile test edilir (Marriott, 1971; Klastorin, 1983; Johnson and Wichern, 1988).

4.3. Hotelling-Lawley İz Ölçütü

Çokdeğişkenli varyans çözümlemesinde ortaya çıkan sonuçların değerlendirilmesinde kullanılan Hotelling-Lawley İz ölçütü, aynı Wilk'slamda ölçütünde olduğu gibi grup ortalama vektörleri arasındaki farklılığın değerlendirilme ölçütü olarak kullanılmaktadır.

Çokdeğişkenli varyans çözümlemesinde Hotelling-Lawley iz ölçütü $(tr(W^{-1}B))$ şeklinde ele alınan bir ölçüttür. Bu ölçütün enbüyüklenmesi gruplar arasındaki varyansın büyüklenmesi olarak ele alınmaktadır ve Hotelling T^2 istatistiğine dayandırılarak önemliliği hesaplanmaktadır. (Bock, 1975; Lee, 1979; Hawkins, et al., 1982).

Hotelling T^2 istatistiği, tek değişkenli Student t testinin bir çok değişkenli genellemesidir.

Hotelling T^2 istatistiğinde, t testindeki ortalama yerine ortalamalar vektörü ve standart sapma değerinin yerine kovaryans matrisi konularak yazılabilir. $VV'=\sum$ gibi gruplariçi kovaryans matrisi $\sum$ 'nın karekök faktörü V' olsun ve bu ardışık matrislerin inversleri V^{-1} ve V'^{-1} olur. Bu durumda $V'^{-1}V^{-1}=\sum^{-1}$ dir.

t istatistiğine benzer olarak

$$T=\sqrt{n}\frac{(\overline{X}-\mu)'}{V'}=\sqrt{n}(\overline{X}-\mu)'V'^{-1} \quad (47)$$

ve T^2 değeri

$$T^2=n(\overline{X}-\mu)'\sum{}^{-1}(\overline{X}-\mu) \quad (48)$$

olur. Burada, $\overline{X}$ örnek için ortalama vektör, μ toplum ortalama vektörü ve $\sum^{-1}=V'^{-1}V^{-1}$ örnek kovaryans matrisinin inversidir (Overall and Klett, 1972; Johnson and Wichern, 1988).

T^2 test istatistiğinin önemliliğini test etmek p ve n-p serbestlik dereceli F dağılımına dayanır ve

$$F=\frac{T^2(n-p)}{(n-1)p} \quad (48)$$

olarak test edilmektedir (Bock, 1975; Lee, 1979; Hawkins et al., 1982).

4.4. Ayırma Çözümlemesi

Çoklu ayırma çözümlemesi, populasyonlar ya da çok sayıdaki grup arasındaki ilişkileri araştırarak ve aynı zamanda grup birimlerinin sınıflandırılmasını sağlayan çokdeğişkenli bir yöntemdir. Birimlerin farklı populasyonlardan çekildiği ve her birim için p sayıda farklı ölçüm olduğu durumlarda uygulanır. p sayıda ölçümün farklı populasyonlar içinde eşit varyans-kovaryans matrisine sahip normal dağılım gösterdiği kabul edilir. Kısaca iki ve daha fazla gruplu durumlarda bireylerin uygun fonksiyonlar yardımıyla minimum hata ile ait oldukları gruplara sınıflandırılmasını sağlar. (Flury and Riedwyl, 1985; Kurtuluş, 1985; Johnson and Wichern, 1988)

Klinik araştırmalarda çoklu ayırma çözümlemesinin önemliliği, p sayıda ölçümde farklı gruplar arasındaki ilişkilerin araştırılması ve farklı gruplardaki birimlerin sınıflandırılması için temel olarak verilebilmesidir (Overall and Klett, 1972; Kieser and Groeneveld, 1989).

Çoklu ayırma çözümlemesi, farklı grupların elemanları arasındaki sınıflandırma için maksimum potansiyele sahip bir ya da daha fazla ağırlık kombinasyonuna, çoklu ölçümlerin indirgenmesi ile sonuçlanır. Farklı grupların bireyleri arasındaki farklılaşmaları bulmak için, çoklu ölçümler maksimum etkinliğe sahip bir ya da daha fazla ağırlıklı bileşenlere indirgeme amaçlanır. İlk ayırma fonksiyonu (DF1 Canonical Variate), grup içi değişkenliğe göre gruplar arasındaki ayrılmanın maksimum ortalamasını ifade eden tüm olası bileşenleri içine alan ağırlıklı bir bileşkendir. İkinci ya da nadiren üçüncü DF ise, ilk DF1 ile gruplararasındaki ayrışmanın iyi yapılmadığı durumları açıklayan ağırlıklı bileşenleri içine alan bir bileşkedir.

Doğrusal bir bileşkenin gruplar arası ve gruplar içi varyansları, gruplar arası ve gruplar içi kovaryanslarla belirlenebilir. Çoklu ayırma çözümlemesi, gruplar içi varyansa göre gruplar arası maksimum varyansa sahip olacak tipte bir yapıyı belirleyecek bileşke kaysayıları seti belirlenmeye çalışılır.

Enbüyüklemeyi sağlayacak fonksiyon;

$$f(a)=\frac{a'Ba}{a'Wa} \tag{50}$$

olarak yazılır. Burada B gruplar arası kovaryans matrisi, W gruplar içi Kovaryans Matrisi'dir.

Kriter fonksiyon uygunluk katsayılarını elde etmek için en büyüklendiğinden

$$\begin{aligned}(B-\lambda W)a&=0\\ &=\left[W^{-1}B-\lambda I\right]a=0\end{aligned} \tag{51}$$

olur ve buradan

$$\begin{aligned}DF_1&=a_1x_1+a_2x_2+\ldots+a_kx_k\\ DF_2&=a_1x_1+a_2x_2+\ldots+a_kx_k\end{aligned} \tag{52}$$

Diskriminant fonksiyon değerleri bulunur (Overall and Klett, 1972; Bock, 1975; Hawkins, et al. 1982; Johnson and Wichern, 1988).

4.5. Kofenetik Korelasyon Katsayısı

Kofenetik korelasyon katsayısı, orijinal benzerlik matrisi S ile ağaç grafiği oluşturulduktan sonra bağlantı katsayılarını içeren katsayılar matrisinin C, karşılıklı elemanları arasındaki ilişkiyi ölçen bir ilişki katsayısıdır ve r_{cs} ile gösterilir (Sneath and Sokal, 1973; Anderberg, 1973; Hawkins, et al. 1982).

Bu ölçüm mikrobiyoloji, biyoloji ve zooloji vd. bilim dallarında taksonomik problemlerde birimlerin, desenlerin, varyantların sınıflandırılmasında geçerliliğini test etmekte popüler bir yöntemdir (Anderberg, 1973; Sneath and Sokal, 1973; Bursalıoğlu, 1985; Bağcı vd., 1991).

S ve C matrislerinin karşılıklı elemanları S_{ij} ve C_{ij}'lerin ilişkileri

$$R(S_p,S_q)=\sum_{i=2}^{i=n}\sum_{j=1}^{j=i-1}W(i,j)\left[S_p(i,j)-S_q(i,j)\right]^2 \qquad (53)$$

olarak hesaplanır. Burada n birim sayısı ve W(i,j) seçilen ağırlık fonksiyonudur. Toplumdaki indisler i satırları, j de sütunları gösterecek şekilde alt üçgenden alınır. r_{cs}'nin önemliliği Pearson r_{xy} ilişki katsayısı ile benzer şekilde test edilir. Eleman sayıları S ve C alt üçgen matris olduğunda n(n-1)/2 olarak alınır.

5. VERİ TÜRETİMİ

Araştırmamızda yararlanılan veriler çokdeğişkenli normal dağılım varsayımından yararlanılarak türetilmiştir. Değişken ve birim sayısı seçilirken, çok uzun bilgisayar zamanı gerektirmeyen ve kümeleme çözümlemesinin felsefesine uygun veri matrisleri hedeflendi. Bu nedenle değişken sayısı p=6, birim sayısı n ise gruplara göre sırasıyla 10 ile 60 arasında değişecek şekilde seçildi.Kümeleme çözümlemesinin felsefesinin altında az sayıda birimin çokdeğişkenli gözlemleri incelenerek birimlerin incelenen konu ile ilgili yapısal farklılaşmalarını kısa zamanda yakalamak yatmaktadır. Kümeleme çözümlemesi, hakkında açık bilgi olmayan bir konuda toplanan gözlemlere ilişkin doğal sınıflarla ilgili neden-sonuç ilişkisini ortaya koymak için yararlanıldığından, birim sayısı az gözlemlerden bilgi edinme ön planda gelmektedir (Sneath and Sokal, 1973; Anderberg, 1973; Kendall, 1980). Bu nedenle veri matrisinde birim sayısı kısıtlı tutuldu.

Çokdeğişkenli normal dağılım varsayımı ile ilk aşamada rastgele 6 değişken için 10, 20, 30, 40 ve 50 birimlik veriler türetildi. Türetilen bu gruplara rasgele için R ve içerdiği birim içinde n gruplar oluşturuldu.

Bu gruplar,

10 birimlik değerler için R10
20 birimlik değerler için R20
30 birimlik değerler için R30
40 birimlik değerler için R40
50 birimlik değerler için R50

olarak tanımlandı.

Rasgele verilerin türetiminde çokdeğişkenli normal dağılımdan çekim yapıldı. Bunun için ortalama vektörü $\mu=[0,0,0,0,0,0]$ olarak ve varyans kovaryans matrisi ise alt üçgen matrisi olarak aşağıdaki gibi alındı,

$$V=\begin{bmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ 0 & 0 & 1 & & & \\ 0 & 0 & 0 & 1 & & \\ & & & & & \\ & & & & & 1 \end{bmatrix}$$

Bratley, Fox and Schrage'de (1983, s. 153) verilen türetim yaklaşımından yararlanılmıştır.

Bu yaklaşımda, $X=(X_1,...X_6)$ değerle $\mu = [\mu_1, ... ,\mu_6]$ ortalamalar vektörlü normal rastgele değişken vektörü olarak alınmıştır. Kovaryans matrisi

$$V=E\left[(X-\mu)(X-\mu)'\right]=\begin{bmatrix} V_{11} & & & V_{1m} \\ & V_{22} & & \\ & & & \\ V_{m1} & & & V_{mm} \end{bmatrix}$$

şeklinde hesaplanmıştır. İşlem alt üçgen matris C hesaplatılarak, bu V=CC' olarak alınan değerle, 6 bağımsız standart normal bileşenleri olan Z vektörü oluşturularak $X=\mu+CZ$] sonucuna ulaşılmıştır. Çokdeğişkenli normal dağılım oluşturmak için kullanılan en iyi yöntem Cholesky faktörizasyonudur ve C 'yi hesaplamak için aşağıdaki yaklaşım kullanılır.

$$i = 1,\ldots,m \qquad j = 1,\ldots,i-1 \qquad c_{ij} = \frac{\left(v_{ij} - \sum_{k=1}^{j-1} c_{ik}\, c_{jk}\right)}{c_{jj}};$$

$$c_{ji} = 0 \qquad c_{ii} = \left(v_{ii} - \sum_{k=1}^{i-1} c_{ik}^{2}\right)^{1/2}$$

P=6 için kümelemede kullanılan değişken değerleri aşağıdaki gibi elde edildi. Türetimlerde ve veri dönüşümlerinde BMDP1D modülünden yararlanıldı.

BMDP 1D modülünde;

```
/INPUT VARIABLES=0.
 CASES=10.
/TRANSF.
 X1=RNDG(1132).
 X2=RNDG(1281).
 X3=RNDG(3456).
 X4=RNDG(2346).
 X5=RNDG(3675).
 X6=RNDG(2560).
 SAVE="R10.DAT".
```

```
FORMAT=FREE.
KEEP X1 TO X6.
/PRINT DATA.
/END.
```

örnek olarak yazılan programla türetimler yapıldı.

İkinci aşamada, koşullu veri olarak tanımlanan ve çokdeğişkenli normal dağılımdan p=6 ve n=10 için rasgele verilerdeki gibi türetilmiş veri matrisi elde edildi. Bu veri matrisi değişkenlere göre aşağıdaki dönüşüme uğratılarak kullanıldı.

```
X1=X1*10+30
X2=X2*10+32
X3=X3*10+34
X4=X4*10+36
X5=X5*10+38
X6=X6*10+40
```

Rasgele (R) gruplarının her birine koşullu n=10, p=6 veri matrisi ayrı ayrı eklenmiş ve yeni Koşullu (K) veri grupları oluşturulmuştur.

Koşullu setleri,

R10+10 birimlik değerler için K20
R20+10 birimlik değerler için K30
R30+10 birimlik değerler için K40
R40+10 birimlik değerler için K50
R50+10 birimlik değerler için K60

olarak tanımlanmıştır.

Koşullu grup açıkça iki farklı vektör ortalamalarına sahip iki çokdeğişkenli normal dağılımın birleştirilmesinden oluşmuştur.

6. VERİ ÇÖZÜMLEME YÖNTEMLERİ

6.1. Kümeleme Yöntemleri

Rastgele ve koşullu veri setleri için, BMDP 2M modülü ve SPSS/PC+ paket programları aracılığı ile Öklid ve kare öklid uzaklığı içeren benzerlik matrisleri belirlendi. BMDP programında KM modülü aracılığı ile Öklid uzaklığı kullanılarak TBK yöntemi ile tüm grupların ağaç grafikleri elde edildi.

BMDP KM modülü aracılığı ile K-Ortalama yönteminden yararlanarak küme sayıları 2, 3, 4 ve 5 olacak şekilde tüm gruplar için birimlerin girdikleri kümeler küme sayılarına göre belirlenmiştir.

6.2. Kümeleme Ölçütleri

Araştırmamızda türetilen rastgele birimlerin oluşturdukları kümelerin uygun olup olmadıklarını test etmek amacıyla g1 ölçütünün g_{1a} ve g_{1b} uygulamaları, g_2, g_3 ölçütünün g_{3a} ve g_{3b} uygulamaları, Çokdeğişkenli varyans çözümlemesi istatistikleri Wilk's Lamda, Hotelling Lawley iz ölçütü, Kümeleme çözümlemesi ile oluşan kümelerin girdikleri küme numaralarını belirten sayılarla tanımlanan gruplamaların doğru sınıflanma olasılıklarını belirlemek için ayırma çözümlemesi ve Kofenetik korelasyon katsayısı ölçütleri kullanılmıştır.

Bu ölçütlerin hesaplanma biçimleri aşağıda verilmiştir.

6.2.1. g_1 ölçütü

Küme içi türdeşlik ve kümeler arası farklılaşma değerini veren g_1 ölçütü için 2 yöntemle

1. $g_{1a}=w_1-b_1$
2. $g_{1b}=w_1/b_1$

şeklinde hesaplama yapılmıştır. w_1 ve b_1 değerlerinin hesaplanması formül (22) ve formül (23)'den sırasıyla BASIC'de yazılan bilgisayar programlarıyla yapılmıştır.

Birimler arası kümeleme istatistiği z'nin, beklenen değeri E(z) ve varyansı Var(z)'nin hesaplanabilmesi için uzaklık ölçüsü dij matrisi ile değişken ve birim sayıları cinsinden uzaklıklar ölçüsü olarak adlandırılan t_{ij} matrisine gereksinim vardır. t_{ij} değerleri

$$t_{ij} = \begin{cases} \dfrac{2}{n_k(n_k-1)} & \text{eğer } i \in G_k \text{ ve } j \in G_k \quad (i \neq j) \\[2ex] \dfrac{-2}{n_k n_l m(m-1)} & \text{eğer } i \in G_k \text{ ve } j \in G_k \quad (k \neq l) \\[2ex] 0 & \text{eğer } \quad (i = j) \end{cases}$$

olarak hesaplanır. Burada n_k ve n_l değerleri her bir küme içindeki birim sayısı, m ise küme sayısıdır. Birinci durumda her bir kümenin kendi içindeki birimlerinin birbirleriyle olan uzaklık değerleri, ikinci durumda oluşturulan kümelerin birbirleriyle olan uzaklıkları, son durumda ise birimlerin birbirlerine eşit olduğu durumu vermektedir. Bu değerler hesaplandıktan sonra d_{ij} matrisi gibi t_{ij} matriside oluşturulmuştur.

6.2.2. g_2 ölçütü

Bu ölçütte, diğer iki ölçüte göre farklı olarak birimlerin uzaklık ölçüsü olarak Kare Öklid uzaklığı kullanılmıştır. Bu uzaklık ölçümü veriler standardize edilerek SPSS/PC+ programında hesaplanmıştır.

Kare Öklid uzaklığı kullanıldığı zaman g_2 ölçütü w_2 değerine indirgenmektedir. Formül (24)'den BASIC'de yazılan bir programla hesaplanmıştır.

g_2 ölçümünde t_{ij} değerleri

$$t_{ij} = \begin{cases} \dfrac{1}{n_k} & \text{eğer } i,j \in G_k \quad (i \neq j) \\ 0 & \text{diğer durumlarda} \end{cases}$$

olarak hesaplanmıştır.

6.2.3. g_3 ölçütü

Bir küme içindeki $\binom{n_k}{2}$ çiftlerinin sayısı ile ağırlıklı bir çiftin arasındaki uzaklığın ortalaması olarak tanımlanan g_3 ölçütü için 2 yöntemle

1. $g_{3a}=w_3-b_3$
2. $g_{3b}=w_3/b_3$

hesaplama yapılmıştır. Formül (30) ve (31)'den BASIC'de yazılan bir programla hesaplanmıştır.

Bu ölçüt için t_{ij} değerleri

$$t_{ij}=\begin{cases} \dfrac{1}{\sum\limits_{k=1}^{m}\binom{n_k}{2}} & \text{eğer } i\in G_k \text{ ve } j\in G_k \quad (i\neq j) \\[2em] \dfrac{-1}{\binom{N}{2}-\sum\limits_{k=1}^{m}\binom{n_k}{2}} & \text{eğer } i\in G_k \text{ ve } j\in G_k \quad (k\neq l) \\[2em] 0 & \text{eğer } \quad (i=j) \end{cases}$$

şeklinde hesaplanmıştır.

