

**UKUROVA NİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

Murat ERİŐOĐLU

DOKTORA TEZİ

**UZAKLIK LÜLERİNİN KÜMELEME ANALİZİNE OLAN
ETKİLERİNİN İNCELENMESİ VE GELİŐTİRİLMESİ**

İSTATİSTİK ANABİLİM DALI

ADANA, 2011

**UKUROVA NİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

DOKTORA TEZİ

Murat ERİŐOĐLU

**UZAKLIK LÜLERİNİN KÜMELEME ANALİZİNE OLAN
ETKİLERİNİN İNCELENMESİ VE GELİŐTİRİLMESİ**

İSTATİSTİK ANABİLİM DALI

ADANA, 2011

**ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**UZAKLIK ÖLÇÜLERİNİN KÜMELEME ANALİZİNE OLAN
ETKİLERİNİN İNCELENMESİ VE GELİŞTİRİLMESİ**

Murat ERİŞOĞLU

DOKTORA TEZİ

İSTATİSTİK ANABİLİM DALI

Bu Tez / / Tarihinde Aşağıdaki Jüri Üyeleri Tarafından
Oybirliği/Oyçokluğu ile Kabul Edilmiştir.

.....
Prof. Dr. Sadullah SAKALLIOĞLU
DANIŞMAN

.....
Prof. Dr. Hamza Erol
ÜYE

.....
Doç. Dr. Aşır GENÇ
ÜYE

.....
Prof. Dr. Selahattin KAÇIRANLAR
ÜYE

.....
Yrd. Doç. Dr. Tayfun SERVİ
ÜYE

Bu Tez Enstitümüz İstatistik Anabilim Dalında hazırlanmıştır.

Kod No:

Prof. Dr. İlhami YEĞİNGİL
Enstitü Müdürü

Bu Çalışma Ç. Ü. Araştırma Projeleri Birimi Tarafından Desteklenmiştir.
Proje No: FEF2008D15

Not: Bu tezde kullanılan özgün ve başka kaynaktan yapılan bildirişlerin, çizelge ve fotoğrafların
kaynak gösterilmeden kullanımı, 5846 sayılı Fikir ve Sanat Eserleri Kanunundaki hükümlere
tabidir.

ÖZ

DOKTORA TEZİ

UZAKLIK ÖLÇÜLERİNİN KÜMELEME ANALİZİNE OLAN ETKİLERİNİN İNCELENMESİ VE GELİŞTİRİLMESİ

Murat ERİŞOĞLU

**ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI**

Danışman : Prof. Dr. Sadullah SAKALLIOĞLU
Yıl: 2011, Sayfa: 199
Jüri : Prof. Dr. Sadullah SAKALLIOĞLU
: Prof.Dr. Selahattin KAÇIRANLAR
: Prof.Dr. Hamza EROL
: Doç. Dr. Aşır GENÇ
: Yrd. Doç. Dr. Tayfun SERVİ

Uzaklık ölçülerinin kümelemeye etkisinin incelendiği bu çalışmada farklı birçok disiplinde yaygın olarak kullanılan ve geniş bir literatüre sahip olan kümeleme analizinde yer alan tekniklerde uygun uzaklık ölçülerinin kullanılmasıyla, tekniklerini kümeleme performanslarının arttırılması amaçlanmıştır. Bu amaçla yeni önerilen iki uzaklık ölçüsü ve literatürde geniş bir kullanıma sahip mevcut beş uzaklık ölçüsünün kümeleme performansları hem farklı özellikler taşıyan gerçek veri setlerinde hem de simülasyon çalışmasıyla incelenmiştir. Çalışmada aşamalı kümeleme teknikleri içerisinde basitliği ve etkinliği açısından en fazla kullanılan ve bilinen k -ortalamlar tekniğinde uygun küme sayısının belirlenmesi ve küme başlangıç merkezlerinin seçimi için yeni algoritmalar önerilmiştir. Önerilen algoritmaların performansı literatürde yer alan mevcut algoritmalar ile karşılaştırılmış ve önerilen algoritmaların performansının daha iyi olduğu görülmüştür. Çalışma kapsamında ayrıca standartlaştırma ve kümeleme öncesi boyut indirgemenin kümelemeye olan etkisi simülasyon çalışması ile ortaya konmuştur.

Anahtar kelimeler: Kümeleme, Uzaklık, Küme sayısı, Boyut indirgeme, Başlangıç merkezi

ABSTRACT

PhD. THESIS

AN INVESTIGATION OF EFFECTS ON CLUSTER ANALYSIS OF DISTANCE MEASUREMENTS AND IMPROVEMENT

Murat ERİŞOĞLU

**ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF STATISTICS**

Supervisor : Prof. Dr. Sadullah SAKALLIOĞLU
Year: 2011, Page: 199
Jury : Prof. Dr. Sadullah SAKALLIOĞLU
: Prof.Dr. Selahattin KAÇIRANLAR
: Prof.Dr. Hamza EROL
: Assoc. Prof. Dr. Aşır GENÇ
: Asst. Prof. Dr. Tayfun SERVİ

Cluster analysis is a multivariate statistical analysis that widely used in many different disciplines and with a wide range of literature. In this study, the effects of the distance measurements on cluster analysis will be investigated. In this study, our goal is clustering techniques to increase performance when the use of appropriate distance measures in clustering. To study the effects on cluster analysis of distance measures used in both simulation and real data sets. k -means algorithm is the most well known and the fast methods in non-hierarchical cluster algorithms. Because of the simplicity of k -means algorithm, this algorithm is used in various fields. Also in this study, a new method is developed for choosing the number of clusters and an algorithm to compute initial cluster centers for k -means algorithm. Moreover in this study we also showed that the effectiveness of clustering high dimensional data using principal components instead of original variables and the effectiveness of clustering high dimensional data using standardized variables instead of original variables

Key Words: Clustering, Distance, Number of cluster, Reduction dimensionality, Initial cluster centers

TEŞEKKÜR

Bu tez çalışması sürecinde yoğun iş temposuna rağmen değerli bilgi ve görüşlerini benimle paylaşan danışmanım sayın Prof. Dr. Sadullah SAKALLIOĞLU'na; tez izleme komitesi üyeleri sayın Prof. Dr. Hamza EROL ve sayın Doç. Dr. Aşır GENÇ'e; hiçbir zaman yardımlarını esirgemeyen Çukurova Üniversitesi İstatistik Bölümü'nde görev yapan çalışma arkadaşlarıma teşekkürlerimi sunarım.