6.2.4. E(z), Var(z) ve z Test İstatistiği

Birimler arası kümeleme istatistiği z için beklenen değer E(z) ve varyans Var(z)

$$E(Z)=\frac{1}{\binom{N}{2}}\left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}t_{ij}\right]\left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}d_{ij}\right]$$

$$Var(Z)=\frac{1}{4N(N-1)}\left[2B_dB_t+\frac{4H_dH_t}{(N-2)}+\frac{K_dK_t}{(N-2)(N-3)}\right]-[E(Z)]^2$$

şeklinde hesaplanır. Burada

$$B_d=\sum_{i=1}^{N}\sum_{j=1}^{N}d_{ij}^2 \qquad B_t=\sum_{i=1}^{N}\sum_{j=1}^{N}t_{ij}^2$$

$$H_d=\sum_{i=1}^{N}\left(\sum_{j=1}^{N}d_{ij}\right)^2-B_d \qquad H_t=\sum_{i=1}^{N}\left(\sum_{j=1}^{N}t_{ij}\right)^2-B_t$$

$$K_d=G_d+2B_d-4\sum_{i=1}^{N}\left(\sum_{j=1}^{N}d_{ij}\right)^2 \qquad K_t=G_t+2B_t-4\sum_{i=1}^{N}\left(\sum_{j=1}^{N}t_{ij}\right)^2$$

şeklindedir. g_1, g_2 ve g_3 ölçütleri için hesaplanacak bu değerler BASIC'de yazılmış bir programla tüm veri setleri için hesaplatılmıştır. Sonradan sırası ile bu ölçümler

$$z(g_r) = \frac{g_r}{\sqrt{Var(g_r)}}$$

ile test edilmiştir.

6.2.5. Çokdeğişkenli varyans çözümlemesi istatistiği Wilk's Lamda

Rasgele ve koşullu verilerin küme sayılarına göre sırasıyla girdikleri kümelere göre SYSTAT paket programında MGLH modülünde, değişkenlere ve belirlenen küme gruplarına göre Wilk's Lamda değeri, F istatistiği ve olasılık değerleri hesaplatılmıştır.

6.2.6. Hotelling-Lawley İz ölçütü

Rasgele ve koşullu veriler, küme sayılarına göre sırasıyla girdikleri kümeler açısından SYSTAT programında MGLH modülünde değişkenlere ve birim sayılarına göre Hotelling-Lawley İz değeri, F istatistiği ve olasılık değerleri hesaplanmıştır.

6.2.7. Ayırma ölçütü

Birimlerin girdikleri her küme sayısına göre belirlenen dağılımları için Minitab paket programında DSCRIMINANT komutu ile daha önce K-Ortalamalar yöntemi ile girdikleri sınıfların numaraları grup numarası alınarak belirlendikten sonra birimlerin bu gruplarda doğru sınıflanrırılma olasılıkları hesaplanmıştır.

6.2.8. Kofenetik korelasyon katsayısı

Tüm birimlerin standardize değerleri hesaplatılarak SPSS/PC paket programında CLUSTER modülünde Öklid uzaklığına göre TBK yöntemi ile bağlantılar elde edilmiş ve bu bağlantıları içeren bağlantı katsayıları, C_{ij}'leri iceren C bağlantı katsayıları matrisi çıkarılmıştır. Benzerlik matrisi S ile katsayılar matrisi C'nin i ve j değerleri aynı olan karşılıklı elamanları arasındaki ilişki basit korelasyon çözümlemesi ile test edilmiştir. r_{cs} olarak ifade edilen kofenetik korelasyon katsayısı Pearson korelasyon katsayısı ile aynı istatistik özellikler taşımaktadır.

7. BULGULAR

7.1. Rasgele Gruplara İlişkin Bulgular

Aşamalı olmayan kümeleme yöntemine göre kümelenen R10, R20, R30, R40 ve R50 gruplarının elemanları olan birimlerin küme sayısı 2, 3, 4 ve 5 olarak seçildiği durumlarda oluşturdukları kümeler Çizelge 7.1'de verilmiştir.

Rasgele türetilen farklı sayıdaki birimlerin oluşturdukları kümeler; küme sayısı arttırıldığında, birbirine çok benzer olan birimlerin oluşturdukları kümeler değişmeden kalırken, yakın benzerlikte olan birimlerin oluşturduğu kümeler dağılma göstermektedir. Bu durum tüm gruplarda gözlenmiştir. Küme sayısı küçülürken çok benzer, benzer ve farklı özellikteki birimlerin bir araya geldikleri, küme sayısı artarken ise türdeşliğin enbüyüklendiği yeni kümelere atanmaların ortaya çıktığı gözlenmiştir. Bu bilgi veri matrisindeki birimlerin birbirleri ile olan benzerlik ve uzaklıkları hakkında ayrıntılı bilgi edinmeyi sağlamaktadır. Veri matrisinin ayrıntılı incelenmesinde araştırıcıya hangi birimler üzerinde daha ayrıntılı durması ve bilinmeyen yapıların çözümlenmesi konusunda hangi birimlerin daha çok yararlı bilgi verebileceğini saptama olanağı vermektedir.

Çizelge 7.1. Rasgele türetilen birimlerin küme sayısına göre oluşturdukları kümeler ve bu kümelerde yer alan birim noları

Grup Adı	Birim Sayısı (n)	Küme Sayısı (m)	Küme Elemanları
R10	10	2	(1,4,5,9,10)(2,3,6,7,8)
		3	(2,5,9)(4,10,6,1)(3,7,8)
		4	(2,5,9)(4,10)(6,1)(3,7,8)
		5	(2,9)(5)(4,10)(6,1)(3,7,8)
R20	20	2	(2,3,5,7,8,9,11,12,14,17,20)(1,6,10,13,15,16,18,19)
		3	(2,9,11,12,14,17)(1,4,6,10,13,16,18,19) (3,5,7,8,20)
		4	(2,9,11,12,14,17)(4,10,13,16,18,19) (3,5,7,8)(1,6,15)
		5	(11,14,17)(4,10,13,16,18,19) (3,5,7,8,20) (1,6,15)(2,9,12)
R30	30	2	(5,6,7,10,11,13,14,17,19,20,26,27,28,8) (1,2,3,4,9,12,15,16,18,21,22-25,29,30)
		3	(5,6,7,10,11,13,14,17,19,20,26,27,28) (1,2,8,9,12,18,22)(3,4,15,16,21,23,24,25,29,30)
		4	(11,17,19,26,27,28)(1,2,8,9,12,18,22) (3,4,15,16,21,23,24,25,29,30)(5,6,7,10,13,14,20)
		5	(11,17,19,26,27,28)(1,2,8,9,12,18,22) (4,15,16,24,30)(5,6,7,10,13,14,20)(3,21,23,25,29)
R40	40	2	(2-5,8,9,11,12,14,17,24,25,28-32,34,35,38,39) (1,6,7,19,13,15,16,18,19-23,26,27,33,36,37,40)
		3	(2,3,8,9,17,24,25,28,30,32,35,38) (1,7,10,13,15,16,18,19,20,21,22,23,26,27,33,36,37,40) (4,5,6,11,12,14,29,31,34,39)
		4	(2,3,8,9,17,24,25,28,30,32,35,38) (1,7,10,13,16,18,22,23,27,36,40) (4,5,11,12,14,29,31,34,39)(6,15,19,20,21,26,33,37)
		5	(3,9,24,38)(1,7,10,13,16,18,22,23,27,36,40) (4,5,11,12,14,29,31,34,39)(6,15,19,20,21,26,33,37) (2,8,17,25,28,30,32,35)
R50	50	2	(2-4,6,7,12,13,15,18,22,24,25,28,31,33,34,36,37,39,40,42,45) (5,8-11,14,16,17,19-23,26,27,29,30,32,35,38,41,43,44,46-50)
		3	(2-4,6,7,12,13,15,18,22,24,25,28,31,33,34,36,37,39,40,42,45) (1,5,8,17,21,23,29,30,35,44,46,47,49,50) (9,10,11,14,16,19,20,26,27,32,38,41,43,48)
		4	(4,7,13,15,18,28,31,33,39,40) (1,5,8,17,21,23,29,30,35,44,46,47,49,50) (9,10,11,14,16,19,20,26,27,32,38,41,43,48) (2,3,6,12,22,24,25,34,36,37,42,45)
		5	(4,7,13,15,18,28,31,39,40) (1,5,8,17,21,23,29,30,35,44,46,47,49,50) (9,10,11,14,16,32,48) (3,6,12,22,24,25,33,36,37,42,45) (2,19,20,26,27,34,38,41,43)

Rastgele grup birimlerin Öklid uzaklığına göre hazırlanmış benzerlik matrisinden TBK yöntemiyle aşamalı bağlantılarını gösteren ağaç grafikleri sırası ile R10 grubu için Şekil 7.1'de, R20 grubu için Şekil 7.2'de, R30 grubu için Şekil 7.3'de, R40 grubu için Şekil 7.4'de ve R50 grubu için Şekil 7.5'de gösterilmiştir.

Bağlantı Katsayıları

```
           Birimler
                          1
           1  4  7  2  9  0  8  3  6  5

0.782      -+-  I  I  I  I  I  I  I  I
2.149       I  -+-  I  I  I  I  I  I
2.485       --+--   I  I  I  I  I  I
2.530         I     I  I  I  I  -+-
2.613         -+----  I  I  I   I
2.641          I      I  -+-    I
2.702          -+------   I     I
2.719            --+-------     I
2.730               --+---------
```

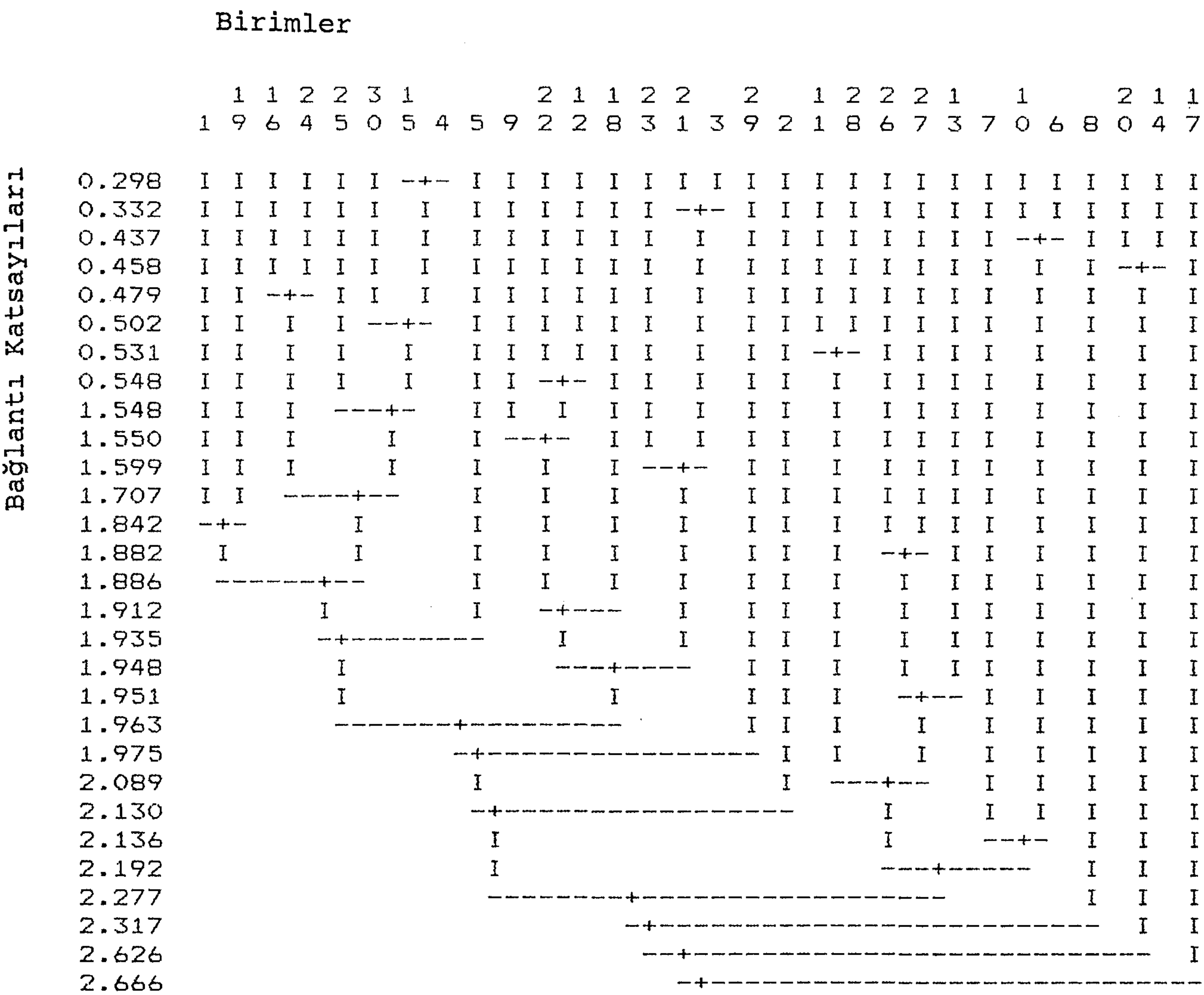
Şekil 7.1. TBK ve Öklid Uzaklığa Göre R10 Grubu Ağaç Grafiği

Birimler

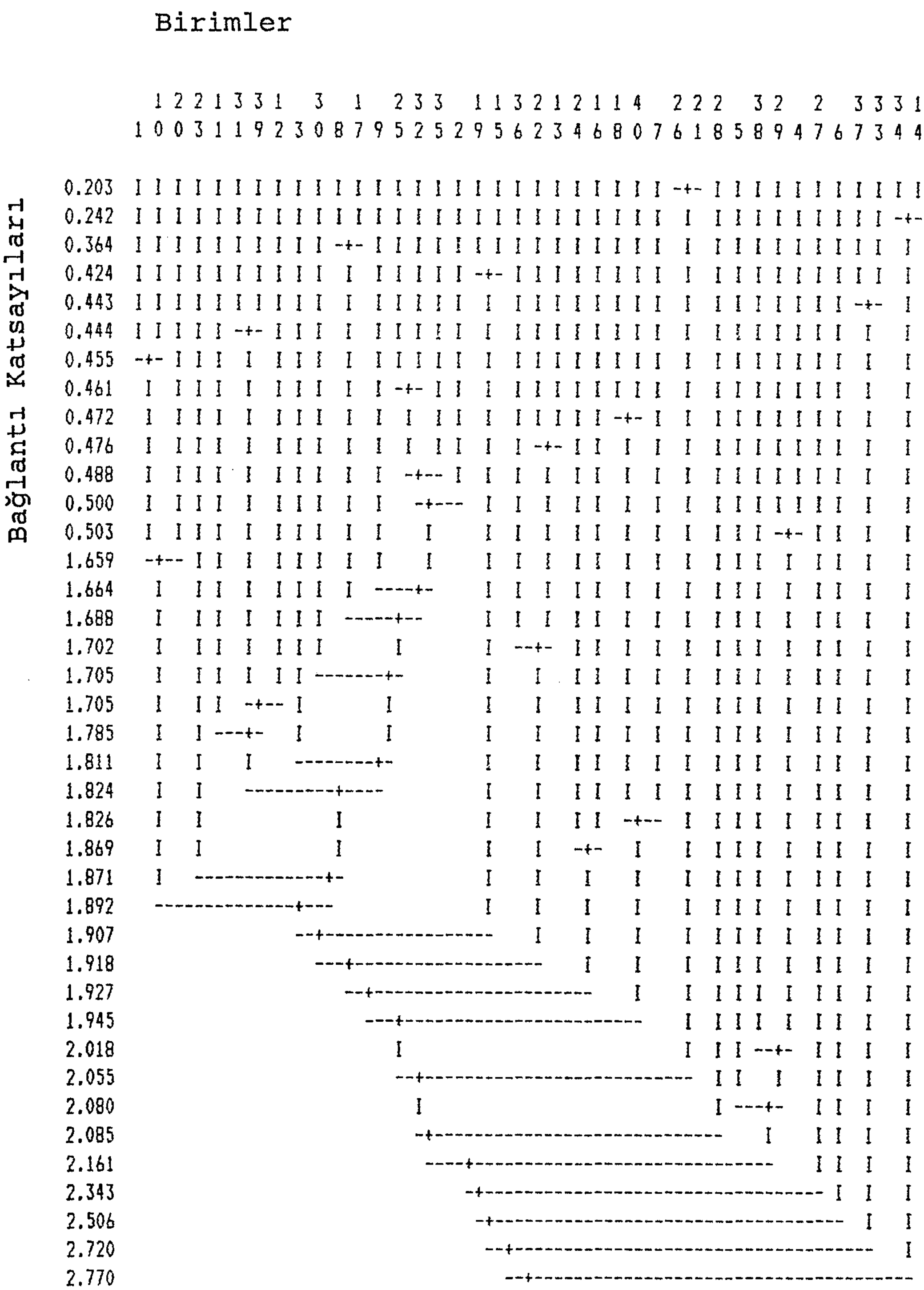
Bağlantı Katsayıları

```
            2 1 1 1 1                   1 1 1 1 1 1
        1 6 0 1 2 9 4 9 2 8 3 5 7 0 6 8 3 4 7 5

0.487   I I I I I I I I I I I I I I I -+- I I I
1.537   I I I I I I I I I I I I I I --+-  I I I
1.886   I I I I -+- I I I I I I I I   I   I I I
1.955   I I I I  I  I I I I I I I ---+-   I I I
1.986   -+- I I  I  I I I I I I I    I    I I I
2.010    I  I I  I  I -+- I I I I    I    I I I
2.033    I  I I  -+--  I  I I I I    I    I I I
2.046    I  I I   I    I  I I I I     -+---- I I
2.066    -+-- I   I    I  I I I I     I     I I
2.087     I   I   --+---  I I I I     I     I I
2.103     I   I     I     I I I -----+-     I I
2.104     I   -----+-     I I I      I      I I
2.180     I        I      I I -------+-     I I
2.190     ------+---      I I       I       I I
2.335           I         -+-       I       I I
2.357           --+---------        I       I I
2.448             -------+-----------       I I
2.658                    -+------------------ I
2.730                     -+-------------------
```

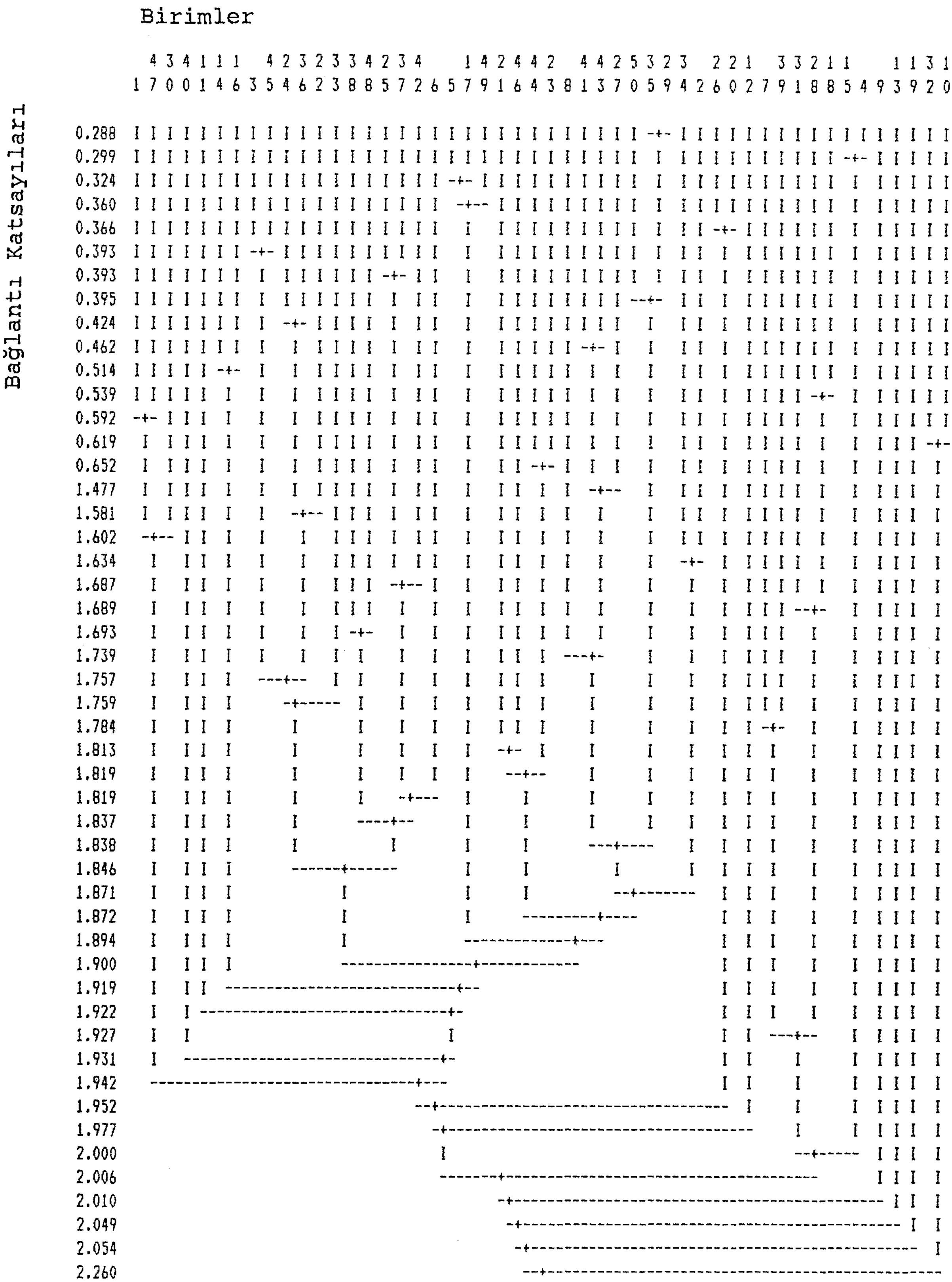
Şekil 7.2. TBK ve Oklid Uzaklığa Göre R20 Grubu Ağaç Grafiği



Şekil 7.3. TBK ve Oklid Uzaklığa Göre R30 Grubu Ağaç Grafiği



Şekil 7.4. TBK ve Oklid Uzaklığa Göre R40 Grubu Ağaç Grafiği



Şekil 7.5. TBK ve Öklid uzaklığa göre R50 grubu ağaç grafiği

Rasgele grupların ağaç grafikleri Şekil 1-5 incelenerek 2 birim bağlantı uzaklığına göre oluşan kümeler ve küme elamanları Çizelge 7.2'de verilmiştir.

Çizelge 7.2'deki 2 birim uzaklığa göre;

R10 grubu 10 ayrı özellikte birim olarak ortaya çıkmıştır.

R20 grubunda 13 küme oluşmuş, bunlardan 10 birim tek başına küme özelliği göstermiştir.

R30 grubunda 8 küme oluşmuş bunlardan 3 tanesinin eleman sayısı 1 birim olmuştur. En büyük elemanlı küme ise 17 birim içermiştir.

R40 grubunda 14 küme oluşmuş ve bunlardan 5 tanesi 1 birimlik kümeler olmuştur.

R50 grubunda 8 küme oluşmuş bunlardan 4 tanesi tek elamanlıdır. Enbüyük küme 37 elamanlı bir kümedir.

Bu ayırım, çokdeğişkenli yapılarda birim sayısı arttıkca kümeleme çözümlemesinin benzer birimleri bulma ve aynı sınıfta gösterme yeteneğinin arttığını ve farklı özellikteki birimlerin ise tek başlarına ya da yakın benzerlikte olan birimlerle aynı kümeye girmelerini sağlamaktadır.

Çizelge 7.2. Rasgele gruplarda TBK yönteminde 2 birim uzaklığa göre oluşan kümeler ve bu küme elemanları

Grup Adı	Birim Sayısı (n)	Küme Sayısı (m)	Küme Elemanları
R10	10	9	(1-4) (2) (3) (5) (6) (7) (8) (9) (10)
R20	20	13	(10,18,16,13,4) (1,6) (12,19) (20)(11)(14)(9)(2)(8)(3)(15)(7)(5)
R30	30	8	(1,9,16,24,25,30,15,4,5,9,22,12,18,23,21,3,29)(26,2,13)(10,6)(20,14)(11,28)(7)(8)(17)
R40	40	14	(1,2,10,20,23,11,31,39,12,3,30,8,17,9,25,32,35)(18,40,7)(36,22,13)(24,16)(19,15)(26,12)(29,34)(37,33)(34,14)(28)(5)(38)(27)(6)
R50	50	8	(1,47,30,40,11,14,16,3,45,24,36,22,33,38,48,25,37,42,6,5,17,49,21,46,44,23,8,41,43,27,50,35,29,34,2,26,20)(7,39,31,28,18)(15,4)10,32)(9)(13)(19)(12)

7.2. Koşullu Verilere İlişkin Bulgular

Koşullu olarak türetilen 10 birimlik verilerin rasgele türetilen R10, R20, R30, R40 ve R50 gruplarına eklenmesiyle elde edilen K20, K30, K40, K50 ve K60 grupları aşamalı olmayan kümeleme yöntemiyle kümelenerek küme sayısı 2 ile 5 arasında seçildiği durumlarda oluşturdukları kümeler Çizelge 7.3'de verilmiştir.

Koşullu olarak türetilen farklı sayıdaki birimlerin oluşturdukları kümelerde, rasgele türetilen R gruplarından gelen birimlerin tüm K gruplarında, küme sayısı ne olursa olsun aynı kümede yer aldıkları gözlenmiştir. Eklenen 10 birimlik veriler, küme sayısına göre kendi içinde parçalanmıştır. Rasgele verilerde olduğu gibi küme sayısı arttırıldığında birbirine çok benzer olan birimlerin oluşturdukları kümeler değişmeden kalırken, yakın benzerlikte olan birimlerin oluşturduğu kümeler dağılma göstermekte ve çok benzer birimlerin oluşturduğu alt kümelere parçalanmaktadır. Tüm gruplarda eklenen birimler küme sayısına göre aynı şekilde parçalanmaktadır. Koşullu verilerde de küme sayısı küçülürken; çok benzer, benzer ve farklı özellikteki birimlerin bir araya geldikleri, küme sayısı artarken ise türdeşliğin enbüyüklendiği yeni kümelere atanma sağlanmaktadır. Bu bilgi, veri matrisindeki birimlerin birbirleri ile olan benzerlik ve uzaklıkları hakkında ayrıntılı bilgi edinmeyi sağlamaktadır.

Çizelge 7.3. Koşullu türetilen birimlerin küme sayısına göre oluşturdukları kümeler ve bu kümelerde yer alan birim noları

Grup Adı	Birim Sayısı (n)	Küme Sayısı (m)	Küme Elemanları
K20	20	2	(1-10) (11-20)
		3	(1-10) (11) (12-20)
		4	(1-10) (11) (12,13,15,16,19,20) (1,4,17,18)
		5	(1-10) (11) (12,13,15,19,20) (1,4,17,18) (16)
K30	30	2	(1-20) (21-30)
		3	(1-20) (21,22,23,25,26,27)(24,28,29,30)
		4	(1-20) (21,22,23,25,26,27)(24,28,29)(30)
		5	(1-20) (21,22,26,29) (23,25,27) (24,28) (30)
K40	40	2	(1-30) (31-40)
		3	(1-30) (31,32,33,35,36,37) (34,38,39,40)
		4	(1-30) (31,32,33,35,36,37) (34,38,39) (40)
		5	(1-30) (31,32,36,39) (33,35,37) (34,38) (40)
K50	50	2	(1-40) (41-50)
		3	(1-40) (41,42,43,45,46,47) (44,48,49,50)
		4	(1-40) (41,42,43,45,46,47) (44,48,49) (50)
		5	(1-40) (41,42,46,49) (43,45,47) (44,48) (50)
K60	60	2	(1-50) (51-60)
		3	(1-50) (51,52,53,55,56,57) (54,58,59,60)
		4	(1-50) (51,52,53,55,56,57) (54,58,59) (60)
		5	(1-50) (51,52,56,59) (53,55,57) (54,58) (60)

Koşullu grup birimlerinin Öklid uzaklığına göre hazırlanmış benzerlik matrisinden TBK yöntemiyle aşamalı bağlantılarını gösteren ağaç grafikleri sırası ile K20 grubu için Şekil 7.6'da, K30 grubu için Şekil 7.7'de, K40 grubu için Şekil 7.8'de, K50 grubu için Şekil 7.9'da ve K60 grubu için Şekil 7.10'da gösterilmiştir.

Birimler

Bağlantı Katsayıları

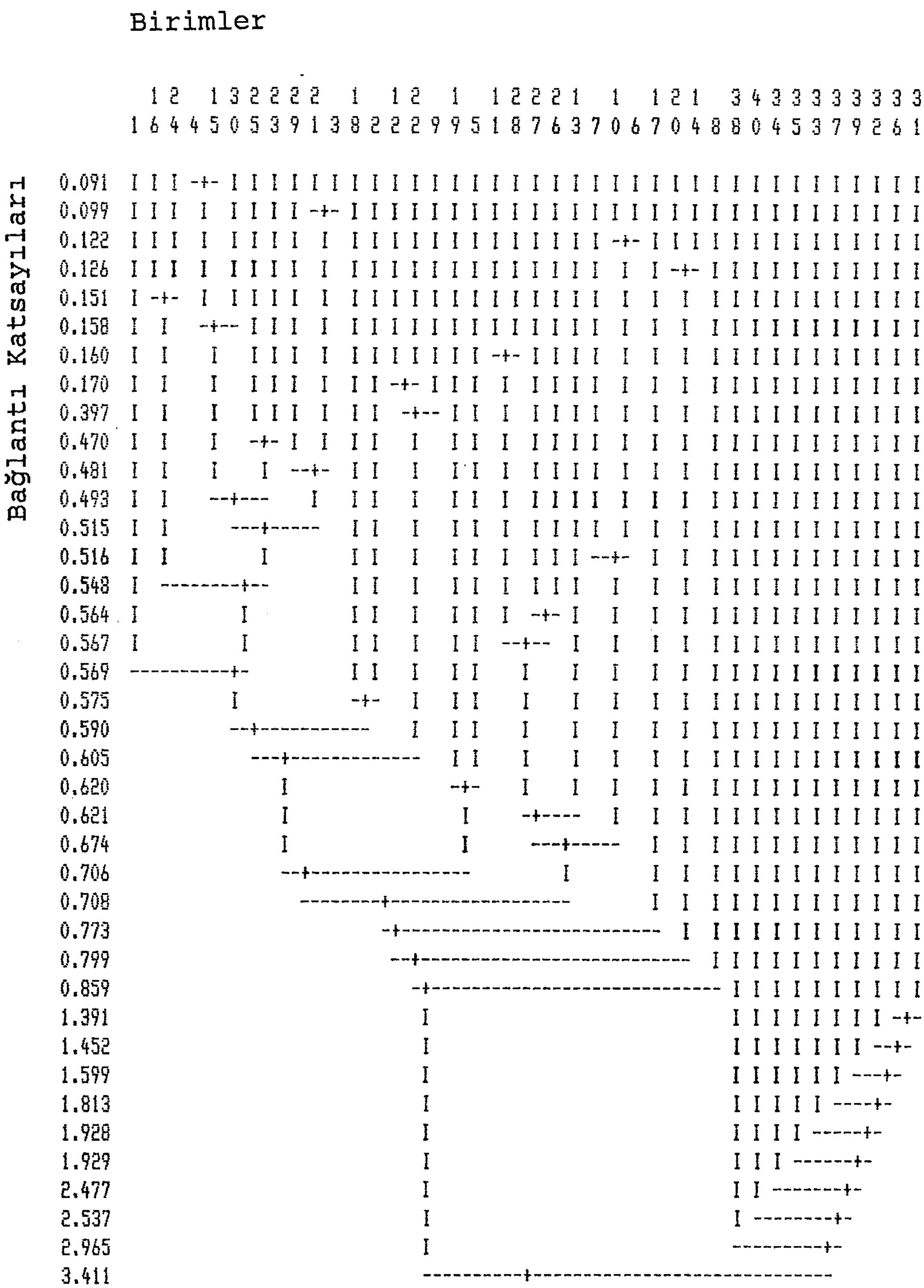
```
           1               1 1 2 1 1 1 1 1 1 1
       1 6 0 4 3 8 5 7 9 2 8 4 0 5 3 7 9 2 6 1

0.450  I I I I I I I I -+- I I I I I I I I I I
0.493  I I I I -+- I I  I  I I I I I I I I I I
0.532  -+- I I  I  I I  I  I I I I I I I I I I
0.544   I  -+-  I  I I  I  I I I I I I I I I I
0.664   I   I   -+-- I  I  I I I I I I I I I I
0.690   I   I    -+---  I  I I I I I I I I I I
0.692   --+--     I     I  I I I I I I I I I I
0.726     I       --+----  I I I I I I I I I I
0.772     ------+----      I I I I I I I I I I
1.163           I          I I I I I I I I -+-
1.249           I          I I I I I I I --+-
1.318           I          I I I I I I ---+-
1.523           I          I I I I I ----+-
1.605           I          I I I I -----+-
1.668           I          I I I ------+-
2.087           I          I I -------+-
2.140           I          I --------+-
2.444           I          ---------+-
3.086           ----------+----------
```

Sekil 7.6. TBK ve Oklid uzakliga gore K20 grubu agac grafigi

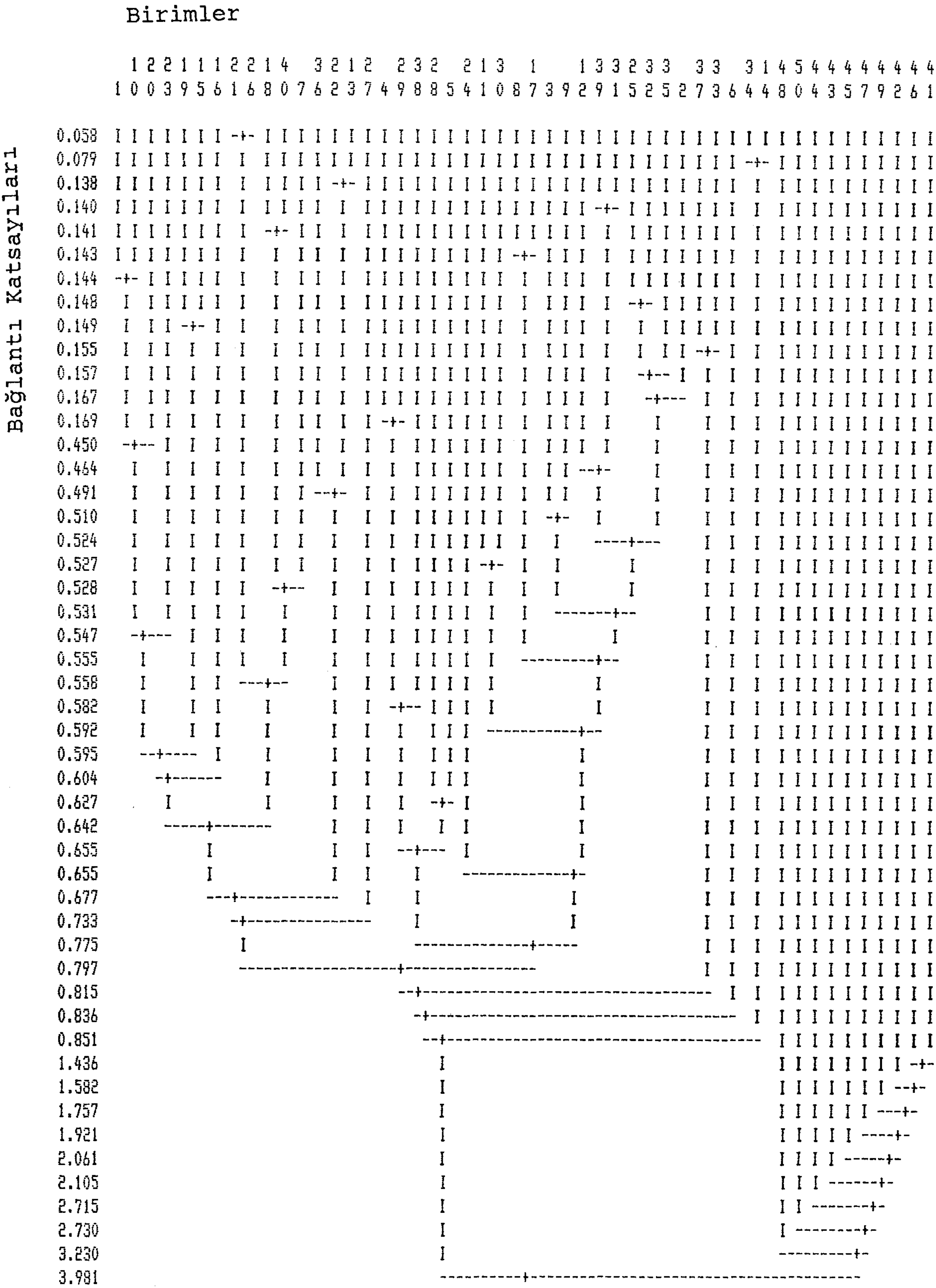
```
       Birimler

         1   1 1 1 1 1       1     1 1 2 1     2 2 3 2 2 2 2 2 2 2
       1 9 6 0 6 8 3 5 4 5 7 7 3 8 2 4 0 1 9 2 8 4 0 5 3 7 9 2 6 1

0.137  I I I I I -+- I I I I I I I I I I I I I I I I I I I I I I I
0.440  I I I I --+-  I I I I I I I I I I I I I I I I I I I I I I I
0.478  I I I ---+-   I I I I I I I I I I I I I I I I I I I I I I I
0.481  I I I    I    I I I I I I I -+- I I I I I I I I I I I I I I
0.486  I I I    I    I I I I I I I  I  I I -+- I I I I I I I I I I
0.489  I I I    I    I I I I I I I  I  I --+-  I I I I I I I I I I
0.503  I I I    -+---- I I I I I I  I  I   I   I I I I I I I I I I
0.505  I I I     I     I I I I I I  -+--   I   I I I I I I I I I I
0.518  I -+-     I     I I I I I I   I     I   I I I I I I I I I I
0.545  I  I      I     I I I I -+-   I     I   I I I I I I I I I I
0.572  --+-      I     I I I I  I    I     I   I I I I I I I I I I
0.586    I       -+----- I I I  I    I     I   I I I I I I I I I I
0.591    I        I      I I I  I    ---+---   I I I I I I I I I I
0.651    ------+---      I I I  I       I      I I I I I I I I I I
0.660          I         I I I  ------+--      I I I I I I I I I I
0.673          I         I -+-        I        I I I I I I I I I I
0.692          I         I  --------+--        I I I I I I I I I I
0.699          I         ----------+-          I I I I I I I I I I
0.704          -----------+---------           I I I I I I I I I I
1.256                     I                    I I I I I I I I -+-
1.361                     I                    I I I I I I I --+-
1.427                     I                    I I I I I I ---+-
1.661                     I                    I I I I I ----+-
1.755                     I                    I I I I -----+-
1.800                     I                    I I I ------+-
2.295                     I                    I I -------+-
2.316                     I                    I --------+-
2.703                     I                    ---------+-
3.270                     ----------+--------------------
```