Ayrıca hayatım boyunca sesini duyarak bile huzur bulduğum anneme ve yaşam kaynağım biricik eşim Ülkü (GÖKAL) ERİŞOĞLU'na sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

SAYFA

ÖZ	I
ABSTRACT	II
TEŞEKKÜR	III
İÇİNDEKİLER	IV
TABLOLAR DİZİNİ	VII
ŞEKİLLER DİZİNİ	XII
1. GİRİŞ	1
2. ÖNCEKİ ÇALIŞMALAR	7
3. KÜMELEME ANALİZİ	23
3.1. Aşamalı Kümeleme Analizi Teknikleri	27
3.1.1. Tek Bağlantı Kümeleme Tekniği	29
3.1.2. Tam Bağlantı Kümeleme Tekniği	32
3.1.3. Ortalama Bağlantı Kümeleme Tekniği	34
3.1.4. Ağırlıklı Ortalama Bağlantı Kümeleme Tekniği	36
3.1.5. Merkezi Bağlantı Kümeleme Tekniği	37
3.1.6. Medyan Bağlantı Kümeleme Tekniği	40
3.1.7. Ward Bağlantı Kümeleme Tekniği	42
3.1.8. Esnek Beta Tekniği	44
3.2. Aşamalı Olmayan Kümeleme Teknikleri	45
3.2.1. k -Ortalamalar Kümeleme Algoritması	47
3.3. Kümelerin Değerlendirilmesi	55
3.3.1. Dışsal (External) Kriterler	55
3.3.1.1. Rand İndeksi	57
3.3.1.2. Jaccard Katsayısı	57
3.3.1.3. Fowlkes ve Mallows İndeksi	58
3.3.1.4. Γ İstatistiği	58
3.3.1.5. Düzeltilmiş Rand İndeksi	58
3.3.2. İçsel (Internal) Kriterler	59
3.3.2.1. Kofenetik (Cophenetic) Korelasyon Katsayısı	60

3.3.2.2. Wilks'in Lambda Test İstatistiği.....	60
3.3.3. Oransal (Relative) Kriterler	61
3.3.3.1. Calinski ve Harabasz İndeksi.....	62
3.3.3.2. Davies-Bouldin İndeksi	62
3.3.3.3. Dunn İndeksi.....	63
4. UZAKLIK ÖLÇÜLERİ	65
4.1. Uzaklık ve Özellikleri	65
4.2. Nicel Değişkenler İçin Uzaklık Ölçüleri	67
4.2.1. Manhattan (City-Block) Uzaklığı.....	67
4.2.2. Chebyshev Uzaklığı.....	69
4.2.3. Öklid (Euclidean) Uzaklığı.....	71
4.2.4. Standartlaştırılmış Öklid Uzaklığı	72
4.2.5. Minkowski Uzaklığı	73
4.2.6. Mahalanobis Uzaklığı.....	75
4.2.7. Nicel Veriler İçin Diğer Uzaklıklar	76
4.3. Önerilen Uzaklık Ölçüleri	79
4.3.1. Geometrik Uzaklık	79
4.3.2. Karma Uzaklık.....	80
5. BOYUT İNDİRGEMENİN KÜMELEMeye ETKİSİNİN İNCELENMESİ.....	83
5.1. Veri Setlerinin Üretimi.....	84
5.2. Temel Bileşenler Analizi.....	86
5.3. Simülasyon Çalışması	87
5.4. Uygulama	88
5.5. Sonuçlar.....	105
6. KÜME SAYISININ TAHMİNİNDE YENİ BİR YÖNTEM	107
6.1. Önemli Bazı Küme Sayısı Tahmin Yöntemleri	107
6.1.1. Marriott İndeksi	107
6.1.2. Varyans Açıklama Oran Kriteri.....	108
6.1.3. Üst Kuyruk Yöntemi	108
6.1.4. Silhouette İstatistiği	109
6.1.5. Gap İstatistiği.....	109

6.2. Küme Sayısı Tahmini İçin Önerilen Yöntem.....	111
6.3. Önerilen Yöntem ile Küme Sayısı Tahminleri.....	119
6.4. Önerilen Yöntemin Diğer Yöntemlerle Karşılaştırılması	122
6.5. Sonuçlar.....	126
7. KÜME BAŞLANGIÇ MERKEZLERİ İÇİN YENİ YÖNTEM.....	127
7.1. Küme Başlangıç Merkezlerinin Seçimi İçin Önerilen Yeni Algoritma	129
7.2. Karşılaştırma Kriterleri.....	132
7.3. Uygulama	133
7.4. Sonuçlar.....	139
8. UZAKLIK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ	141
8.1. Aşamalı Kümeleme Tekniklerinde Uzaklık Ölçülerinin Etkisi	141
8.1.1. Aşamalı Kümeleme Tekniklerinde Uzaklık Ölçülerinin Etkisinin Gerçek Veri Setleri Üzerinde İncelenmesi	141
8.1.2. Aşamalı Kümeleme Tekniklerinde Uzaklık Ölçülerinin Etkisinin Waveform Simülasyon Veri Setlerinde İncelenmesi.....	154
8.2. Aşamalı Olmayan Kümeleme Tekniklerinde Uzaklık Ölçülerinin Etkisi	164
8.2.1. k – Ortalamalar Tekniğinde Uzaklık Ölçülerinin Etkisinin Gerçek Veri Setleri Üzerinde İncelenmesi.....	164
8.2.2. k – Ortalamalar Tekniğinde Uzaklık Ölçülerinin Etkisinin Simülasyonla İncelenmesi	167
8.3. Sonuçlar.....	177
9. SONUÇLAR VE ÖNERİLER	179
KAYNAKLAR	185
ÖZGEÇMİŞ	199