Bağlantı Katsayıları

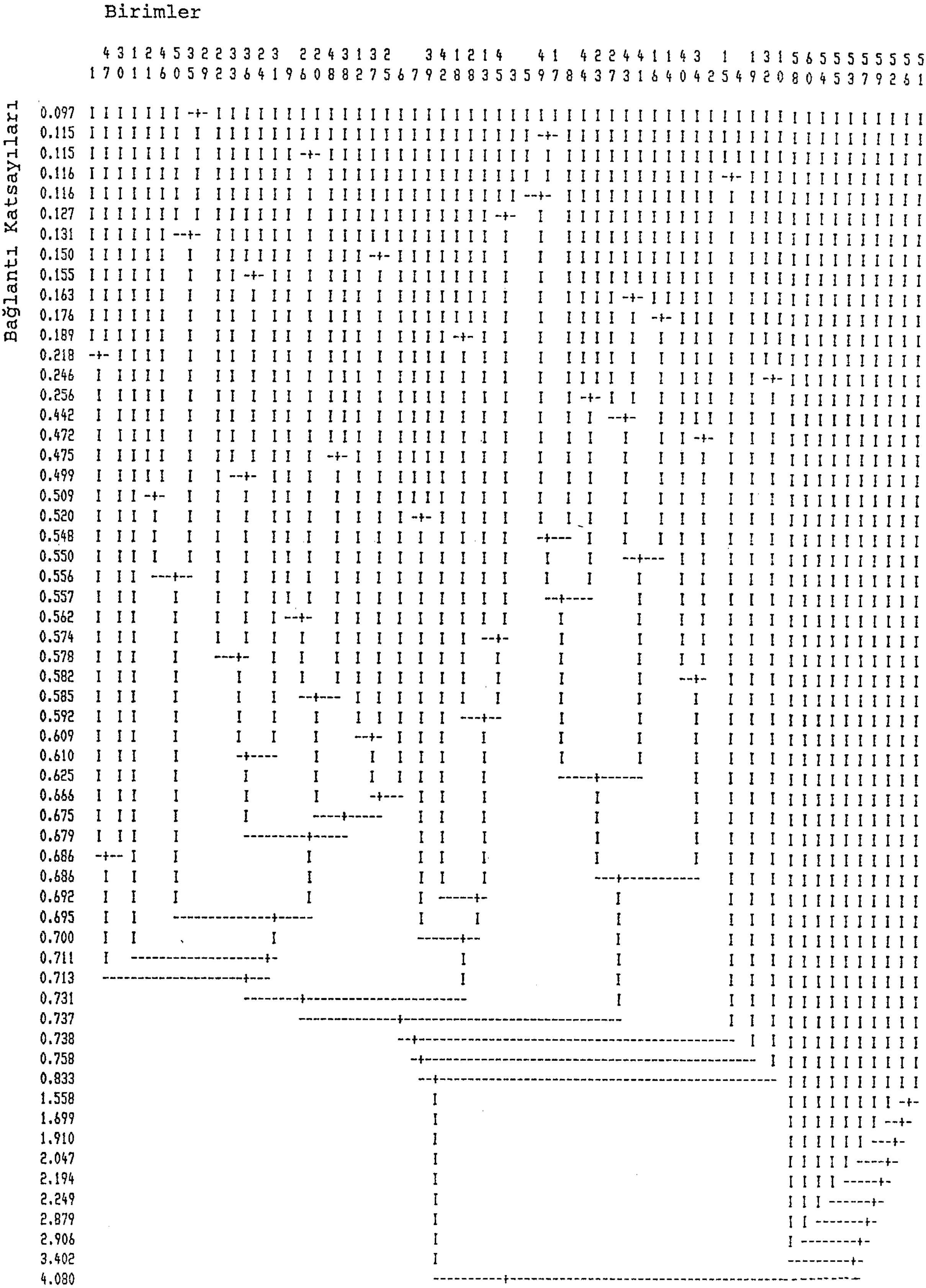
Şekil 7.7. TBK ve Öklid uzaklığa göre K30 grubu ağaç grafiği



Şekil 7.8. TBK ve Öklid uzaklığa göre K40 grubu ağaç grafiği



Şekil 7.9. TBK ve Öklid uzaklığa göre K50 grubu ağaç grafiği


Birimler
Bağlantı Katsayıları
0.097
0.115
0.115
0.116
0.116
0.127
0.131
0.150
0.155
0.163
0.176
0.189
0.218
0.246
0.256
0.442
0.472
0.475
0.499
0.509
0.520
0.548
0.550
0.556
0.557
0.562
0.574
0.578
0.582
0.585
0.592
0.609
0.610
0.625
0.666
0.675
0.679
0.686
0.686
0.692
0.695
0.700
0.711
0.713
0.731
0.737
0.738
0.758
0.833
1.558
1.699
1.910
2.047
2.194
2.249
2.879
2.906
3.402
4.080

Koşullu grupların ağaç grafikleri Şekil 6-10 incelenerek 2 birim bağlantı uzaklığına göre oluşan kümeler ve küme elamanları Çizelge 7.4'de verilmiştir.

Çizelge 7.4'deki 2 birim uzaklığa göre;

K20 grubu 2 ayrı küme özelliği göstermektedir. R10 grubundan gelen birimler tek bir kümede toplanmış, koşullu türetilen birimlerde başka bir kümede toplanmıştır.

K30 grubunda 3 küme oluşmuş, bunlardan R20 grubundan gelen birimler bir kümede toplanmış, koşullu türetilen birimler ise iki ayrı küme oluşturmuştur.

K40 grubu 3 ayrı küme özelliği göstermektedir. R30 grubundan gelen birimler tek bir kümede toplanmış, koşullu türetilen birimlerde iki ayrı kümede toplanmıştır.

K50 grubu 3 ayrı küme özelliği göstermektedir. R40 grubundan gelen birimler tek bir kümede toplanmış, koşullu türetilen birimlerde iki ayrı kümede toplanmıştır.

K60 grubu 3 ayrı küme özelliği göstermektedir. R50 grubundan gelen birimler tek bir kümede toplanmış, koşullu türetilen birimlerde iki ayrı kümede toplanmıştır.

Koşullu gruplarda TBK yöntemi ile yapılan kümelemenin 2 birim bağlantı uzaklıklarına göre oluşturdukları kümeler Çizelge 7.3'de verilen K-ortalamalar yöntemine göre elde edilen kümelemelerle benzerlik göstermektedir.

Çizelge 7.4. Koşullu gruplarda TBK yönteminde 2 birim uzaklığa göre oluşan kümeler ve bu kümelerde yer alan birim noları

Grup Adı	Birim Sayısı (n)	Küme Sayısı (m)	Küme Elemanları
K20	20	2	(1-10) (11-20)
K30	30	3	(1-20) (28,24) (21,22,23,25,26,27,29,30)
K40	40	3	(1-30) (38,40) (34,35,33,37,39,32,36,31)
K50	50	3	(1-40) (48,50,44,43) (45,47,49,42,46,41)
K60	60	3	(1-50) (58,60,54,55,53) (57,59,52,56,51)

7.3. g_1 Ölçütü Bulguları

7.3.1. $g_{1a}=w_1-b_1$ değerlerine ilişkin bulgular

Rasgele gruplar:

R10, R20, R30, R40 ve R50 gruplarının küme sayısı 2-5 olarak seçildiği durumlarda oluşturdukları kümelere göre w_1, b_1, g_{1a}, var(g_{1a}) ve g_{1a}'nın önemliliğini test eden z test istatistiği değerleri Çizelge 7.5' de verilmiştir.

R10 grubunda küme belirleme ölçütü g_{1a}, küme sayısı 2-5 durumlarında birimlerin kümelemelere atanmalarını önemli bir kümelenme olarak vermemiştir.

R20 grubunda rasgele verilerin 4 kümeye ayrılmasının önemli bir parçalanma olduğu belirlenmiştir ($P<0.001$)

R30 grubunda birimlerin 5 kümeye ayrılmalarının önemli bir parçalanma olduğu ortaya çıkmıştır ($P<0.001$)

R40 grubunda birimlerin 3 kümeye ayrılması önemli parçalanma olarak belirlenmiştir ($P<0.001$)

R50 grubunda 5 gruba birimlerin ayrılması önemli bulunmuştur ($P<0.001$).

Tüm rasgele olarak türetilen verilerde önemli bir kümelenme görülmemesi gerekirken g_{1a} ölçütü gruptaki birim sayısı 20 ve daha fazla olduğunda 0.05 yanılma düzeyinde önemli kümelenme belirtisi vermiştir ($P<0.05$).

Çizelge 7.5. Rasgele gruplarda g_{1a} ölçütü değerleri

Grup adı	Küme sayısı	w_1	b_1	g_{1a}	var(g_{1a})	z test ist.
R10	2	3.38	3.38	0	0.551	-
	3	2.95	3.53	-0.58	0.280	$-1.10^{n.s}$
	4	2.40	3.58	-1.18	0.897	$-1.25^{n.s}$
	5	1.71	3.11	-1.40	1.577	$-1.12^{n.s}$
R20	2	3.06	3.62	-0.56	0.109	$-1.70^{n.s}$
	3	2.88	3.59	-0.75	0.108	-2.16^{*}
	4	2.75	3.59	-0.84	0.047	-3.87^{***}
	5	2.66	3.58	-0.92	0.071	-3.45^{***}
R30	2	2.95	3.72	-0.77	0.312	$-1.38^{n.s}$
	3	2.55	3.60	-1.05	0.219	-2.24^{*}
	4	2.62	3.62	-1.00	0.094	-3.26^{**}
	5	2.30	3.52	-1.22	0.010	-12.2^{***}
R40	2	2.26	3.56	-1.30	0.156	-3.29^{***}
	3	2.83	3.60	-0.77	0.006	-9.94^{***}
	4	2.72	3.59	-0.87	0.032	-4.86^{***}
	5	2.55	3.56	-1.01	0.239	-2.07^{*}
R50	2	2.75	3.68	-0.93	0.125	-2.63^{**}
	3	2.80	3.57	-0.77	0.056	-3.27^{***}
	4	2.65	3.61	-0.96	0.292	$-1.78^{n.s}$
	5	2.52	3.60	-1.08	0.004	-17.08^{***}

n.s : $P>0.05$
* : $P<0.05$
** : $P<0.01$
*** : $P<0.001$

Koşullu gruplar:

K20, K30, K40, K50 ve K60 gruplarının elemanları olan birimlerin küme sayısı 2-5 olarak seçildiği durumlarda oluşturdukları kümelere göre w_1, b_1, g_{1a}, var(g_{1a}) ve g_{1a}'nın önemliliğini test eden z test istatistiği değerleri Çizelge 7.6' da verilmiştir.

K20 grubunda küme belirleme ölçütü g_{1a}, küme sayısı 2-5 durumlarında birimlerin kümelere atanmalarını önemli bir kümelenme olarak vermiştir (P<0.01) ve parçalanmanın küme içi türdeşliği enbüyükleyen küme sayısının 4 olduğunu göstermiştir (P<0.001).

K30 grubunda g_{1a}, küme sayısı 2-5 durumlarında birimlerin kümelere atanmalarını önemli bir kümelenme olarak vermiştir (P<0.001) ve parçalanmanın küme içi türdeşliği enbüyükleyen küme sayısının 4 olduğunu göstermiştir (P<0.001).

K40 grubunda g_{1a}, küme sayısı 2-5 durumlarında birimlerin kümelere atanmalarını önemli bir kümelenme olarak vermiştir (P<0.001) ve parçalanmanın küme içi türdeşliği enbüyükleyen küme sayısının 3 olduğunu göstermiştir (P<0.001).

K50 grubunda küme belirleme ölçütü g_{1a}, küme sayısı 2-5 durumlarında birimlerin kümelere atanmalarını önemli bir kümelenme olarak vermiştir (P<0.001) ve parçalanmanın küme içi türdeşliği enbüyükleyen küme sayısının 3 olduğunu göstermiştir (P<0.001).

Çizelge 7.6. Koşullu gruplarda g_{1a} ölçütü değerleri

Grup adı	Küme sayısı	w_1	b_1	g_{1a}	var(g_{1a})	z test ist.
K20	2	2.96	4.53	-1.57	0.118	-4.57***
	3	0.60	3.49	-2.89	1.190	-2.65**
	4	0.39	3.13	-2.74	0.202	-6.10***
	5	0.28	2.90	-2.62	0.229	-5.47***
K30	2	1.73	4.81	-3.08	1.089	-2.95**
	3	1.90	4.11	-2.21	0.065	-8.67***
	4	0.49	3.64	-3.15	0.086	-10.74***
	5	0.32	3.50	-3.18	0.172	-7.67***
K40	2	1.89	5.27	-3.38	0.356	-5.66***
	3	2.06	4.49	-2.43	0.059	-10.00***
	4	0.54	3.98	-3.44	0.817	-3.81**
	5	0.34	3.83	-3.49	0.729	-4.09***
K50	2	2.05	5.69	-3.64	0.609	-4.66***
	3	2.22	4.86	-2.64	0.048	-12.05***
	4	0.58	4.40	-3.82	1.003	-3.81***
	5	0.37	4.14	-3.77	0.574	-4.98***
K60	2	1.71	6.07	-4.36	1.072	-4.21***
	3	2.38	5.20	-2.82	0.232	-5.85***
	4	0.61	4.60	-3.99	0.711	-4.73***
	5	0.40	4.43	-4.03	0.978	-4.08***

n.s : $P>0.05$
* : $P<0.05$
** : $P<0.01$
*** : $P<0.001$

K60 grubunda küme belirleme ölçütü g_{1a}, küme sayısı 2-5 durumlarında birimlerin kümelere atanmalarını önemli bir kümelenme olarak vermiştir ($P<0.001$) ve parçalanmanın küme içi türdeşliği enbüyükleyen küme sayısının 3 olduğunu göstermiştir ($P<0.001$).

İki farklı yapıda çok değişkenli normal dağılımlardan çekilmiş örnekleri içeren koşullu verilerde g_{1a} ölçütü, tüm gruplarda ve tüm küme sayılarında önemli kümelenmeyi işaret etmektedir ($P<0.01$).

7.3.2. $g_{1b}=w_1/b_1$ değerlerine ilişkin bulgular

Rasgele gruplar:

R10, R20, R30, R40 ve R50 gruplarının elemanları olan birimlerin küme sayısı 3, 4 ve 5 olarak seçildiği durumlarda oluşturdukları kümelere göre w_1, b_1, g_{1b}, var(g_{1b}) ve g_{1b}'nin önemliliğini test eden z test istatistiği değerleri Çizelge 7.7'de verilmiştir.

R10 grubunda küme belirleme ölçütü g_{1b}, küme sayısı 2-5 durumlarında birimlerin kümelemelere atanmalarını önemli bir kümelenme olarak vermemiştir ($P>0.05$).

R20 grubunda rasgele verilerin 4 kümeye ayrılmasının önemli bir parçalanma olduğu belirlenmiştir ($P<0.001$).

R30 grubunda birimlerin 5 kümeye ayrılmalarının önemli bir parçalanma olduğu ortaya çıkmıştır ($P<0.001$).

Çizelge 7.7. Rasgele gruplarda g_{1b} ölçütü değerleri

Grup adı	Küme sayısı	w_1	b_1	g_{1b}	var(g_{1b})	z test ist.
R10	2	3.38	3.38	1	0.551	$1.35^{n.s}$
	3	2.95	3.58	0.84	0.280	$1.56^{n.s}$
	4	2.40	3.58	0.67	0.897	$0.71^{n.s}$
	5	1.71	3.11	0.55	1.577	$0.44^{n.s}$
R20	2	3.06	3.62	0.85	0.109	2.56^{**}
	3	2.88	3.59	0.80	0.108	2.44^{*}
	4	2.75	3.59	0.77	0.047	3.53^{***}
	5	2.66	3.58	0.74	0.071	2.79^{**}
R30	2	2.95	3.72	0.79	0.312	$1.42^{n.s}$
	3	2.55	3.60	0.71	0.219	1.51^{*}
	4	2.62	3.62	0.72	0.094	2.36^{*}
	5	2.30	3.52	0.65	0.010	6.53^{***}
R40	2	2.26	3.56	0.63	0.156	$1.61^{n.s}$
	3	2.83	3.60	0.79	0.006	10.15^{***}
	4	2.72	3.59	0.76	0.032	4.24^{***}
	5	2.55	3.56	0.72	0.239	$1.47^{n.s}$
R50	2	2.75	3.68	0.75	0.125	2.11^{**}
	3	2.80	3.57	0.78	0.056	3.31^{**}
	4	2.65	3.61	0.73	0.292	$1.36^{n.s}$
	5	2.52	3.60	0.70	0.004	11.07^{***}

n.s : P>0.05
* : P<0.05
** : P<0.01
*** : P<0.001

R40 grubunda birimlerin 3 kümeye ayrılması önemli parçalanma olarak belirlenmiştir ($P<0.001$)

R50 grubunda 5 gruba birimlerin ayrılması önemli bulunmuştur ($P<0.001$).

Tüm rasgele olarak türetilen verilerde önemli bir kümelenme görülmemesi gerekirken g_{1b} ölçütü gruptaki birim sayısı 20 ve daha fazla olduğunda 0.05 yanılma düzeyinde önemli kümelenme belirtisi vermiştir ($P<0.05$).