Tablo 3.1. Eklemeli aşamalı kümeleme tekniklerinin gösteriminde kullanılacak örnek veri seti.....	29
Tablo 3.2. Birimler arasındaki Öklid uzaklığına göre elde edilen uzaklık matrisi....	29
Tablo 3.3. Tek bağlantı kümeleme tekniğine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri	31
Tablo 3.4. Tam bağlantı kümeleme tekniğine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri	33
Tablo 3.5. Ortalama bağlantı kümeleme tekniğine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri.....	35
Tablo 3.6. Ağırlıklı ortalama bağlantı kümeleme tekniğine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri.....	36
Tablo 3.7. Merkezi bağlantı kümeleme tekniğine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri.....	39
Tablo 3.8. Medyan bağlantı kümeleme tekniğine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri.....	41
Tablo 3.9. Ward bağlantı kümeleme tekniğine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri	43
Tablo 3.10. Lance ve Williams'ın önerdiği genel formüldeki parametrelerin bağlantı tekniğine göre değerleri.....	45
Tablo 3.11. k -ortalamalar algoritmasının uygulaması için oluşturulan örnek veri seti	50
Tablo 3.12. Örnek veri setinin k -ortalamalar algoritması ile kümelenmesinde I. adımda gerçekleştirilen hesaplamalar ve gözlemlerin kümelere atanması	51
Tablo 3.13. Örnek veri setinin k -ortalamalar algoritması ile kümelenmesinde II. adımda gerçekleştirilen hesaplamalar ve gözlemlerin kümelere atanması	52

Tablo 3.14.Örnek veri setinin k -ortalamalar algoritması ile kümelenmesinde III. adımda gerçekleştirilen hesaplamalar ve gözlemlerin kümelere atanması	53
Tablo 3.15.Şekil 3.19'daki örnek C ve P kümelenmesi için hesaplanan a,b,c ve d değerleri ve açıklamaları.....	56
Tablo 5.1. Veri setlerinin üretiminde göz önünde tutulacak durumlar.....	87
Tablo 5.2. KA uygulamasının gerçekleştirileceği veri matrisleri.....	87
Tablo 5.3. Durum 1 için doğru sınıflandırma yüzdesine ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin birikimli varyans açıklama oranları	89
Tablo 5.4. Durum 2 için doğru sınıflandırma yüzdesine ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin ortalama birikimli varyans açıklama oranları	90
Tablo 5.5. Durum 3 için doğru sınıflandırma yüzdesine ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin ortalama birikimli varyans açıklama oranları	91
Tablo 5.6. Durum 4 için doğru sınıflandırma yüzdesine ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin ortalama birikimli varyans açıklama oranları	92
Tablo 5.7. Durum 1 için eşleştirme katsayısına ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin ortalama birikimli varyans açıklama oranları	98
Tablo 5.8. Durum 2 için eşleştirme katsayısına ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin ortalama birikimli varyans açıklama oranları	99
Tablo 5.9. Durum 3 için eşleştirme katsayısına ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin ortalama birikimli varyans açıklama oranları	100
Tablo 5.10.Durum 4 için eşleştirme katsayısına ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin ortalama birikimli varyans açıklama oranları	101

Tablo 6.1. İncelenen veri setlerinin örneklem büyüklükleri, değişken sayıları ve küme sayıları	119
Tablo 6.2. Veri setleri için iç içe kümelemenin olmadığı varsayımı altında gerçek küme sayıları(k),özdeğerler, temsil değerlerinin standart sapmaları(σ_{sv}), iç içe kümelemenin olmadığı varsayımı altında özel değer $\gamma = 3.0$ alınarak elde edilen beklenen minimum farklar (EMD) ve ardışık minimum fark yöntemi ile gerçekleştirilen küme sayısı tahminleri ($\hat{k}$)	120
Tablo 6.3. Veri setleri için küme sayıları(k), özdeğerler, temsil değerlerinin standart sapmaları(σ_{sv}), $\gamma = 2.5$ alınarak elde edilen beklenen minimum farklar (EMD) ve ardışık minimum fark yöntemi ile gerçekleştirilen küme sayısı tahminleri ($\hat{k}$).....	121
Tablo 6.4. Veri setleri için küme sayıları(k), özdeğerler, temsil değerlerinin standart sapmaları(σ_{sv}), $\gamma = 1.2$ alınarak elde edilen beklenen minimum farklar (EMD) ve ardışık minimum fark yöntemi ile gerçekleştirilen küme sayısı tahminleri ($\hat{k}$).....	122
Tablo 6.5. İncelenen veri setlerinde Öklid uzaklığının kullanıldığı “Tek bağlantı” ve “Ward” kümeleme algoritmalarında farklı c sabiti değerleri için üst kuyruk yöntemi ile elde edilen küme sayısı tahminleri	123
Tablo 6.6. Marriott indeksi, varyans açıklama oranı ve Silhouette indeksine göre gerçekleştirilen küme sayısı tahminleri.....	124
Tablo 6.7. İncelenen veri setlerinde karesel Öklid uzaklığının kullanıldığı “Tam bağlantı” ve “Ward” kümeleme algoritmalarında Gap istatistiği ile elde edilen küme sayısı tahminleri.....	125
Tablo 7.1. İris verisindeki değişkenler için aritmetik ortalama, standart sapma ve değişim katsayısı değerleri	134
Tablo 7.2. Taç yaprağı genişliği değişkeni ile diğer değişkenler arasındaki korelasyonlar	134

Tablo 7.3. Önerilen algoritma ile rassal başlangıçlı k -ortalamlar algoritmasının hatalı sınıflandırma yüzdesi, Rand indeks ve Wilks'ın lambda istatistiğine göre karşılaştırılması.....	138
Tablo 8.1. Aşamalı kümeleme tekniklerinde uzaklık fonksiyonlarının etkisinin inceleneceği veri setlerinin özellikleri	142
Tablo 8.2. İlgilenilen veri setlerinde farklı aşamalı kümeleme tekniği ve farklı uzaklık fonksiyonlarına göre elde edilen kofenetik korelasyon katsayı değerleri	145
Tablo 8.3. İlgilenilen 13 gerçek veri setinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık fonksiyonları kullanılarak kümelenmesi sonrası elde edilen kofenetik korelasyon katsayı değerlerinin ortalaması ve standart hatası	148
Tablo 8.4. İlgilenilen veri setlerinde farklı aşamalı kümeleme tekniği ve farklı uzaklık fonksiyonlarına göre elde edilen Wilks' lambda test istatistiği değerleri.....	149
Tablo 8.5. İlgilenilen 13 gerçek veri setinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık fonksiyonları kullanılarak kümelenmesi sonrası elde edilen Wilks' lambda test istatistiği değerlerinin ortalaması ve standart hatası	152
Tablo 8.6. Wafeform simülasyon veri setlerinin ($n = 75$) farklı aşamalı kümeleme tekniklerinde farklı uzaklık fonksiyonları kullanılarak kümelenmesi sonrası elde edilen kofenetik korelasyon katsayısı değerlerinin ortalama ve standart hataları	156
Tablo 8.7. Wafeform simülasyon veri setlerinin ($n = 150$) farklı aşamalı kümeleme tekniklerinde farklı uzaklık fonksiyonları kullanılarak kümelenmesi sonrası elde edilen kofenetik korelasyon katsayısı değerlerinin ortalama ve standart hataları	157
Tablo 8.8. Wafeform simülasyon veri setlerinin ($n = 300$) farklı aşamalı kümeleme tekniklerinde farklı uzaklık fonksiyonları kullanılarak kümelenmesi sonrası elde edilen kofenetik korelasyon katsayısı değerlerinin ortalama ve standart hataları	158

Tablo 8.9. Wafeform simülasyon veri setlerinin ($n = 75$) farklı aşamalı kümeleme tekniklerinde farklı uzaklık fonksiyonları kullanılarak kümelenmesi sonrası elde edilen Wilks' lambda test istatistiği değerlerinin ortalama ve standart hataları	160
Tablo 8.10.Wafeform simülasyon veri setlerinin ($n = 150$) farklı aşamalı kümeleme tekniklerinde farklı uzaklık fonksiyonları kullanılarak kümelenmesi sonrası elde edilen Wilks' lambda test istatistiği değerlerinin ortalama ve standart hataları	161
Tablo 8.11.Wafeform simülasyon veri setlerinin ($n = 300$) farklı aşamalı kümeleme tekniklerinde farklı uzaklık fonksiyonları kullanılarak kümelenmesi sonrası elde edilen Wilks' lambda test istatistiği değerlerinin ortalama ve standart hataları	162
Tablo 8.12.Farklı yapılardaki gerçek veri setlerinin k -ortalamalar tekniğinde farklı uzaklık fonksiyonları kullanılarak kümelenmesi sonrası elde edilen Wilks' lambda test istatistiği değerleri.....	165
Tablo 8.13.Simülasyon çalışmasında çok değişkenli normal dağılımdan rassal olarak üretilecek örneklemelerin oluşturulmasında kullanılacak kovaryans matrisleri	168
Tablo 8.14.Durum I için ilgilenilen 7 uzaklık fonksiyonu ile elde edilen Wilks'in lambda test istatistiği değerlerine ait ortalama ve standart hatalar.....	169
Tablo 8.15.Durum II için ilgilenilen 7 uzaklık fonksiyonu ile elde edilen Wilks'in lambda test istatistiği değerlerine ait ortalama ve standart hatalar.....	171
Tablo 8.16.Durum III için ilgilenilen 7 uzaklık fonksiyonu ile elde edilen Wilks'in lambda test istatistiği değerlerine ait ortalama ve standart hatalar.....	173
Tablo 8.17.Durum IV için ilgilenilen 7 uzaklık fonksiyonu ile elde edilen Wilks'in lambda test istatistiği değerlerine ait ortalama ve standart hatalar.....	175

Şekil 1.1. Yıllara göre içerisinde “kümeleme analizi” kavramı geçen çalışma sayıları	3
Şekil 3.1. 16 kartın kümelenmesinden elde edilebilecek farklı bazı kümeler	25
Şekil 3.2. Kümeleme analizi tekniklerinin sınıflandırılması	26
Şekil 3.3. Aşamalı kümelemenin ağaç diyagramı(dendrogram) ile grafiksel gösterimi	27
Şekil 3.4. Eklemeli aşamalı kümeleme algoritmalarının akış şeması	28
Şekil 3.5. Tek bağlantı kümeleme tekniğine göre iki küme arasındaki uzaklığın belirlenmesi	30
Şekil 3.6. Tek bağlantı kümeleme tekniği ile birimlerin bağlanmasının (a) iki boyutlu düzlemde ve (b) ağaç grafiği ile gösterimi	31
Şekil 3.7. Tam bağlantı kümeleme tekniğine göre iki küme arasındaki uzaklığın belirlenmesi	32
Şekil 3.8. Tam bağlantı kümeleme tekniği ile birimlerin bağlanmasının (a) iki boyutlu düzlemde ve (b) ağaç grafiği ile gösterimi	33
Şekil 3.9. Ortalama bağlantı kümeleme tekniğine göre iki küme arasındaki uzaklığın belirlenmesi	34
Şekil 3.10. Ortalama bağlantı kümeleme tekniği ile birimlerin bağlanmasının (a) iki boyutlu düzlemde ve (b) ağaç grafiği ile gösterimi	35
Şekil 3.11. Ağırlıklı ortalama bağlantı kümeleme tekniği ile birimlerin bağlanmasının (a) iki boyutlu düzlemde ve (b) ağaç grafiği ile gösterimi	37
Şekil 3.12. Merkezi bağlantı kümeleme tekniğine göre iki küme arasındaki uzaklığın belirlenmesi	38
Şekil 3.13. Merkezi bağlantı kümeleme tekniği ile birimlerin bağlanmasının (a) iki boyutlu düzlemde ve (b) ağaç grafiği ile gösterimi	39
Şekil 3.14. Medyan bağlantı kümeleme tekniği ile birimlerin bağlanmasının (a) iki boyutlu düzlemde ve (b) ağaç grafiği ile gösterimi	41