Koşullu gruplar:

K20, K30, K40, K50 ve K60 gruplarının elemanları olan birimlerin küme sayısı 2-5 olarak seçildiği durumlarda oluşturdukları kümelere göre w_1, b_1, g_{1b}, var(g_{1b}) ve g_{1b}'in önemliliğini test eden test istatistiği z değerleri Çizelge 7.8' de verilmiştir.

K20, K30, K40 ve K60 gruplarında küme belirleme ölçütü g_{1b}, küme sayısı 2-5 durumlarında birimlerin kümelere atanmalarında önemli bir kümelenme vermemiştir ($P>0.05$)

K50 grubunda küme belirleme ölçütü g_{1b}, küme sayısı 2-5 durumlarında birimlerin kümelere atanmalarını önemli bir kümelenme olarak vermiştir ($P<0.05$) ve parçalanmanın küme içi türdeşliği enbüyükleyen küme sayısının 3 olduğunu göstermiştir ($P<0.05$).

İki farklı yapıda çok değişkenli normal dağılımlardan çekilmiş örnekleri içeren koşullu verilerde g_{1b} ölçütü K50

Çizelge 7.8. Koşullu gruplarda g_{1b} ölçütü değerleri

Grup adı	Küme sayısı	w_1	b_1	g_{1b}	var(g_{1b})	z test ist.
K20	2	2.96	4.53	0.65	0.118	1.90$^{n.s}$
	3	0.60	3.49	0.17	1.190	0.16$^{n.s}$
	4	0.39	3.13	0.12	0.202	0.28$^{n.s}$
	5	0.28	2.90	0.10	0.229	0.21$^{n.s}$
K30	2	1.73	4.81	0.36	1.089	0.3$^{4n.s}$
	3	1.90	4.11	0.46	0.065	1.81$^{n.s}$
	4	0.49	3.64	0.13	0.086	0.46$^{n.s}$
	5	0.32	3.50	0.09	0.172	0.22$^{n.s}$
K40	2	1.89	5.27	0.36	0.356	0.60$^{n.s}$
	3	2.06	4.49	0.46	0.059	1.89$^{n.s}$
	4	0.54	3.98	0.14	0.817	0.15$^{n.s}$
	5	0.34	3.83	0.09	0.729	0.10$^{n.s}$
K50	2	2.05	5.69	0.36	0.609	0.46$^{n.s}$
	3	2.22	4.86	0.46	0.048	2.08 *
	4	0.58	4.40	0.13	1.003	0.13$^{n.s}$
	5	0.37	4.14	0.09	0.574	0.12$^{n.s}$
K60	2	1.71	6.07	0.28	1.072	0.27$^{n.s}$
	3	2.38	5.20	0.46	0.232	0.9$^{5n.s}$
	4	0.61	4.60	0.13	0.711	0.16$^{n.s}$
	5	0.40	4.43	0.09	0.978	0.09$^{n.s}$

n.s : P>0.05
* : P<0.05
** : P<0.01
*** : P<0.001

grubu dışında tüm gruplarda ve tüm küme sayılarında önemli kümelenme göstermemiştir ($P>0.05$).

7.4. g_2 Ölçütü

Rasgele gruplar

g_2 ölçütü w_2 değerlerini kullanarak hesaplanmaktadır. w_2 değerleri benzerlik matrisi kare öklid uzaklıklarını içerdiği durumlar için belirlenebilmektedir. Bu nedenle kare Öklid uzaklığı kullanılarak R10, R20, R30, R40 ve R50 gruplarının, küme sayısı 2-5 olarak seçildiği durumlarda oluşturdukları kümelere göre w_2, g_2, var(g_2) ve g_2'nin önemliliğini test eden z test istatistiği değerleri Çizelge 7.9'da verilmiştir.

R10, R20, R30, R40 ve R50 gruplarında g_2, küme sayısı 2-5 durumlarında birimlerin kümelemelere atanmalarını önemli bir kümelenme olarak vermemiştir ($P>0.05$).

Koşullu gruplar:

K20, K30, K40, K50 ve K60 gruplarının küme sayısı 2-5 olarak seçildiği durumlarda oluşturdukları kümelere göre w_2, g_2, var(g_2) ve g_2'nin önemliliğini test eden z test istatistiği değerleri Çizelge 7.10' da verilmiştir.

Rastgele verilerde olduğu gibi K20, K30, K40, K50 ve K60 gruplarında küme belirleme ölçütü g_2 küme sayısı 2-5 durumlarında birimlerin kümelemelere atanmalarını önemli bir kümelenme olarak vermemiştir ($P>0.05$).

Çizelge 7.9. Rasgele gruplarda g_2 ölçütü değerleri

Grup adı	Küme Sayısı	w_2	var(g_2)	z test ist.
R10	2	63.56	202700.20	$0.14^{n.s}$
	3	40.41	330477.00	$0.07^{n.s}$
	4	25.34	201531.70	$0.06^{n.s}$
	5	15.76	257832.70	$0.03^{n.s}$
R20	2	82.48	12509.46	$0.74^{n.s}$
	3	74.52	11406.39	$0.70^{n.s}$
	4	64.05	23177.45	$0.42^{n.s}$
	5	57.59	8417.73	$0.63^{n.s}$
R30	2	122.15	32456.00	$0.72^{n.s}$
	3	111.12	26352.00	$0.68^{n.s}$
	4	95.76	24626.53	$0.61^{n.s}$
	5	82.50	22788.47	$0.55^{n.s}$
R40	2	183.76	65451.33	$0.72^{n.s}$
	3	166.05	52675.00	$0.72^{n.s}$
	4	146.50	48370.00	$0.67^{n.s}$
	5	131.50	15995.48	$1.04^{n.s}$
R50	2	216.93	78543.16	$0.77^{n.s}$
	3	208.44	76867.40	$0.75^{n.s}$
	4	176.90	74564.37	$0.65^{n.s}$
	5	155.04	78932.66	$0.55^{n.s}$

n.s : $P>0.05$
* : $P<0.05$
** : $P<0.01$
*** : $P<0.001$

Çizelge 7.10. Koşullu gruplarda g_2 ölçütü değerleri

Grup adı	Küme Sayısı	w_2	$var(g_2)$	z test ist.
K20	2	34.52	12112.36	$0.36^{n.s}$
	3	27.57	10323.44	$0.27^{n.s}$
	4	24.49	34170.53	$0.13^{n.s}$
	5	23.89	9564.71	$0.24^{n.s}$
K30	2	40.22	27452.86	$0.24^{n.s}$
	3	31.23	26469.84	$0.19^{n.s}$
	4	28.77	26441.47	$0.18^{n.s}$
	5	21.28	22551.93	$0.14^{n.s}$
K50	2	45.18	59315.23	$0.19^{n.s}$
	3	42.19	51005.57	$0.19^{n.s}$
	4	39.24	51197.33	$0.17^{n.s}$
	5	30.70	47210.23	$0.14^{n.s}$
K50	2	59.16	82503.08	$0.21^{n.s}$
	3	56.09	79608.06	$0.20^{n.s}$
	4	52.71	76278.48	$0.19^{n.s}$
	5	42.72	69604.15	$0.16^{n.s}$
K60	2	79.18	40998.30	$0.39^{n.s}$
	3	71.14	396241.20	$0.11^{n.s}$
	4	67.43	383170.52	$0.11^{n.s}$
	5	56.16	372452.53	$0.09^{n.s}$

n.s : $P>0.05$
* : $P<0.05$
** : $P<0.01$
*** : $P<0.001$

Koşullu verilerde birimlerin açıkca enaz iki kümeye ayrıldığı bilinirken g_2 bu ayırıma duyarlı değildir.

7.5. g_3 Ölçütü Bulguları

7.5.1. $g_{3a}=w_3-b_3$ değerlerine ilişkin bulgular

Rasgele gruplar:

R10, R20, R30, R40 ve R50 gruplarının elemanları olan birimlerin küme sayısı 2-5 olarak seçildiği durumlarda oluşturdukları kümelere göre w_3, b_3, g_{3a}, var(g_{3a}) ve g_{3a}'nın önemliliğini test eden z test istatistiği değerleri Çizelge 7.11' de verilmiştir.

R10 grubunda küme belirleme ölçütü g_{3a}, küme sayısı 2-5 durumlarında birimlerin kümelere atanmalarında 3 kümeye ayrılmayı önemli bir kümelenme olarak vermiştir ($P<0.001$).

R20 grubunda rasgele verilerin 4 kümeye ayrılmasının önemli bir parçalanma olduğu belirlenmiştir ($P<0.001$)

R30 grubunda birimlerin 3 kümeye ayrılmalarının önemli bir parçalanma olduğu ortaya çıkmıştır ($P<0.001$)

R40 grubunda birimlerin 5 kümeye ayrılması önemli parçalanma olarak belirlenmiştir ($P<0.001$)

R50 grubunda birimlerin 5 kümeye ayrılmaları önemli bulunmuştur ($P<0.001$).

Çizelge 7.11. Rasgele gruplarda g_{3a} ölçütü değerleri

Grup adı	Küme sayısı	w_3	b_3	g_{3a}	$var(g_{3a})$	z test ist.
R10	2	3.38	3.38	0	0.161	-
	3	2.99	3.54	-0.55	0.055	-2.35**
	4	2.63	3.56	-0.93	1.732	-0.71n.s
	5	2.16	3.39	-1.23	3.857	-0.63n.s
R20	2	2.76	3.52	-0.76	0.036	-4.01***
	3	2.89	3.61	-0.72	0.022	-4.85***
	4	2.69	3.58	-0.89	0.016	-7.04***
	5	2.64	3.56	-0.92	0.022	-6.20***
R30	2	2.72	3.48	-0.76	0.056	-3.21**
	3	2.64	3.64	-1.00	0.007	-11.95***
	4	2.38	3.61	-1.23	0.046	-5.73**
	5	2.46	3.55	-1.09	0.194	-2.47*
R40	2	2.46	3.11	-0.65	0.110	-1.96n.s
	3	2.90	3.62	-0.72	0.009	-7.59***
	4	2.68	3.58	-0.90	0.011	-8.58***
	5	2.58	3.56	-0.98	0.004	-15.50***
R50	2	2.85	3.90	-1.05	0.023	-6.92n.s
	3	2.90	3.61	-0.71	0.012	-6.48***
	4	2.65	3.60	-0.95	0.010	-9.50***
	5	2.51	3.58	-1.07	0.005	-15.13***

n.s : P>0.05
* : P<0.05
** : P<0.01
*** : P<0.001

Tüm rasgele olarak türetilen verilerde önemli bir kümelenme görülmemesi gerekirken g_{3a} ölçütü gruptaki birim sayısı 20 ve daha fazla olduğunda 0.05 yanılma düzeyinde önemli kümelenme belirtisi vermiştir (P<0.05).

Koşullu gruplar:

K20, K30, K40, K50 ve K60 gruplarının elemanları olan birimlerin küme sayısı 2-5 olarak seçildiği durumlarda oluşturdukları kümelere göre w_3, b_3, g_{3a}, var(g_{3a}) ve g_{3a}'nın önemliliğini test eden z test istatistiği değerleri Çizelge 7.12'de verilmiştir.

K20 grubunda küme belirleme ölçütü g_{3a}, küme sayısı 2-5 durumlarında birimlerin kümelere atanmalarını önemli bir kümelenme olarak vermiştir (P<0.001) ve parçalanmanın küme içi türdeşliği enbüyükleyen küme sayısının 3 olduğunu göstermiştir (P<0.001).

K30 grubunda g_{3a}, küme sayısı 2-5 durumlarında birimlerin kümelere atanmalarını önemli bir kümelenme olarak vermiştir (P<0.001) ve parçalanmanın küme içi türdeşliği enbüyükleyen küme sayısının 4 olduğunu göstermiştir (P<0.001).

K40 grubunda g_{3a}, küme sayısı 2-5 durumlarında birimlerin kümelere atanmalarını önemli bir kümelenme olarak vermiştir (P<0.001) ve parçalanmanın küme içi türdeşliği enbüyükleyen küme sayısının 4 olduğunu göstermiştir (P<0.001).

Çizelge 7.12. Koşullu gruplarda g_{3a} ölçütü değerleri

Grup adı	Küme sayısı	w_3	b_3	g_{3a}	var(g_{3a})	z test ist.
K20	2	1.78	3.36	-1.58	0.161	-3.94***
	3	1.26	4.33	-3.07	0.059	-12.64***
	4	1.06	3.99	-2.93	0.056	-12.38***
	5	1.47	3.93	-2.46	0.113	-7.32***
K30	2	1.96	5.01	-3.05	0.173	-7.33***
	3	1.07	4.60	-3.53	0.148	-9.18***
	4	1.05	4.51	-3.46	0.141	-9.21***
	5	0.97	4.45	-3.48	0.159	-8.73***
K40	2	1.12	4.11	-2.99	0.196	-3.05***
	3	1.10	3.17	-2.07	0.187	-4.79***
	4	1.08	3.16	-2.08	0.187	-4.81***
	5	1.05	3.15	-2.10	0.201	-4.68***
K50	2	1.18	5.93	-4.75	0.344	-8.10***
	3	1.17	5.56	-4.39	0.236	-9.04***
	4	1.16	5.56	-4.40	0.245	-8.89***
	5	1.14	5.50	-4.36	0.246	-8.79***
K60	2	1.36	6.01	-4.65	0.259	-9.14***
	3	1.22	5.96	-4.74	0.256	-9.37***
	4	1.21	5.96	-4.75	0.264	-9.24***
	5	1.20	5.92	-4.72	0.286	-8.83***

n.s : $P>0.05$
* : $P<0.05$
** : $P<0.01$
*** : $P<0.001$

K50 grubunda g_{3a}, küme sayısı 2-5 durumlarında birimlerin kümelere atanmalarını önemli bir kümelenme olarak vermiştir (P<0.001) ve parçalanmanın küme içi türdeşliği enbüyükleyen küme sayısının 3 olduğunu göstermiştir (P<0.001).

K60 grubunda g_{3a}, küme sayısı 2-5 durumlarında birimlerin kümelere atanmalarını önemli bir kümelenme olarak vermiştir (P<0.001) ve parçalanmanın küme içi türdeşliği enbüyükleyen küme sayısının 3 olduğunu göstermiştir (P<0.001).

İki farklı yapıda çok değişkenli normal dağılımlardan çekilmiş örnekleri içeren koşullu verilerde g_{3a} ölçütü, tüm gruplarda ve tüm küme sayılarında önemli kümelenmeyi işaret etmektedir (P<0.001).

7.5.2. $g_{3b}=w_3/b_3$ değerlerine ilişkin bulgular

Rasgele gruplar:

R10, R20, R30, R40 ve R50 gruplarının elemanları olan birimlerin küme sayısı 2-5 olarak seçildiği durumlarda oluşturdukları kümelere göre w_3, b_3, g_{3b} , var(g_{3b}) ve g_{3b}'nin önemliliğini test eden z test istatistiği değerleri Çizelge 7.13'de verilmiştir.

R10 grubunda küme belirleme ölçütü g_{3b}, küme sayısı 2-5 durumlarında birimlerin kümelere atanmalarında 3 kümeye ayrılmasını önemli bir kümelenme olarak vermiştir (P<0.001).

Çizelge 7.13. Rasgele gruplarda g_{3b} ölçütü değerleri

Grup adı	Küme sayısı	w_3	b_3	g_{3b}	var(g_{3b})	z test ist.
R10	2	3.38	3.38	1	2.211	$0.67^{n.s}$
	3	2.99	3.54	0.85	0.055	3.60^{***}
	4	2.63	3.56	0.74	1.732	$0.56^{n.s}$
	5	2.16	3.39	0.64	3.857	$0.32^{n.s}$
R20	2	2.76	3.52	0.78	0.036	4.13^{***}
	3	2.89	3.61	0.80	0.022	5.40^{***}
	4	2.69	3.58	0.75	0.016	5.94^{***}
	5	2.64	3.56	0.74	0.022	5.00^{***}
R30	2	2.72	3.48	0.78	0.056	3.30^{***}
	3	2.64	3.64	0.73	0.007	8.67^{***}
	4	2.38	3.61	0.66	0.046	3.07^{**}
	5	2.46	3.55	0.69	0.194	$1.57^{n.s}$
R40	2	2.46	3.11	0.79	0.110	2.28^{*}
	3	2.90	3.62	0.80	0.009	8.44^{***}
	4	2.68	3.58	0.75	0.011	7.14^{***}
	5	2.58	3.56	0.73	0.004	11.46^{***}
R50	2	2.85	3.90	0.73	0.023	4.82^{***}
	3	2.90	3.61	0.80	0.012	7.33^{***}
	4	2.65	3.60	0.74	0.010	7.36^{***}
	5	2.51	3.58	0.70	0.005	9.92^{***}

n.s : P>0.05
* : P<0.05
** : P<0.01
*** : P<0.001

R20 grubunda rasgele verilerin 4 kümeye ayrılmasının önemli bir parçalanma olduğu belirlenmiştir ($P<0.001$)

R30 grubunda birimlerin 3 kümeye ayrılmalarının önemli bir parçalanma olduğu ortaya çıkmıştır ($P<0.001$)

R40 grubunda birimlerin 5 kümeye ayrılması önemli parçalanma olarak belirlenmiştir ($P<0.001$)

R50 grubunda birimlerin 5 kümeye ayrılması önemli bulunmuştur ($P<0.001$).

Tüm rasgele olarak türetilen verilerde g_{3b} ölçütü, gruptaki birim sayısı 20 ve daha fazla olduğunda 0.05 yanılma düzeyinde önemli kümelenme belirtisi vermiştir ($P<0.05$).

Koşullu gruplar:

K20, K30, K40, K50 ve K60 gruplarının elemanları olan birimlerin küme sayısı 2-5 olarak seçildiği durumlarda oluşturdukları kümelere göre w_3, b_3, g_{3b}, var(g_{3b}) ve g_{3b}'nin önemliliğini test eden z test istatistiği değerleri Çizelge 7.14'de verilmiştir.