Şekil 3.15. Ward bağlantı kümeleme tekniği ile birimlerin bağlanması (a) iki boyutlu düzlemde ve (b) ağaç grafiği ile gösterimi	43
Şekil 3.16. k -ortalamalar algoritmasının akış şeması	48
Şekil 3.17. k -ortalamalar algoritmasında küme başlangıç merkezlerinin seçiminin kümeleme etkisi.....	49
Şekil 3.18. Örnek veri setinin k -ortalamalar algoritması ile kümelenmesinde küme merkezleri ve küme üyeliklerinin değişiminin gösterimi.....	54
Şekil 3.19. C ve P kümelenmesi için gözlem çiftlerinin muhtemel dört durumu için örnek gösterim.....	56
Şekil 4.1. Manhattan uzaklığının iki boyutlu düzlemde hesaplanması	68
Şekil 4.2. Manhattan uzaklık fonksiyonunun üç boyutlu uzayda gösterimi	69
Şekil 4.3. Chebsyhev uzaklığının iki boyutlu düzlemde hesaplanması	70
Şekil 4.4. Chebsyhev uzaklık fonksiyonunun üç boyutlu uzayda gösterimi.....	70
Şekil 4.5. Öklid uzaklığının iki boyutlu düzlemde hesaplanması	71
Şekil 4.6. Öklid uzaklık fonksiyonunun üç boyutlu uzayda gösterimi	72
Şekil 4.7. İki boyutlu düzlemde orijinden bir birim uzaklıktaki noktalardan oluşan yüzeyde L_1, L_2 ve L_∞ metrikleri.	74
Şekil 4.8. Farklı λ değerleri Minkowski uzaklık fonksiyonunun üç boyutlu uzayda gösterimi	74
Şekil 4.9. Gözlemlerin konumlarının Canberra uzaklığındaki etkisine ait gösterimi.....	78
Şekil 4.10. Geometrik uzaklık fonksiyonunun üç boyutlu uzayda grafiği.....	79
Şekil 5.1. Farklı dört durum için ortalama doğru sınıflandırma yüzdesi.....	93
Şekil 5.2. t . tekrarda ham veri ile elde edilen doğru sınıflandırma yüzdesi ile standartlaştırılmış veri ile doğru sınıflandırma yüzdesi arasındaki farkların değişimi	94
Şekil 5.3. Ham veri ile kümelemenin standartlaştırılmış veri ile kümelemeyle doğru sınıflandırma yüzdesi bakımından karşılaştırılması	95
Şekil 5.4. Kümelemede kullanılan bileşen sayısına ait birikimli varyans açıklama oranları ile kümelemede kullanılan bileşen sayısına göre elde edilen doğru sınıflandırma yüzdesi arasındaki korelasyonlar.....	96

Şekil 5.5. Farklı dört durum için ortalama eşleştirme katsayısı değerleri.....	102
Şekil 5.6. Kümelemelerden elde edilen eşleştirme katsayıları ile kümelemelerde kullanılan bileşenlerin birikimli varyans açıklama oranları arasındaki saçılım grafikleri	103
Şekil 5.7. Ortalama eşleştirme katsayıları ile ortalama birikimli varyans açıklama oranları arasındaki saçılım grafikleri	104
Şekil 5.8. Kümelemede orijinal değişkenler yerine ilk sekiz temel bileşenin kullanılmasıyla elde edilen eşleştirme katsayılarının değişimi.....	104
Şekil 6.1. Ardışık minimum fark yöntemi ile ilk beş temsil değerinin incelenmesi.....	117
Şekil 6.2. Ardışık farklar yöntemi ile küme sayısı tahmini için yazılmış Matlab kodu.....	118
Şekil 6.3. İris verisinin taç yaprağı genişliği ve çanak yaprağı genişliğine göre saçılımı	121
Şekil 7.1. Birinci aday küme başlangıç merkezinin seçimi.....	130
Şekil 7.2. İkinci aday küme başlangıç merkezinin seçimi	131
Şekil 7.3. Üçüncü aday küme başlangıç merkezinin seçimi	132
Şekil 7.4. Taç yaprağı genişliği ile çanak yaprağı genişliğine göre saçılım	135
Şekil 7.5. İris verisinde önerilen yöntemle birinci aday küme başlangıç merkezinin seçimi	136
Şekil 7.6. İris verisinde önerilen yöntemle ikinci aday küme başlangıç merkezinin seçimi	136
Şekil 7.7. İris verisinde önerilen yöntemle üçüncü aday küme başlangıç merkezinin seçimi	137
Şekil 8.1. Şarap verisinin tam bağlantı kümeleme tekniğinde farklı uzaklık fonksiyonlarına göre elde edilen ağaç grafikleri.....	144
Şekil 8.2. Şarap verisinin ağırlıklı ortalama bağlantı kümeleme tekniğinde farklı uzaklık fonksiyonlarına göre elde edilen ağaç grafikleri	144

Şekil 8.3. Gerçek veri setlerinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık fonksiyonlarına göre kümelenmesi ile elde edilen ortalama kofenetik korelasyon katsayısı ve ortalama Wilks' lambda test istatistiğine göre değerlendirilmesi	153
Şekil 8.4. Rassal olarak üretilen 300 birimlik Waveform simülasyon verisinin ortalama bağlantı kümeleme tekniğinde farklı uzaklık fonksiyonlarına göre elde edilen ağaç grafikleri.....	155
Şekil 8.5. Rassal olarak farklı örneklem hacimlerinde üretilen Waveform simülasyon verilerinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık fonksiyonlarına göre kümelenmesi ile elde edilen ortalama kofenetik korelasyon katsayısına göre değerlendirilmesi	159
Şekil 8.6. Rassal olarak farklı örneklem hacimlerinde üretilen Waveform simülasyon verilerinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık fonksiyonlarına göre kümelenmesi ile elde edilen ortalama Wilks' lambda test istatistiğine göre değerlendirilmesi	163
Şekil 8.7. Farklı yapılarıdaki gerçek veri setlerinin k -ortalamalar tekniğinde farklı uzaklık fonksiyonları kullanılarak kümelenmesi sonrası elde edilen Wilks' lambda test istatistiğine göre değerlendirilmesi	166
Şekil 8.8. Durum I'de ilgilenilen 7 farklı uzaklık fonksiyonuna göre elde edilen ortalama Wilks' lambda değerlerinin değişimi	170
Şekil 8.9. Durum II'de ilgilenilen 7 farklı uzaklık fonksiyonuna göre elde edilen ortalama Wilks' lambda değerlerinin değişimi	172
Şekil 8.10. Durum III'de ilgilenilen 7 farklı uzaklık fonksiyonuna göre elde edilen ortalama Wilks' lambda değerlerinin değişimi	174
Şekil 8.11. Durum IV'de ilgilenilen 7 farklı uzaklık fonksiyonuna göre elde edilen ortalama Wilks' lambda değerlerinin değişimi	176

1. GİRİŞ

Küreselleşme ve hızla gelişen teknolojinin yaşamımızdaki etkileri iyice belirgin hale gelmiştir. Bu etki yaşam alanımızda yer alan olguların etkileşimini de arttırdığından gerçekleştirilen bilimsel çalışmaların çoğunda göz önüne alınması gereken değişken sayısını da arttırmıştır. Ayrıca bu gelişmelere paralel olarak, bilişim sektöründeki gelişmeler, iletişim alanında elde edilen kazanımlar günümüzde hızlı bir veri akışının oluşumunu sağlamıştır.

Veri boyutundaki bu artış beraberinde hangi veriyi ve bu verinin nasıl kullanılması gerektiği problemini ortaya çıkarmıştır. Yapılan çalışmalardan optimum sonuçlar alınabilmesi için bu veri yığınının işlenmesi ve görünenin arkasına bakabilmek gerekmektedir. Verinin işlenmesinde, veriden bilgi elde edilmesinde kullanılan en önemli araçlar istatistiksel tekniklerdir.