Çizelge 7.14. Koşullu gruplarda g_{3b} ölçütü değerleri

Grup adı	Küme sayısı	w_3	b_3	g_{3b}	var(g_{3b})	z test ist.
K20	2	1.78	3.36	0.53	0.161	$1.32^{n.s}$
	3	1.26	4.33	0.29	0.059	$1.20^{n.s}$
	4	1.06	3.99	0.27	0.056	$1.12^{n.s}$
	5	1.47	3.93	0.37	0.113	$1.11^{n.s}$
K30	2	1.96	5.01	0.39	0.173	$0.94^{n.s}$
	3	1.07	4.60	0.23	0.148	$0.60^{n.s}$
	4	1.05	4.51	0.23	0.141	$0.60^{n.s}$
	5	0.97	4.45	0.22	0.159	$0.55^{n.s}$
K40	2	1.12	4.11	0.27	0.196	$0.62^{n.s}$
	3	1.10	3.17	0.35	0.187	$0.80^{n.s}$
	4	1.08	3.16	0.34	0.187	$0.79^{n.s}$
	5	1.05	3.15	0.34	0.201	$0.74^{n.s}$
K50	2	1.18	5.93	0.20	0.344	$0.34^{n.s}$
	3	1.17	5.56	0.21	0.236	$0.43^{n.s}$
	4	1.16	5.56	0.21	0.245	$0.42^{n.s}$
	5	1.14	5.50	0.21	0.246	$0.42^{n.s}$
K60	2	1.36	6.01	0.23	0.256	$0.44^{n.s}$
	3	1.22	5.96	0.21	0.256	$0.40^{n.s}$
	4	1.21	5.96	0.20	0.264	$0.40^{n.s}$
	5	1.20	5.92	0.20	0.286	$0.38^{n.s}$

n.s : P>0.05
* : P<0.05
** : P<0.01
*** : P<0.001

K20, K30, K40, K50 ve K60 gruplarında küme belirleme ölçütü g_{3b} küme sayısı 2-5 durumlarında birimlerin kümelere atanmalarında önemli bir kümelenme vermemiştir (P>0.05)

İki farklı yapıda çokdeğişkenli normal dağılımlardan çekilmiş örnekleri içeren koşullu verilerde g_{3b} ölçütü, tüm gruplarda ve tüm küme sayılarında önemli kümelenme göstermemiştir (P>0.05).

7.6. Çok Değişkenli Varyans Çözümlemesine Göre Wilk's Lamda Ölçütü ve F Değerlerine İlişkin Bulgular

Rasgele gruplar:

R10, R20, R30, R40 ve R50 gruplarının elemanları olan birimlerin küme sayısı 2-5 olarak seçildiği durumlarda oluşturdukları kümelere göre çokdeğişkenli analiz çözümlemesi yapılarak Wilk's Lamda ölçütü ve F test istatistikleri Çizelge 7.15'de verilmiştir.

Çokdeğişkenli varyans çözümlemesi sonuçlarının bir istatistiği olan lamda grup ortalama vektörleri arasındaki farklılığı değerlendirme ölçütüdür. Wilk's lamda değerlerinin F test istatistiklerinin büyüklüğü ise grup ortalama vektörleri arasındaki türdeş olmayışın bir ölçüsü olarak alındığında Çizelge 7.13'e göre rasgele verilerde grupların içerdiği birim sayısı arttıkça birimler arasındaki çok az benzerlikler bile kümeleme analizi tarafından değerlendirilmekte ve gruplar içi varyans matrisinin çok küçük

Çizelge 7.15. Rasgele gruplarda çokdeğişkenli varyans çözümlemesi Wilk's lamda ölçütü, F test istatistiği ile olasılık ve önemlilik düzeyleri

Grup adı	Küme sayısı	Wilk's Lambda	F test ist.	Önemlilik P
R10	2	0.016	3.51	n.s
	3	0.0002	16.32	***
	4	0.0000004	13.60	***
	5	-	-	-
R20	2	0.013	16.61	***
	3	0.00007	53.46	***
	4	0.000002	68.39	***
	5	0.000001	41.38	***
R30	2	0.018	24.57	***
	3	0.0004	52.11	***
	4	0.0002	33.36	***
	5	0.00009	25.42	***
R40	2	0.007	11.10	***
	3	0.0024	37.71	***
	4	0.0002	47.97	***
	5	0.0001	36.19	***
R50	2	0.007	79.91	***
	3	0.0004	98.04	***
	4	0.0002	64.71	***
	5	0.0005	31.31	***

n.s : P>0.05   * : P<0.05   ** : P<0.01   *** : P<0.001

değerler içermesi nedeniyle küme sayısının arttığı görülmektedir. R10, R30 ve R50 gruplarında uygun küme sayısı 3 iken R20 ve R40 gruplarında 4 olarak belirlenmiştir.

Koşullu gruplar:

K20, K30, K40, K50 ve K60 gruplarının elemanları olan birimlerin küme sayısı 2-5 olarak seçildiği durumlarda oluşturdukları kümelere göre Wilk's Lamda ölçütü değerleri Çizelge 7.16'de verilmiştir.

Koşullu grupların tümünde Wilk's lamda değerleri önemli düzeyde kümelenmeyi belirterek kümelere ayrılan bireylerin birbirlerine göre önemli düzeyde farklılaşma gösterdiklerini göstermiştir. Her grup için birimlerin oluşturacakları küme sayısını belirlemek amacıyla gruplarda seçilen küme sayılarına göre elde edilen F değerleri karşılaştırılırsa Çizelge 7.16'ya göre tüm ayırımların 2-5 küme sayısı için önemli olduğu görülmektedir. Fakat en uygun küme sayısının tüm gruplar için 2 olduğu belirtilebilir ve diğer 3 kümenin gerçek sınıflama olan iki ana kümenin alt kümeleri olabileceği yargısı çıkarılabilir. Bu sonuç Çizelge 7.3'de varılan sonuç ile benzerlik göstermektedir.

Bu sonuç; veri matrisi birbirleriyle çok farklı özellik gösteren çokdeğişkenli verileri içeriyor ise birlikte değerlendirmeye alınan farklı grupların farklı kümelere atanmasına karşı Wilk's lamda istatistiğinin çok duyarlı olduğunu göstermektedir. Birleştirilen veri seti kadar ana kümeye ayrılmanın işaretini lamda verebilmektedir.

Çizelge 7.16. Koşullu gruplarda çokdeğişkenli varyans çözümlemesi Wilk's lamda ölçütü, F test istatistiği ile olasılık ve önemlilik düzeyleri

Grup adı	Küme sayısı	Wilk's Lambda	F test ist.	Önemlilik P
K20	2	0.0100	19.45	***
	3	0.0059	9.80	***
	4	0.0032	6.90	***
	5	0.0028	4.64	***
K30	2	0.0150	27.67	***
	3	0.0034	22.56	***
	4	0.0013	17.81	***
	5	0.0002	20.82	***
K40	2	0.0190	34.93	***
	3	0.0055	26.66	***
	4	0.0025	20.82	***
	5	0.0004	25.53	***
K50	2	0.0200	43.45	***
	3	0.0074	31.00	***
	4	0.0043	22.60	***
	5	0.0012	23.92	***
K60	2	0.0100	56.42	***
	3	0.0067	39.90	***
	4	0.0040	28.91	***
	5	0.0013	28.59	***

n.s : P>0.05   * : P<0.05   ** : P<0.01   *** : P<0.001

7.7. Çok Değişkenli Varyans Çözümlemesine Göre Hotelling-Lawley İz Ölçütü ve F Değerlerine İlişkin Bulgular

Rasgele gruplar:

R10, R20, R30, R40 ve R50 gruplarının elemanları olan birimlerin küme sayısı 2-5 olarak seçildiği durumlarda oluşturdukları kümelere göre çokdeğişkenli analiz çözümlemesi yapılarak Hotelling Lawley İz ölçütü ve F test istatistikleri Çizelge 7.17'de verilmiştir.

Çok değişkenli varyans çözümlemesi sonuçlarının bir istatistiği olan Hotelling-Lawley trace (iz) ölçütü çok değişkenli varyans çözümlemesinde gruplar ortalama vektörünün farklılığı değerlendirme ölçütüdür. İz değerlerinin F test istatistiklerinin büyüklüğü ise grup ortalama vektörleri arasındaki türdeş olmayışın bir ölçüsü olarak alındığında Çizelge 7.17'ye göre rasgele verilerde grupların içerdiği birim sayısı arttıkça birimler arasındaki çok az benzerlikler bile kümeleme analizi tarafından değerlendirilmekte ve gruplar içi varyans matrisinin çok küçük değerler içermesi nedeniyle küme sayısının arttığı görülmektedir. R10 ve R40 gruplarında uygun küme sayısı 4 iken R20, R30 ve R50 gruplarında 3 olarak belirlenmiştir. Bu sonuç lamda değerine R10 sınıfı dışında benzerlik göstermektedir.

Çizelge 7.17. Rasgele gruplarda çokdeğişkenli varyans çözümlemesi Hotelling-Lawley İz Ölçütü, F test istatistiği ile olasılık ve önemlilik düzeyleri

Grup adı	Küme sayısı	Hotelling-Lawley İz Değeri	F test ist.	Önemlilik P
R10	2	30.75	5.13	n.s
	3	2143.97	74.91	***
	4	4942.01	102.96	***
	5	-	-	-
R20	2	31.68	31.68	***
	3	648.65	384.39	***
	4	784.49	310.53	***
	5	790.05	221.21	***
R30	2	18.19	33.35	***
	3	158.76	182.28	***
	4	162.85	132.32	***
	5	166.85	102.34	***
R40	2	82.01	34.17	***
	3	38.87	66.22	***
	4	208.79	256.64	***
	5	227.97	215.81	***
R50	2	40.33	141.15	***
	3	188.14	425.06	***
	4	192.91	317.49	***
	5	38.45	49.22	***

n.s : P>0.05 * : P<0.05 ** : P<0.01 *** : P<0.001

Koşullu gruplar:

K20, K30, K40, K50 ve K60 gruplarının elemanları olan birimlerin küme sayısı 2-5 olarak seçildiği durumlarda oluşturdukları kümelere göre Hotelling-Lawley iz ölçütü değerleri Çizelge 7.18'de verilmiştir.

Koşullu grupların tümünde iz değerleri tüm küme sayıları için önemli kümelenmeyi belirterek kümelere ayrılan bireylerin birbirlerine göre önemli düzeyde farklılaşma gösterdiklerini göstermiştir. Her grup için birimlerin oluşturacakları küme sayısını belirlemek amacıyla gruplarda seçilen küme sayılarına göre elde edilen F değerleri karşılaştırılırsa Çizelge 6.18'e göre tüm ayırımların 2-5 küme sayısı için önemli olduğu görülmektedir. Fakat en uygun küme sayısının tüm gruplar için 2 olduğu belirtilebilir ve diğer 3 kümenin gerçek sınıflama olan iki ana kümenin alt kümeleri olabileceği yargısı çıkarılabilir. Bu sonuç, Çizelge 7.3 ve Çizelge 7.16'daki Wilk's lamda ölçütüne göre bulunan değerler ile uyuşum göstermektedir.

İz değeri, lamda değerleri ile önemli düzeyde yüksek benzer sonuçlar vermektedir.

7.8. Ayırma Çözümlemesi

Rasgele R10, R20, R30 ve R50 gruplarında 2-5 küme oluşturulması durumunda birimlerin uygun kümelerde yer aldığını göstermiştir. Kümeleme çözümlemesine göre oluşan

Çizelge 7.18. Koşullu gruplarda çokdeğişkenli varyans çözümlemesi Hotelling-Lawley İz Ölçütü, F test istatistiği ile olasılık ve önemlilik düzeyleri

Grup adı	Küme sayısı	Hotelling-Lawley İz Değeri	F test ist.	Önemlilik P
K20	2	87.06	89.46	***
	3	104.55	61.96	***
	4	106.04	41.97	***
	5	108.25	30.31	***
K30	2	59.98	109.96	***
	3	68.64	78.80	***
	4	83.71	68.01	***
	5	120.07	73.64	***
K40	2	52.14	139.04	***
	3	62.13	105.84	***
	4	72.27	88.83	***
	5	97.85	92.63	***
K50	2	43.58	152.52	***
	3	53.64	121.19	***
	4	57.94	95.36	***
	5	71.12	91.03	***
K60	2	44.32	192.05	***
	3	52.71	148.37	***
	4	57.47	118.53	***
	5	67.16	108.34	***

n.s : P>0.05   * : P<0.05   ** : P<0.01   *** : P<0.001

kümelerde birimlerin doğru sınıflandırma olasılıkları tüm gruplar için % 100 olarak belirlenmiştir. Bu nedenle bulgulara ilişkin bir çizelge hazırlanarak burada verilmemiştir.

Koşullu gruplarda da rasgele gruplarda elde edilen % 100 doğru sınıflama sonuçları ile benzer sonuçlar elde edilmiştir. Yine koşullu gruplarda yapılan ayırma çözümlemesi sonuçlarına ilişkin bir çizelge verilmemiştir.

Her küme sayısı için aynı doğru sınıflandırma oranlarına ulaşılması küme sayısının belirlenmesi açısından ayırma çözümlemesinin yararlı ölçü getirmediği ortaya çıkmaktadır. Ayırma çözümlemesi, kümeleme çözümlemesi ile ortaya çıkan kümelerin oluşumunu denetleme açısından yararlanılacak yardımcı bir yöntem olmaktadır.

7.9. Kofenetik Korelasyon Katsayısı

Öklid uzaklığı içeren S ile TBK yöntemi ile bağlanan kümelerin bağlantı katsayıları matrisi C elamanları arasındaki ilişkiyi ölçen kofenetik korelasyon katsayıları rasgele gruplarda Çizelge 7.19'de verilmiştir. Aşamalı bağlantılar küme sayısından etkilenmeksizin birim uzaklık düzeylerinde birimlerin birbirlerine bağlanmalarını yaptığından küme sayısına bakılmaksızın tek bir C matrisi hesaplanmaktadır. Bu nedenle her grup için tek bir kofenetik korelasyon katsayısı hesaplanmıştır.

Çizelge 7.19. Rasgele gruplarda TBK yöntemiyle aşamalı olarak kümelenen birimlerin Kofenetik ilişki katsayıları ve önemlilikleri

Grup Adı	r_{cs}	Önemlilik
R10	0.996	***
R20	0.417	n.s.
R30	0.564	***
R40	0.259	n.s
R50	0.123	n.s.

n.s : P>0.05
*** : P<0.001

R10 ve R30 gruplarında birimlerin birbirleriyle bağlantı düzeyleri ile benzerlik düzeyleri arasında önemli düzeyde ilişki bulunmuştur (P<0.001***). Buna karşılık R20, R40 ve R50 gruplarında benzerlikler ile bağlantı düzeyleri arasında önemli ilişki saptanamamıştır (P>0.05).

Koşullu gruplarda kofenetik korelasyon katsayıları Çizelge 7.20'de verilmiştir.

Çizelge 7.20. Koşullu gruplarda TBK yöntemiyle aşamalı olarak kümelenen birimlerin Kofenetik ilişki katsayıları ve önemlilikleri

Grup Adı	r_{cs}	Önemlilik
K20	0.934	***
K30	0.336	n.s
K40	0.954	***
K50	0.848	***
K60	0.964	***

n.s : P>0.05
*** : P<0.001

K20, K40, K50 ve K60 gruplarındaki benzerlikler ile bağlantı düzeyleri arasında önemli ilişki saptanmıştır (P<0.001), buna karşılık K20 grubunda, R20 grubunda olduğu gibi bir ilişki bulunamamıştır (P>0.05).

8. TARTIŞMA VE SONUÇ

Kümeleme çözümlemesi, n bireyden elde edilen X_i değişkenlerini dikkate alarak bireylerin alt kümelere ve gruplara ayrılmasını sağlayan çokdeğişkenli bir çözümleme yöntemidir. Kümeleme çözümlemesinde, değişkenler arası uzaklıklardan giderek birimlerin tüm değişkenler yönünden ait oldukları kümelere atanmaları amaçlanmaktadır.

Farklı kümeleme yaklaşımlarının etkinliklerinin karşılaştırılması; grup sayısı, değişken sayısı, kümelerin şekilleri, küme şekillerindeki benzerlikler, kümelerin göreli büyükleri, kümelerin ayrımı ve mevcut olan toplam örnek büyüklüğü gibi birçok faktörü kapsamaktadır (Symons, 1981).

Kümeleme çözümlemesinde en yaygın yöntemler Aşamalı Kümeleme Yöntemi ile Aşamalı Olmayan Kümeleme Yöntemleridir. Aşamalı yöntemlerin karşılaştırılmasıyla ilgili bugüne kadar yapılan araştırmalarda, Tek Bağlantı Kümeleme (TBK) yönteminin diğer yöntemlere göre doğal sınıflamaya en uygun sonuçları veren yöntem olduğu belirlenmiştir (Everitt, 1979; Mardia et.al., 1979; Jambu and Lebeaux, 1986; Özdamar, 1988). Aşamalı Olmayan Kümeleme Yöntemlerinin de karşılaştırılmasında yapılan çalışmalarda K-Ortalamalar yöntemi diğer yöntemlere göre en uygun yaklaşımı veren yöntem olarak belirlenmiştir ve SYSTAT, SPSS/PC+, BMDP vb. paket programlarda default prosedür olarak K-Ortalamalar yöntemi ele alınmıştır (Anderberg, 1973; Punj and Stewart, 1983; Johnson and Wichern, 1988). Araştırmamızda, bu iki yöntem birlikte kullanılarak veri matrisinin kümelere ayrılması sağlanmıştır.

Seçilen yönteme göre birimlerin kümelenmesinde, benzerlik düzeylerini belirleyen uzaklık ölçütlerinden Öklid uzaklığı, en etkin ve yaygın olarak kullanılan ölçüttür (Kendall, 1980; Punj and Stewart, 1983; Özdamar, 1988).

Bu yöntemlerle belirlenen küme sayılarının uygunluğu için; birimlerin kendi içinde türdeş, diğerleriyle türdeş olmayan kümelere ayrılabilmesinde ve küme sayısının geçerliliğinin test edilmesinde bir çok ölçüt tanımlanmıştır (Anderberg, 1973; Sneath and Sokal, 1973; Everitt, 1979; Klastorin, 1983). Bu ölçütlerin geçerliliğini, rasgele ve koşullu türetilmiş verilerde etkinliklerini test etmek amacıyla bu araştırma düzenlenmiştir.

Bu araştırmada, çokdeğişkenli normal dağılımdan türetilmiş farklı sayı ve koşul içeren rasgele R10, R20, R30, R40 ve R50; koşullu K20, K30, K40, K50 ve K60 gruplarında yer alan 6 değişken değerine sahip birimlerin Öklid uzaklığı içeren benzerlik matrisleri hesaplanmıştır. Bu benzerlik matrisinden yararlanılarak birimler TBK ve K-Ortalamalar yöntemleri aracılığı ile kümelere ayrılmış 2 ile 5 arasında küme sayısı ve bağlantı uzaklıklarına göre birimlerin oluşturdukları kümeler ve bu kümelerde yer alan birimler (küme elemanları) belirlenmiştir.

Oluşturulan kümelerin uygun bir küme olup olmadığını denetlemek için g_{1a}, g_{1b}, g_{2}, g_{3a}, g_{3b} ölçütleri hesaplanmıştır. Kümeleme sonucu her birimin girdiği kümelere birer numara verilmiş ve her küme sayısı için çokdeğişkenli varyans çözümlemesi yapılarak birimlerin ait oldukları kümelerin özelliklerinin, kendi içinde türdeş ve kendi aralarında türdeş olmayan bir yapıda olup olmadıklarını denetlemek için Wilk's Lamda ve Hotelling-Lawley iz ölçütleri hesaplanmıştır. Ayrıca her kümenin 6 değişken için elde edilen ayırma fonksiyonlarına göre doğru sınıflandırma olasılıkları hesaplanmıştır. Benzerlik matrisi ile bağlantı katsayıları matrisinin benzerliğini belirten Kofenetik korelasyon katsayısı hesaplanmıştır.

Rasgele türetilmiş ve 10 birim içeren R10 grubu kümelemesinde; g_{1a}, g_{1b} ve g_2 ölçütleri önemli kümelenme olmadığını belirtirken; g_{3a}, g_{3b} ve Wilk's Lamda ölçütleri 3 kümeye parçalanmayı önemli kümelenme olarak vermiştir. Hotelling-Lawley iz ölçütü ise 4 kümeye ayrılmayı önemli kümelenme olarak bulmuştur. TBK, 2 birim bağlantı uzaklığına göre 8 birimi bağımsız olarak nitelemiş, diğer iki birim bir küme oluşturmuştur. Ayırma çözümlemesinde sınıflandırma olasılıkları tüm küme sayıları için %100 olarak bulunmuştur. Kofenetik korelasyon katsayısı, küme sayısına duyarlı olmadığı için tek değer olarak bulunmuş ve önemli düzeyde ilişki saptanmıştır ($P<0.001$). R10 grubunda ölçütlerin birbirinden değişik sonuçlar vermesini, birim sayısının yetersizliğine bağlamak gerekir. Nitekim, Johnson ve Wichern (1988) çokdeğişkenli yapılar çözümlenirken birim sayısının fazla olmasının yararlı olacağını belirtmiştir.

R20 grubunda; g_{1a}, g_{1b}, g_{3a}, g_{3b}, Wilk's Lamda ölçütü 4 kümeye, Hotelling-Lawley iz ölçütü 3 kümeye ayrılmayı önemli kümelenme olarak bulmuştur. TBK yöntemi, 2 birim uzaklık ölçütüne göre, 10 birimi bağımsız olarak nitelerken diğer 10 birimi 3 kümede toplamıştır. g_2 ölçütü ve Kofenetik korelasyon katsayısı önemli kümelenme belirtmemişlerdir ($P>0.05$). Ayırma çözümlemesinde sınıflandırma olasılıkları tüm küme sayıları için %100 olarak bulunmuştur. Bu rasgele grupta da ölçütlerin ortaya koyduğu sonuçlar uyum göstermemiştir. Fakat g_{1a}, g_{1b}, g_{3a}, g_{3b}, Wilk's Lamda ölçütü R20'nin kümelenmesinde benzer sonuçlar vermiştir. Bu yöntemler, küme içi ve kümeler arası varyans-kovaryans matrislerinin farklarını ya da oranlarını ele alan yöntemlerdir.

R30 grubunda; g_{1a} ve g_{1b} ölçütlerinde 5 kümenin uygun olduğu ortaya çıkarken, g_{3a}, g_{3b} ve Wilk's Lamda ve Hotelling-Lawley iz ölçütlerinde 3 kümeye parçalanma önemli kümelenme olarak bulunmuştur. TBK yönteminde 2 birim uzaklık için, 3 birim bağımsız olarak nitelenmekte diğer birimler ise 5 kümeye atanmaktadır. Kofenetik korelasyon katsayısında önemli düzeyde ilişki bulunmuştur ($P<0.001$). g_2 ölçütünde önemli kümelenme bulunamamıştır. Ayırma çözümlemesinde sınıflandırma olasılıkları tüm küme sayıları için %100 olarak bulunmuştur. R30 grubunda ölçütler arasındaki benzerlikler azalmış fakat Wilk's Lamda ve Hotelling-Lawley iz ölçütü benzer sonuç vermiştir.

R40 grubunda; g_{1a} ve g_{1b} ölçütlerinde 3 küme oluşturulması önemli iken; Wilk's Lamda ve Hotelling-Lawley iz ölçütlerinde 4 kümeye parçalanma önemli kümelenme olarak bulunmuştur. TBK yönteminde 2 birim uzaklıkta, 5 birim bağımsız iken diğer birimler 9 küme oluşturmaktadırlar. Kofenetik korelasyon katsayısında önemli bir ilişki saptanamamıştır ($P>0.05$). g_{3a} ve g_{3b} ölçütünde ise 5 kümeye ayrılma önemli kümelenme vermiştir. g_2 ölçütünde önemli kümelenme bulunamamıştır. Ayırma çözümlemesinde sınıflandırma olasılıkları tüm küme sayıları için %100 olarak bulunmuştur. R40 grubunda da ölçütlerin verdiği sonuçlarda farklılaşmalar ortaya çıkmıştır.

R50 grubunda; g_{1a}, g_{1b}, g_{3a} ve g_{3b} ölçütlerinde 5 küme oluşturulması önemli iken, Wilk's Lamda ve Hotelling-Lawley iz ölçütlerinde 3 kümeye parçalanma önemli kümelenme olarak bulunmuştur. TBK yönteminde 2 birim uzaklık için, 4 birim bağımsız diğerleri ise 4 kümede toplanmıştır. Kofenetik korelasyon katsayısında önemli bir ilişki saptanamamıştır ($P>0.05$). g_2 ölçütünde önemli kümelenme bulunamamıştır. Ayırma çözümlemesinde sınıflandırma olasılıkları tüm küme sayıları için %100 olarak bulunmuştur. R50 grubunda da ölçütler farklı uygun küme sayılarını işaret etmişler ve uyum gözlenmemiştir.

Rasgele gruplarda birim sayısı arttıkça g_{1a}, g_{1b}, g_{3a} ve g_{3b} ölçütleri daha fazla sayıda kümenin parçalanma için uygunluk belirtirken, Wilk's Lamda ve Hotelling-Lawley iz ölçütü benzer sonuçları ve 3 dolayında küme sayısını belirtme eğilimi göstermiştir. TBK ve Kofenetik korelasyon katsayısı yaklaşımları grupta tek başına küme oluşturan bağımsız birim sayısı arttıkça önemliliğini kaybetme eğilimi göstermiş ve 3 ve daha fazla birim içeren küme sayısını, Wilk's Lamda ve Hotelling-Lawley iz değerlerine benzer şekilde belirtme eğilimi göstermiştir.

Koşullu türetilmiş ve 20 birim içeren K20 grubu kümelemesinde; Wilk's Lamda ve Hotelling-Lawley iz ölçütlerinde 2 kümeye parçalanma, g_{1a} 4 kümeye, g_{3a} ise 3 kümeye ayrılmayı önemli kümelenme olarak belirtmiştir. TBK yönteminde 2 kümeye ayrılmayı uygun olarak belirtmiş ve

Kofenetik korelasyon katsayısı bu parçalanmayı önemli olarak nitelemektedir ($P<0.001$). g_{1b}, g_{3b} ve g_2 ölçütlerinde önemli kümelenme bulunamamıştır. Ayırma çözümlemesinde sınıflandırma olasılıkları tüm küme sayıları için %100 olarak bulunmuştur. Bu grupta grup içi varyans-kovaryans matrisinin gruplar arası varyans-kovaryans matrisine oranını hedef alan ölçütlerde birliktelik gözlenmiş ve bu sonuca TBK ve Kofenetik korelasyon katsayıları da uyumluluk göstermiştir.

K30 grubunda; g_{1a} ve g_{3a} ölçütlerinde 4 kümeye parçalanma önemli iken Wilk's Lamda ve Hotelling Lawley iz ölçütlerinde 2 kümeye parçalanma önemli kümelenme olarak bulunmuştur. TBK, 3 kümeyi uygun olarak nitelemekte ve Kofenetik korelasyon katsayısı önemsiz kümelenmeyi işaret etmektedir ($P>0.05$). g_{1b}, g_2 ve g_{3b} ölçütlerinde önemli kümelenme bulunamamıştır. Ayırma çözümlemesinde sınıflandırma olasılıkları tüm küme sayıları için %100 olarak bulunmuştur. Bu grupta, grup içi varyans-kovaryans matrisinin gruplar arası varyans-kovaryans matrisine oranını hedef alan ölçütlerde birliktelik kaybolmuş ve Wilk's Lamda ve Hotelling-Lawley iz ölçütü benzer sonuç vermiştir. TBK bu ölçütlerin sonucuna yakın sonuç vermiştir.

K40 grubunda; g_{1a} ölçütünde 3 kümeye parçalanma önemli iken, g_{3a} ölçütünde 4 kümeye parçalanma önemli bulunmuştur. Wilk's Lamda ve Hotelling-Lawley iz ölçütlerinde 2 kümeye parçalanma önemli kümelenme olarak bulunmuştur. TBK yönteminde, 2 birim uzaklıkta, 3 küme oluşumu ortaya çıkmıştır. g_{1b}, g_2 ve g_{3b} ölçütlerinde önemli kümelenme bulunamamıştır. Ayırma çözümlemesinde sınıflandırma olasılıkları tüm küme sayıları için %100 olarak bulunmuştur. Kofenetik korelasyon katsayısında önemli düzeyde ilişki bulunmuştur ($P>0.001$).

Wilk's Lamda ve Hotelling-Lawley iz ölçütleri benzer sonuçlar verirken aynı yaklaşımı kullanan diğer ölçütler farklı sonuçlar vermiştir.

K50 grubunda; g_{1a}, g_{1b} ve g_{3a} ölçütlerinde 3 kümeye parçalanma önemli iken, Wilk's Lamda ve Hotelling-Lawley iz ölçütlerinde, 2 kümeye parçalanma önemli kümelenme olarak bulunmuştur. TBK yöntemi 3 kümeyi işaret etmektedir. Kofenetik korelasyon katsayısında önemli düzeyde ilişki bulunmuştur. g_2 ve g_{3b} ölçütlerinde önemli kümelenme bulunamamıştır. Ayırma çözümlemesinde sınıflandırma olasılıkları tüm küme sayıları için %100 olarak bulunmuştur. Bu grupta Wilk's Lamda ve Hotelling-Lawley iz ölçütlerinde birbirleriyle uyumlu ve TBK yöntemi, g_{1a}, g_{1b} ve g_{3a} ölçütleri fazla küme sayısı belirtme eğilimi göstermişlerdir.

K60 grubunda; g_{1a} ve g_{3a} ölçütlerinde 3 kümeye parçalanma önemli iken Wilk's Lamda ve Hotelling-Lawley iz ölçütlerinde, 2 kümeye parçalanma önemli kümelenme olarak bulunmuştur. TBK yöntemi 3 kümeye parçalanmayı göstermektedir. g_{1b}, g_2 ve g_{3b} ölçütlerinde önemli kümelenme bulunamamıştır. Ayırma çözümlemesinde sınıflandırma olasılıkları tüm küme sayıları için %100 olarak bulunmuştur. Kofenetik korelasyon katsayısında ileri düzeyde önemli ilişki saptanmıştır ($P<0.001$).

Rasgele ve koşullu gruplarda küme sayısını belirlemeye yönelik ölçütlerde birliktelik ve uyumluluk gözlenmemiştir. Rasgele gruplarda birim sayısı artarken, Wilk's Lamda ve Hotelling-Lawley iz ölçütleri dışındaki ölçütlerin küme sayılarında azalma artma biçiminde dengesiz dalgalanmalar gözlenmiştir. Koşullu verilerde iki farklı parametreli çokdeğişkenli normal dağılım gösteren birimlerin toplanmasına rağmen seçilen tüm küme sayılarında g_{1a}, g_{3a} ve TBK yöntemi 3 kümeye parçalanmayı belirtirken, Wilk's Lamda ve Hotelling-Lawley iz ölçütleri 2 küme belirtmişlerdir. Bu iki ölçüt iki farklı dağılımdan çekilmiş birimleri %100 doğru sınıflandırma yapan 2 küme sayısını en uygun küme olarak belirtmektedir. Koşullu olarak alınan bu durumu en iyi yansıtan ölçü Wilk's Lamda ve Hotelling-Lawley iz ölçütleri olmuştur. g_{1a}, g_{1b} ve g_{3a} ve g_{3b} ölçütleri aynı özellikleri taşıyan veri setlerinde değişik sonuçlar verme eğilimi göstermiştir. TBK, seçilen 2 birim uzaklıktaki bağlantılarda Wilk's Lamda ve Hotelling-Lawley iz değerlerine yakın kümeleri belirtmiştir. Kriter olarak alınacak bağlantı uzaklığı 2.5, 3 bağlantı katsayısı ya da daha büyük değerler olarak alındığında küme sayısı Wilk's Lamda ve Hotelling-Lawley iz değerlerinde belirtilenle benzerlik gösterebilecektir.

g_2 ölçütü rasgele ve koşullu verilerde hiç bir önemlilik belirtisi göstermemiştir. Açıkça farklı olarak türetildikleri bilinen koşullu setlerde g_2'nin bu koşula duyarlılık göstermemesi bu ölçütün uygun kümeleri belirleme açısından yeterli olmadığını ortaya koymaktadır.

g_{1a} ve g_{1b} rasgele gruplarda kendi aralarında uyumlu sonuçlar vermesine rağmen, koşullu gruplarda en azından 2 farklı yapının bulunduğu durumlarda tutarsız sonuçlar vermiştir. Aynı şekilde g_{3a} ve g_{3b} rasgele gruplarda kendi aralarında uyumlu sonuçlar verirken koşullu gruplarda bu tutarlılığını kaybederek önemsiz ya da fazla kümelenmeyi gösteren sonuçlar verme eğilimi göstermiştir.

Bu kümeleme ölçütlerini birbirleriyle karşılaştırarak etkinliklerini ortaya koyan bir çalışmaya rastlamadığımızdan sonuçlarımızı ancak çok sayıda veri setinde tartışarak ortaya koyma gereği duyulmuştur. Fakat yaklaşım olarak varyans-kovaryans matrislerinin farkı, oranı ya da determinantlarının farkı ve bölümü biçiminde bir yaklaşımı izleyen bu yöntemlerin veri sayısı ve koşullar karşısında tutarsız sonuçlar verdiği gözlenmiştir. Buna karşılık Wilk's Lamda ve Hotelling-Lawley iz ölçütleri tutarlı sonuçlar ortaya koymuşlardır. Bu son iki ölçüt, toplam varyans arttıkça küme içi varyansında çokdeğişkenli varyans çözümlemesinde artış göstermesine karşılık çok küçük değerler olarak belirlenmişler fakat olasılık değerleri önemli düzeyde düşme göstermiştir. Araştırmamızda Wilk's Lamda ve Hotelling-Lawley iz ölçütlerinin F test istatistiklerini ve olasılık değerlerini dikkate alarak küme sayısılarının uygunluğunu belirlemeyi seçtik. Olasılık olarak enküçük değeri sahip Wilk's Lamda ve Hotelling-Lawley iz ölçütünün belirlendiği küme sayısını uygun küme sayısı olarak aldık. Seçilen küme sayılarında olasılıklar $P<0.001$ düzeyinde idi.

Kofenetik korelasyon katsayısı rasgele gruplarda, gruplardaki birim sayısı arttığı durumlarda düzensiz değerler olarak ortaya çıkmıştır. ve R10 ile R30 grubunda önemli kümelenmeyi belirtirken diğer R gruplarında önemli kümelenmeyi belirtmemiştir. TBK yönteminin sonuçları ile r_{cs} sonuçları karşılaştırıldığında kümelerde bağımsız birim sayısı arttığı ve küme sayısının çoğaldığı durumlarda artış göstermekte, 3 ve daha fazla birim içeren küme sayısı arttığında azalma göstermektedir. r_{cs} rasgele verilerde koşullara uyumluluk göstermemektedir. Koşullu gruplarda ise koşullara duyarlılık gösterek TBK'daki bağlantıların önemliliğini işaret etmektedir.

Küme sayısının belirlenmesinde kullandığımız ölçütlerden Wilk's Lamda ölçütü birim sayısı 30'un üzerinde olduğunda duyarlılığı diğer ölçütlere göre enyüksek düzeye ulaşmaktadır. Everitt (1979) ve Anderberg (1973) gruplar arası varyansın gruplar içine göre enbüyüklendiği durumları belirleme açısından Wilk's Lamda ölçütünün çok duyarlı olduğunu belirtmişlerdir. Ayrıca kümeleme ölçütlerinde küme içi türdeşliğinin, kümeler arası farklılığa göre değerlendirilmesinde Wilk's Lamda ölçütünün etkinliğini belirtmişlerdir. Anderberg (1973), Klastorin (1983) ve Johnson ve Wichern (1988) gruplararası farklılıkların ortaya konmasında Hotelling-Lawley iz ölçütünün de önemli bir istatistik olduğunu belirtmişlerdir.

Symons (1981) ve Avanzolini ve ark. (1990) aralıklı ya da oransal ölçekle elde edilmiş veri yapılarının en uygun sayıda kümelere ayrılmasında Wilk's Lamda ölçütünün en uygun ölçüt olabileceğini belirtmişlerdir. Bizde, bu araştırmamızda koşullara duyarlılığı bakımından Wilk's Lamdanın özellikle 30 ve daha fazla birim içeren ve neden-sonuç ilişkilerinin tam belirlenemediği veri yapılarını uygun kümelere ayırmada yeterli bir ölçüt olabileceği kanısına vardık.

Kofenetik korelasyon katsayısı aşamalı sınıflamalarda bağlantıların uyumluluğunu belirtme açısından ileri sürülmüş bir yaklaşımdır. Sneath ve Sokal (1973) taksonomik problemlerin çözümünde uygun sınıflamanın belirlenmesi için

Kofenetik korelasyon katsayısının uygun bir ölçüt olduğunu ileri sürmüşlerdir. Fakat Kofenetik korelasyon katsayısı, belirlenen küme sayısına duyarlı değildir. Veri matrisinin aşamalı olarak tek bir bağlantısı elde edilmekte ve bu bağlanmaya ilişkin tek bir Kofenetik korelasyon katsayısı hesaplanabilmektedir. Halbuki, diğer küme içi varyans ve kümeler arası varyans hesaplamalarına dayalı ölçütlerde uygulamacının değişik alternatifleri sınayarak veri matrisini 2, 3, 4, ... vd. gibi değişik sayılarda kümelere ayırarak verileri hakkında daha değişik açılardan bilgi alma olanağı vermekte ve bu küme sayılarına göre hesaplanan küme içi türdeşliği ve kümeler arası farklılığı belirten en uygun durumu belirleme olanağı vermektedir. Kofenetik korelasyon katsayısı bu olanağı araştırıcıya sunmamaktadır.