Yaşanılan bu gelişmeler doğrultusunda istatistiksel tekniklerde çeşitlilik göstermiş ve karşılaşılan ihtiyaçların giderilmesi uğraşı istatistiksel tekniklerinde büyük gelişimler elde edilmesine imkân tanımıştır. Verinin özetlenmesinde kullanılan istatistiksel yaklaşımlar genel olarak iki kategori altında ele alınabilir. Bu yaklaşımlardan biri verinin elde edildiği birimlerin benzerliklerine göre gruplandırılmasıdır. Diğer yaklaşım ise, ele alınan değişkenler arasındaki ilişki yapısını göz önünde bulundurarak boyutun indirgenmesidir.

Benzer birimlerin bir araya getirilmesi yaklaşımını kullanan tekniklerden biri kümeleme analizidir. Kümeleme analizi genel olarak, analiz öncesi doğal bir kümelenmesi olmayan birimlerin göz önüne alınan değişkenlere göre oluşturulan benzerliklerine göre bir araya getirilmesi şeklinde tanımlanmaktadır. Kümelemedeki amaç, benzer olarak bir araya getirilen birimlerden oluşan kümelerde küme içi değişimin minimum, kümeler arası değişimin ise maksimum olmasıdır.

Kümeleme, günümüzde veri analizinde en yaygın kullanılan ve önemi gün geçtikçe artan çok önemli bir istatistiksel tekniktir. Bu önemli istatistiksel araç farklı birçok bilim dalında kullanılmaktadır. Kümeleme analizinin uygulama alanlarına örnek olarak aşağıdaki alanlar verilebilir:

Mühendislik (Yapay zeka, makine öğrenmesi, görüntü tanıma, makine mühendisliği, elektrik mühendisliği): Biyometrik tanıma, radar sinyal analizi, bilgi sıkıştırma, konuşma tanıma ve gürültü giderme mühendislik alanında kümeleme analizinin kullanıldığı tipik uygulamalardır.

Bilgisayar bilimleri: Bilgisayar bilimlerinde web madenciliği, mekânsal veri analizi, bilgi erişimi, metinsel belge toplama ve görüntü bölümlendirme uygulamaları kümeleme analizinin kullanıldığı örnekler olarak verilebilir.

Yaşam ve tıp bilimleri (genetik, biyoloji, mikrobiyoloji, paleontoloji, psikiyatri, klinik, evrime, patoloji): Canlıların sınıflandırılması (taksonomi), gen ve protein fonksiyonlarının tanımlanması, hastalığın tanı ve tedavisi gibi birçok alanda kümeleme analizi kullanılmaktadır. Kümeleme analizinin en yaygın kullanıldığı bu alanlarda kümeleme analizi ileride de önemli bir araç olmaya devam edecektir.

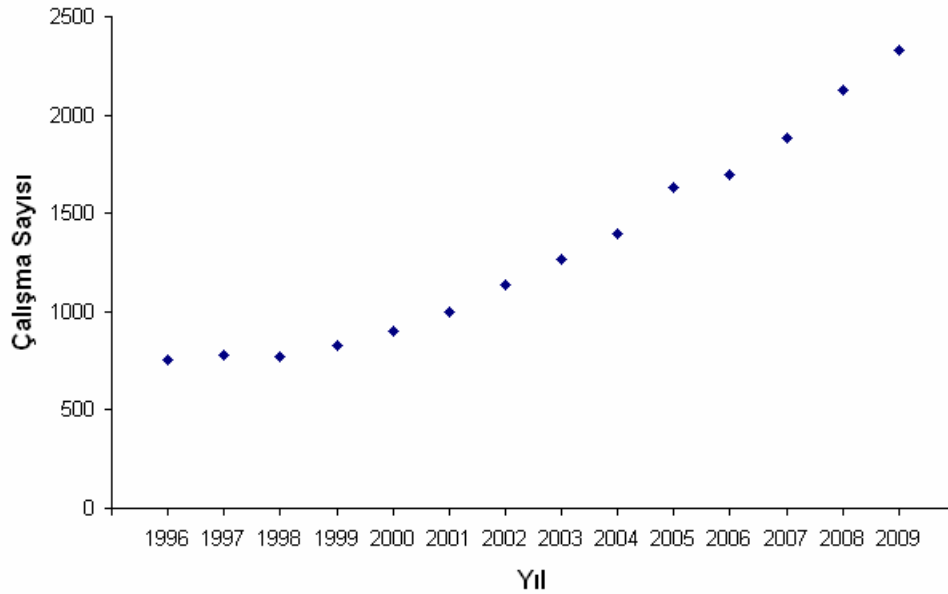
Astronomi ve yer bilimleri (coğrafya, jeoloji, uzaktan algılama): Kümeleme, analizi; yıldızlar ve gezegenlerin sınıflandırılması, yer şekillerinin tanımlanması, şehir planlaması, nehir ve dağ sistemleri ile ilgili çalışmalar da kullanılmaktadır.

Sosyal bilimler (sosyoloji, psikoloji, arkeoloji, antropoloji, eğitim): Davranış biçimi analizi, farklı kültürler arasındaki ilişkilerin belirlenmesi, dilin evrimsel tarihi, sosyal ağlar, arkeolojik bulgu ve artefakt sınıflandırma analizi ve suç psikolojisi çalışmaları gibi farklı uygulamalarda kümeleme analizi kullanılmaktadır.

Ekonomi (pazarlama, işletme): Firmaların gruplandırılması, müşteri özelliklerinin ve satın alma alışkanlıklarının tanımlanması, hisse senetlerinin trend analizinde kümeleme analizinden yararlanılmaktadır.

Farklı birçok disiplinde yaygın olarak kullanılan kümeleme analizinin literatürde ki ilk örneği, 1854 yılında Londra'daki bir kolera salgını sırasında John Snow adında bir görevlinin yaptığı çalışma gösterilmektedir (Gilbert 1958). John Snow, kolera salgını sırasında kendisine rapor edilen vakaları üzerinde işaretlemek için özel bir harita oluşturmuş ve kullanmıştır. Harita üzerinde işaretlenen vakaları inceleyen görevli, merkezdeki bir cadde üzerindeki yerleşim yerlerinde hastalığın yoğunlaştığını ve diğer bölgelere buradan dağıldığı gerçeğini tespit etmiştir. Veri kümeleme analizi kavramı olarak ise ilk kez Tryon (1939) tarafından kullanılmıştır. 1940'lı yılların sonunda ve 1950'li yıllarda verilerin aşamalı olarak gruplanması

üzerinde çalışılmış, 1960’lı yıllarda ise aşamalı olmayan kümeleme yöntemleri geliştirilmeye başlanmıştır. Kümeleme analizi oldukça geniş ve sürekli artan bir literatüre sahiptir. Thomson Reuters tarafından hazırlanan SCI, SCI-Expanded, SSCI ve AHCI indekslerince taranan çalışmalara ulaşılabilen ISI Web of KnowledgeSM (<http://apps.isiknowledge.com>) arama sayfasında arama kutucuğuna “cluster analysis” terimi yazılıp yıllara göre arama yapılarak 1996 ile 2009 yılları arasında yıllara göre içerisinde kümeleme analizi kavramı geçen çalışma sayıları Şekil 1.1’de verilmiştir. Yıllara göre çalışma sayıları incelendiğinde içerisinde kümeleme analizi kavramı geçen çalışma sayısının giderek arttığı görülmektedir. 1996 yılında kümeleme analizi kavramı geçen çalışma sayısı 756 iken bu sayı 2009 yılında 2332 olmuştur. Bu çalışmalar 3000’in üzerinde farklı dergilerde yayınlanmıştır.



Şekil 1.1. Yıllara göre içerisinde “kümeleme analizi” kavramı geçen çalışma sayıları

Kümeleme analizinde gerçekte bir eniyileme elde edilmeye çalışıldığından bu alanda birçok farklı yaklaşım geliştirilmiş ve hala geliştirilmeye çalışılmaktadır. Kümeleme analizi içerisinde yer alan tekniklerin farklı sınıflandırmaları bulunmakla beraber en yaygın olarak benzer birimleri bir araya getirmede izledikleri yönteme göre teknikler aşamalı ve aşamalı olmayan kümeleme teknikleri olarak iki kategori altında toplanabilir.

Kümeleme analizi teknikleri ister aşamalı, isterse aşamalı olmayan teknikler olsun, kümelemede en temel adım benzerliklerin belirlenmesidir. Birimler arasındaki benzerliğin belirlenmesinde genelde uzaklık ölçüleri kullanılmaktadır. İki birim arasındaki uzaklık ne kadar küçükse o iki birim o kadar benzerdir. Kümeleme analizi içerisindeki hangi teknik kullanılırsa kullanılsın amaç benzer birimleri bir araya getirerek küme içi değişimi minimum, kümeler arası değişimi ise maksimum yapmak olduğundan kümeleme performansı seçilen uzaklık ölçüsü ile doğrudan ilişkilidir. Kümelenen veri setlerinde kümelemeye tabi tutulacak birim sayısı, göz önüne alınan değişken sayısı, değişkenler arasındaki ilişki yapısı, küme sayısı, ölçek türü, terim büyüklükleri gibi birçok özellik farklılık gösterebileceğinden herhangi bir kümeleme tekniğinin her zaman herhangi bir uzaklık türü ile en iyi sonucu vereceği şeklinde bir genelleme yapılamaz. Bununla birlikte belli özelliklere sahip veri setleri için hangi kümeleme tekniğinde hangi uzaklık ölçüsünün kullanılabileceği önerilebilir. Bu çalışmada, mevcut uzaklık ölçülerinin kümeleme üzerine olan etkilerini incelenecektir. Uzaklık ölçülerinin kümeleme üzerine etkisi incelenirken veri tipi, değişken sayısı, değişkenler arasındaki ilişki yapısı ve gözlem sayıları gibi özellikler göz önünde bulundurularak gerçekleştirilen simülasyonlar sonucunda belirlenen performans kriterlerine göre en iyi olan kombinasyonlar belirlenecektir. Bu çalışma kapsamında mevcut yaygın kullanılan uzaklık ölçüleri yanında önerilecek olan alternatif uzaklık ölçülerinin de kümeleme performansı incelenecektir.

Kümeleme analizinde seçilecek uzaklık ölçüsünün yanında, analiz öncesi boyut indirgemenin kümelemeye olan etkisi, standartlaştırmanın kümeye olan etkisi, uygun küme sayısının belirlenmesi ve k -ortalamlar algoritmasında küme başlangıç merkezlerinin seçimi gibi temel problemler hala araştırmacılar tarafından incelenen konuları oluşturmaktadır.

Çalışma kapsamında kümeleme öncesi boyut indirgemenin ve standartlaştırmanın kümelemeye olan etkisi simülasyon ile incelenecektir. Çok değişkenli normal dağılımların karmasından oluşturulan ve gerçek küme üyelikleri bilinen veri setleri farklı boyut sayılarına göre kümelendirilecektir. Ayrıca standartlaşmanın etkisini incelemek için her veri seti bir kerede standartlaştırılarak kümelendirilecektir. Elde edilen kümeleme sonuçları doğru sınıflandırma yüzdesi,

ham veri ile kümeleme sonuçlarına uyum bakımından karşılaştırılarak boyut indirgemenin ve standartlaştırmanın kümelemeye olan etkisi ortaya konmaya çalışılacaktır.

Kümeleme analizinde en temel problemlerden biri olan uygun küme sayısının belirlenmesi için bu çalışmada yeni bir yöntem önerilmiştir. Literatürde küme sayısının belirlenmesi için önerilen teknikler deneme-yanılma yöntemine dayanan tekniklerdir. Dolayısıyla bu teknikler ile önerilen küme sayıları seçilen kümeleme algoritması ve seçilen uzaklık ölçüsüne göre farklılık gösterebilmektedir. Önerilen teknik, küme sayısını belirlemede herhangi bir kümeleme analizine ihtiyaç duymadığı için kümeleme algoritmasından ve uzaklık ölçülerinden etkilenmemektedir. Küme sayısı için önerilecek tekniğin literatürde ki teknikler karşısındaki performansı, kümeleme analizi ile ilgili çalışmalarda yaygın bir kullanıma sahip ve çok iyi bilinen gerçek veri setleri kullanılarak gösterilecektir.