Overall ve Klett (1972) Hotelling-Lawley iz ölçütünün çokdeğişkenli varyans çözümlemesinde gruplar ortalama vektörünün farklılığının değerlendirilmesinde etkin bir yaklaşım olduğunu belirtmiştir. Bizde bu ölçütün Wilk's Lamda ölçütü ile aynı sonuçlar verdiğini bulduk.

Kümeleme yöntemleriyle küme sayısı 2-5 olduğu durumlarda oluşan kümelere atanan birimlerin çokdeğişkenli özellikleri ve bu özelliklere dayalı olarak ortaya konan küme özelliğinin yapısı Ayırma Çözümlemesi ile incelendiğinde tüm küme sayılarında birimlerin ait oldukları sınıflara %100 doğru olasılıkla atandıklarını göstermektedir. Bu durum kümeleme çözümlemesi ile oluşturulan yapıların gerçekte doğal sınıflamaları yansıtma açısından önemli bilgiler verdiğini göstermektedir. Kurtuluş (1985), kümeleme sonuçlarının ayırma analizi ile denetlenmesinin veri yapıları hakkında ayrıntılı bilgi edinme açısından önemli olduğunu vurgulamıştır. Biz bu araştırmamızda veri matrisinin farklı küme sayılarına göre kümelenmesinde elde edilen sonuçların %100 doğru sınıflandırma verdiğini saptadık. Bu durumda ayırma analizi kümeleme çözümlemesi sonuçlarının bir kontrolü, doğrulaması niteliğinden öteye bir farklı bilgi vermemiştir. Ancak aynı konuda yapılan ek gözlemlerin hangi sınıflamalara atanacağını belirtmesi ve atama ayırma fonksiyonları tanımlaması açısından önem taşıyabilir.

TBK yöntemi en yaygın kullanılan aşamalı kümeleme yöntemi olarak belirlenmiştir. Sneath ve Sokal (1973), Anderberg (1973), Jambu ve Lebeaux (1986), Özdamar (1988) ve Bağcı ve ark. (1991) yaptıkları yöntem karşılaştırmalarında TBK'nın diğer aşamalı yöntemler içinde doğal sınıflamaya en yakın kümelenmeler verdiğini bulmuşlardır. Fakat TBK'da kümeler ve alt kümeler belirli bağlantı uzaklıklarına göre araştırıcı tarafından ortaya konmaktadır. Kesin bir sınırlama olmaması nedeniyle araştırıcının verilerini iyi tanımasına bağlı kalmaktadır. Sneath ve Sokal (1973), TBK'den yararlanırken araştırıcının verilerini çok iyi tanımasını önermişlerdir.

TBK, birimlerin birbirleriyle benzerliklerini bağlantı uzaklıklarına göre aşamalı olarak belirtmesi açısından taksonomik problemlerde sınıf, tür, alt tür gibi sistematik sınıflandırmalarda birimlerin ya da objelerin hangi sınıflama içine girebileceğini belirleme açısından ek bilgi verme özelliğine sahiptir (Mariott, 1971; Özdamar; 1988; Bağcı ve ark. 1991).

Öklid uzaklığı ve ilişki türü ölçüler içeren benzerlik matrisi elemanları değişkenlerin ölçü birimlerinden bağımsızdırlar (Mariott, 1974; Kendall, 1975). Fakat değişkenlerin aralıklı ya da oransal ölçekli olarak değerlerinin ölçülmesi ve veri matrislerinin oluşturulması gerekir. Nitel özellikler üzerinde çalışırken bu özellikleri mümkün olduğunca sıralı ya da aralıklı bir ölçeğe dönüştürerek sayısallaştırmak gerekir. Bunun için geliştirilmiş Thursten ölçeği, Likert ölçeği, Q tipi ölçek gibi uygun ölçeklerden birini kullanmak gerekir (Kurtuluş, 1985). Bu yapıldığında çokdeğişkenli varyans çözümlemesi, elde edilen kümelerin uygun sayısının belirlenmesinde daha etkin olarak kullanılabilecektir. Bazı taksonomik problemlerde serolojik özelliklere dayalı nominal bilgilerden objelerin sınıflandırılmaları yapılmıştır. Ölçüm ve tartımla elde edilen ya da uygun ölçekler seçilerek skor değişkenlerle bilgilerin ifade edilmesi halinde taksonomik yaklaşımlardan daha etkin sonuçlar alınabilir. Bağcı ve ark. (1991), serolojik bilgilere dayalı sınıflandırmalarla elde edilmiş sınıflandırmaların antibiyotik disk duyarlılık testlerine göre yapılacak sınıflamaları karşılaştırmak için antibiyotik disk duyarlılık testinde inhibisyonzonlarını mm cinsinden ölçerek elde edilen veri matrisini kümeleyerek serolojik sonuçlarla benzer sınıflandırmalar elde etmişlerdir. Bu yaklaşımla sınıfların (küme) özelliklerini ve bu özelliklerin parametre tahminlerini elde etmişlerdir.

Araştırmamızda benzerlik matrisinin kümelere ayrılmasında en uygun kümeye ayrılma sayısının test edilmesinde Wilk's Lamda ve Hotelling-Lawley iz ölçütlerinin en etkin ve birbirlerine benzer yöntemler olduğu bulunmuştur.

TBK, yöntemi birimlerin küme oluştururken birbirleriyle bağlanma uzaklıklarını belirtme özelliği vardır ve bu nedenle küme sayısı yanında birimler arası ilişkiler yönünden ek bilgiler almak gerektiğinde başvurulması gereken bir yöntem olduğu kanısına varılmıştır. Kofenetik korelasyon katsayısı aşamalı kümelenmelerin değerlendirilmesinde ikinci derecede öneme sahip bir ölçüt olarak görünmektedir.

Sneath ve Sokal (1973) ve Özdamar (1988), karmaşık veri yapılarının incelenmesinde Aşamalı ve Aşamalı Olmayan Kümeleme Yöntemlerinin bir arada kullanılabileceğini belirtmişlerdir. K-Ortalamalar yönteminin küme istatistiklerini belirtme ve 2, 3, 4, ... gibi artan sayıda birimleri kümelere ayırarak değişik alternatifleri deneme olanağı vermesi önemli bir özelliktir. TBK yöntemi ise birimler arası ilişkileri ve bağlantıları göstermesi bakımından önemli bir özelliğe sahiptir. Karmaşık veri yapılarını çözümlemede araştırıcı bu iki yöntemi birlikte ele alarak veri analizinde kullanması uygun olacaktır. Böylece araştırıcı, küme özellikleri, birimler arası ilişkiler ve yeni yapıların tanımlanması konularında önemli ipuçları elde edbilecektir. Bu bilgileri iki yöntemin vermesi bakımından birlikte kullanılmalarının yararlı olacağı açıktır.

Bilimsel araştırmalarda, klinik, laboratuvar ve saha çalışmalarında uygulamacı ve bilim adamlarının elde ettikleri verilerini çokdeğişkenli olarak ele almalarının yararlı olacağı, veri matrislerinin karmaşık yapıları gösterdiği durumlarda kümeleme çözümlemesine başvurulmasının uygun olacağı açıktır. Bu yüzden; Tıp, Biyoloji vd. alanlarda, birimler, değişkenler ya da hem birimler hem değişkenler hakkında ayrıntılı bilgiler elde edilmesinde çokdeğişkenli yöntemlerden yararlanmak gerekir. Veri analizlerinde yöntem seçiminde Biyoistatistikçilerle ortak çalışmalar yapılması doğru kararlara ulaşmada önemli ve gerekli bir yaklaşım olacaktır.

Kümeleme çözümlemesi, Tıp, Biyoloji, Siyasal Bilimler ve Yer Bilimleri gibi alanlarda karmaşık veri yapılarını çözümlemede yararlanılabilecek önemli bir yöntemdir. Kümeleme yöntemleriyle, seçilen küme sayısında, birimlerin uygun kümelenme gösterip göstermediğinin araştırılarak uygun küme sayısının belirlenmesinde kullanılan kümeleme ölçütlerinden, Wilk's Lamda ve Hotelling-Lawley iz ölçütleri hem birbirleriyle hem de kendi içlerinde uyumlu sonuçlar vererek, araştırılan ölçütler içinde en uygun olanları olarak belirlenmiştir. TBK yöntemi ile birimler arası ilişkilerin kontrol edilmesi, K-Ortalamalar yöntemi ile elde edilemeyen bilgiler açısından gerekli görülmektedir.

KAYNAKLAR DİZİNİ

1. Anderberg, M.R.: Cluster analysis for aplications. Academic Press, New York and London, 1973.

2. Armitage, P.: Statistical methods in medical research. John Wiley, New York, 1983.

3. Arnold, S.J.: A test for cluster. Journal of Marketing Research, XVI: 545-551, 1979.

4. Avanzolini, G., Barbini, P. and Gnudi, G. : Unsupervised learning and discriminant analysis applied to identification of high risk postcardiac patients. Int. J. Biomed. Comput., 25: 207-221, 1990.

5. Bağcı, H., Shareef, S.R. ve Özdamar, K.: Bacillus Thuringiensis varyetelerinin sınıflandırılmasında sayısal taksonominin uygulanması. Doğa-Tr. J. of Biology, 15: 70-81, 1991.

6. Benjamin, Y. and Igbaria, M.: Clustering categories for better prediction of computer resources utilizations. Applied Statist, 40(2): 295-307, 1991.

7. Bock, R.D.: Multivariate statistical methods in behavioral research. McGraw-Hill Book Company, New York, 1975.

KAYNAKLAR DİZİNİ (devam ediyor)

8. Bools, C., Foster, J., Brown, I. and Berg, I.: The identification of psychiatric disorders in children who fail to attend scholl: A cluster analysis of a non-clinical population. Psychological Medicine, 20: 171-181, 1990.

9. Bratley, P., Fox, B.L. and Schrage, L.E.: A quide to simulation. Spinger-Verlag, New York, 1983.

10. Bursalıoğlu, M.H.: Ülkemizde çeşitli habibatlarda izole edilen bacillus türlerinin bazı zararlı böceklerin biyokontrolünde kullanılma olanakları üzerine bir araştırma. Yüksek Lisans Tezi, Ege Üniv. Fen Fak., İzmir, 1985.

11. Chatfield, C. and Collins, A.J.: Introduction to multivariate analysis. Chapman and Hall, London, 1980.

12. Corsten, L.C.A. and Denis, J.B.: Structuring interaction in two-way tables by clustering. Biometrics, 46: 207-215, 1990.

13. Dawson- Saunders, B. and Trapp, R.G.: Basic and clinical bioistatistics. Prentice-Hall International Inc., 1990.

KAYNAKLAR DİZİNİ (devam ediyor)

14. Day, G.S. and Heeler R.M.: Using cluster analysis to improve marketing experiments. Journal of marketing Research, 8: 340-347, 1971.

15. Diggle, P.J., Besag, J. and Gleaves, J.T.: Statistical analysis of spatial point patterns by means of distance methods, Biometrics, 32: 659-667, 1976.

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

17. Flury, B. and Riedwly, H.: T^2 tests linear two-group discriminant function and their computation by linear regression. The American Statistician, 39:(2), 20-25, 1985

18. Funkhouser, G.R.: A note on the reliability of certain clustering algorithms. Journal of Marketing Research, 10: 99-102, 1983.

19. Gaboriaud, C., Uze, G., Lutfolla, G. and Mogensen, K.: Hydrobhobic cluster analysis reveals duplication in the external structure of human - interferon receptor and homology with - interferon receptor external domain. Biomedical Division, 269:(1), 1-3, 1990.

KAYNAKLAR DİZİNİ (devam ediyor)

20. Garety, P.A., Everitt, B.S. and Hemsley, D.R.: The characteristics of delusions: A cluster analysis of deluded subjects. Evr. Arch. Psychiatry Neurol Sci., 237: 112-114, 1988.

21. Green, P.E. and Tull, S.T.: Research for marketing decisions. Prentice-Hall of India Private Limitet, NewDelhi, 1973.

22. Green, P.E., Frank, R.E. and Robinson, P.J.: Cluster analysis In test market selection. Management Science, 13:(8), 387-400, 1967.

23. Hartigan, J.A. and Engelman, L.: BMDP Inc., California, 1988.

24. Hawkins, D.M., Müller, M.W. and Krooden, J.A.: Topics in applied multivariate analysis. Cambridge University Press, USA, 1982.

25. Jambu, M. and Lebeaux, M.O.: Cluster analysis and data analysis. Nord-Holland Pub., Amsterdam, 1986.

KAYNAKLAR DİZİNİ (devam ediyor)

26. Jamison, R.N., Rock, D.L. and Parris, W.C.V.: Empirically derived symptom checklist 90 subgroups of chronic pain patients: A cluster analysis. Journal of Behavioral Medicine, 11:(82), 147-158, 1988.

27. John, S.M., Traupe, H., Gutberlet, I. and Hellhammer, D.: Identification of andrologic patient groups by cluster analysis. Fertility and Sterility, 50:(6), 945-948, 1988.

28. Johnson, R.A. and Wichern, D.W.: Applied multivariate statistical analysis. Prentice Hall International Inc. New Jersey, 1988.

29. Kendall, M.: Multivariate analysis. Charles Griffin & Company Ltd., London and High Wycombe, 1980.

30. Kieser J.A. and Groeneveld, H.T.: Allocation and discrimination based on human odontometric data. American Journal of Physical Antropology, 79: 331-337, 1989.

31. Klastorin, T.D.: Assessing cluster analysis results. Journal of Marketing Research, 20: 92-98, 1983.

KAYNAKLAR DİZİNİ (devam ediyor)

32. Kristeller, J.L. and Robin, J.: Identifying eating patterns in male and female undergraduates using cluster Analysis. Addictive Behaviors, 14: 631-642, 1989.

33. Krzanowski, W.J. and Lai, Y.T.: A Criterion for determining the number of groups in a data set using sum-of squares clustering. Biometrics, 44: 23-34, 1988.

34. Kurtuluş, K.: Pazarlama araştırmaları. İşletme İktisadi Enstitüsü, 30. Yıl Yayınları No: 6, 1985.

35. Leach, C.: GRAN: A computer program for the cluster analysis of a repertory grid. British Journal of Clinical Psychology, 27: 173-174, 1988.

36. Lee, K.L.: Multivariate tests for clusters. Journal of The American Statsistical Associatiation, 74:(367), 708-714, 1979.

37. Ling, R.F.: A probability theory of cluster analysis. Journal of the American Statistical Association, 68:(341), 159-164, 1973.

KAYNAKLAR DİZİNİ (devam ediyor)

38. Lindsey, J.C. Herzberg, A.M. and Watts, D.G.: A method for cluster analysis based on projections and quantile-quantile plots. Biometrics, 43: 327-341, 1987.

39. Lipkus, A.H., Lenk, T.J., Chittur, K.K. and Gendreau, R.M.: Cluster analysis of protein fourier transform infrared spectra. Biopolymers, 27: 1831-1838, 1988.

40. Mardia, K.V., Kent, J.T. and Bibby, J.M.: Multivariate analysis. Academic Press, London, 1979.

41. Maronna, R. and Jacovkis, P.M. Multivariate clustering procedures with variable metrics. Biometrics, 30: 499-505, 1974.

42. Marriott, F.H.C.: Practical problems in a method of cluster analysis. Biometrics, 27: 501-514, 1971.

43. Marriott, F.H.C.: The interpretation of multiple observations. Academic Press, London, 1974.

KAYNAKLAR DİZİNİ (devam ediyor)

44. Mornon, I.P., Bissery, V., Gaborioud, C., Thomas, A., Ojasoo, T. and Raynaud, J.P.: Hydrophobic cluster analysis (HCA) of the hormone-binding domain of receptor proteins. J. Steroid. Biochem., 34:(1-6) 355-361, 1989.

45. Murtagh, F.: A probability theory of hierarchic clustering using random dendograms. J. Statist. Comput. Simul. 13: 145-157, 1983.

46. Norusis, M.J.: SPSS Inc., Chicago, 1984.

47. Overall, J.E. and Klett, J.C.: Applied multivariateanalysis. McGraw-Hill Book Company, New York, 1972

48. Özdamar, K: Yer-zaman kümeleri ile ilgili belli başlı yöntemleri doğal ve yapay veriler kullanılarak karşılaştırılmaları üzerine bir araştırma. Doçentlik Tezi, Anadolu Üniv., 1981 (yayınlanmamış).

49. Özdamar, K.: Düşük yoğunluklu hastalıkların yer-zaman kümelenmesi analizinde David-Barton beta yaklaşımının etkinliğinin türetilmiş verilerde araştırılması ve bir uygulama. Anadolu Tıp Dergisi, 5, 145-157, 1983.

50. Özdamar, K.: Hastalık olgularının incelenmesinde kümeleme çözümlemesinin kullanılması. T.C. Anadolu Üniversitesi Yayınları No:295, Tıp Fakültesi Yayınları No:25, Eskişehir, 1988.

51. Paykel, E.S.: Classification of depressed patients: a cluster analysis derived grouping. Brit. J. Psychiat., 118: 275-288, 1971.

52. Punj, G. and Stewart, D.W.: Cluster analysis in marketing research: rewiew and suggestions for application. Journal of Marketing Research, 20: 134-148, 1983.

53. Scott, A.J. and Knott, M.: A cluster analysis method for variance. Biometrics, 30: 507-512, 1974.

54. Scott, A.J. and Symons, M.J.: Clustering methods based on likelihood ratio criteria. Biometrics, 387-397, 1971.

55. Sethi, S.P.: Comparative cluster analysis for word markets. Journal of Marketing Research, 8: 348-354, 1971.

KAYNAKLAR DİZİNİ (devam ediyor)

56. Sneath, P.A.H. and Sokal, R.R.: Numerical taxonomy. W.H. Freeman and Company, San Francisso, 1973.

57. Sözer, E.E., Çelik, Y. and Kutsal, A.: Kümeleme çözümlemesinde çekirdek noktaların seçimi. İstatistik Sempozyumu 86, Orta Doğu Teknik Üniversitesi ve Ankara Üniversitesi, 15-17 Eylül, 1986.

58. Symons, M.J.:Clustering criteria and multivariate normal mixtures. Biometrics, 37: 35-43, 1981.

59. SYSTAT INC: SYSTAT tutorial, V 5.0, 1990.

60. Tryon, R.C. and Bailey, D.E.: Cluster analysis. McGraw Hill Book Company, New York 1970.

61. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of The American Statsitical Association, 236-244, 1963.

62. Wilkinson, L.: SYSTAT V 2.0 Manuel, Systat Inc., Evanston, 1984.

63. Vogt, W., Nagel, D. and Sator, H.: Cluster analysis in clinical chemistry: A model. John Wiley & Sons, NewYork, 1987.