Aşamalı kümeleme teknikleri içerisinde en yaygın kullanılan k – ortalamalar tekniğinde küme başlangıç merkezlerinin oluşturulması kümeleme sonucunu doğrudan etkileyen faktörlerden biridir. Küme başlangıç merkezlerinin rassal olarak seçilmesi farklı kümeleme sonuçları ile karşılaşma riskinden dolayı pek tercih edilmemektedir. Bununla birlikte literatürde küme başlangıç merkezlerinin seçimine yönelik çok sayıda algoritma bulunmaktadır. Bu algoritmaların bazıları çok karmaşık algoritmalara sahip iken basit algoritmalarda yeterince etkin bir kümeleme performansına sahip değildir. Bu çalışmada k – ortalamalar tekniğinde küme başlangıç merkezlerinin oluşturulması için oldukça basit ve kümeleme performansı yüksek bir algoritma oluşturulmuştur.

Çalışmadaki genel amaç, kullanım alanı yaygın olan kümeleme analizinde eniyilemeye katkı sağlayarak kümeleme tekniklerinin kümeleme performansını geliştirmektir. Böylelikle analizin kullanıldığı çalışmalarda çalışma sonucunun etkinliğinin artırılması çalışmanın temel hedefini oluşturmaktadır.

Çalışma kapsamında yapılan uygulamaların özetlendiği ve çalışmanın amacı ile öneminin belirtildiği bu bölümün ardından, ikinci bölümde kümeleme analizi ile ilgili önceki çalışmalar verilecektir. Çalışmanın üçüncü bölümünde kümeleme kavramı, kümeleme teknikleri, geçerlilik kriterleri incelenecektir. Çalışmanın

dördüncü bölümünde kümeleme analizinde uzaklık ölçüleri ve özellikleri anlatılacaktır. Çalışmanın beşinci bölümünde yüksek boyutlu verilerin kümelenmesinde boyut indirgemenin ve standartlaştırmanın etkisi incelenecektir. Çalışmanın altıncı bölümünde küme sayısının tahmini için yeni bir yöntem önerilecektir. Çalışmanın yedinci bölümünde k -ortalamalar tekniği için küme başlangıç merkezlerinin seçimi ile ilgili yeni bir algoritma verilecektir. Çalışmanın sekizinci bölümünde aşamalı kümeleme tekniklerinde ve k -ortalamalar tekniğinde uzaklık ölçülerinin kümelemeye olan etkisi hem gerçek veri setleri kullanılarak, hem de simülasyon verileri kullanılarak incelenecektir. Sonuç bölümünde kümeleme performansını eniyilemek için elde edilen bulgular değerlendirilerek önerilerde bulunulacaktır.

2. ÖNCEKİ ÇALIŞMALAR

İnsanoğlu var olduğundan bu yana etrafında bulunan nesneleri bazı özelliklerine göre sınıflara ayırma eğiliminde olmuştur. Örneğin bitkileri yenilir yenmez diye ilkel çağlarda sınıflayan insanoğlu, gün geçtikçe sınıflandırmanın amacını ve kapsamını değiştirmiştir. Birimlerin sayısı arttıkça birimleri sınıflandırmak daha da zorlaşmış ve yeni teknikler bulmayı gerektirmiştir. Bu gereksinim sonucu kümeleme analizi kavramı ortaya çıkmıştır.

Veri kümeleme analizi kavram olarak ilk kez Tryon (1939) tarafından kullanılmıştır. Clements (1954), antropolojik verinin kümelenmesi çalışmasında veri kümeleme analizi kavramını kullanmıştır. Cox (1957) ve Fisher (1958), kümeleme analizi kavramını gruplama adını verdikleri çalışmalarında kullanmışlardır. Sokal ve Sneath (1963), sayısal sınıflandırmanın prensipleri adını verdikleri kitaplarında kümeleme analizini anlatmışlardır.

Kümeleme analizi ile ilgili literatür incelendiğinde, aşamalı kümeleme teknikleri ile ilgili ilk çalışmaların 1940'lı yılların sonu ve 1950'li yıllarda olduğu görülmektedir. Aşamalı olmayan kümeleme analizi teknikleri ile ilgili ilk çalışmalar ise bu çalışmalardan sonra gerçekleştirilmiştir.

Sorensen (1948), benzer birimlerin birleştirilmesi ile oluşturulan yeni kümeye olan uzaklıkların hesaplanmasında maksimum uzaklığın kullanıldığı tam bağlantı kümeleme tekniğini önermiştir. Aşamalı kümeleme teknikleri içerisinde en fazla bilinen tek bağlantı kümeleme tekniği ilk olarak Florek ve ark (1951) tarafından önerilmiştir. Florek ve ark (1951)'nin çalışmasından bağımsız olarak McQuitty (1957) ve Sneath (1957) tarafından da birimlerin sınıflanması için bu teknik önerilmiştir. Sokal ve Michener (1958)'in çalışmalarında ortalama bağlantı ve merkez bağlantı kümeleme teknikleri önerilmiştir. Ward (1963) ile Ward ve Hook (1963)'un çalışmalarında kümeleme için kayıp fonksiyonunu minimum yapacak şekilde birimlerin aşamalı olarak birleştirildiği Ward kümeleme tekniği önerilmiştir. Yine aşamalı kümeleme analizi teknikleri içerisinde yer alan medyan bağlantı tekniği ilk kez Gower (1966) tarafından önerilmiştir. Lance ve Williams (1967), kümelemede birimlerin yeni oluşturulan kümeye olan uzaklıkların

hesaplanmasındaki farklılıklara göre oluşturulan teknikler arasındaki bağıntıyı α, β ve γ parametrelerini kullanarak açıklamışlardır.

Aşamalı olmayan kümeleme teknikleri veya başka bir deyişle parçalamaya dayalı kümeleme yöntemleri içerisinde en çok bilinen ve kullanılan k – ortalamalar tekniği farklı bilimsel alanlarda birbirlerinden bağımsız bir şekilde Steinhaus (1955), Lloyd (1957), Ball ve Hall (1965) ile MacQueen (1967)’in çalışmalarında ilk olarak önerilmiştir. MacQueen (1967)’in çalışmasında “ k – ortalamalar” kavramını kullandığı için çoğu kaynakta algoritmayı ilk öneren olarak onun ismi geçmektedir. Ayrıca Hartigan ve Wong (1979)’un k – ortalamalar tekniği ile ilgili çalışması yöntemin gelişimine önemli katkılar sağlamıştır. Günümüzden yaklaşık 50-55 yıl öncesi önerilen k – ortalamalar algoritması basitlik, uygulanabilirlik, deneysel çalışmalardaki başarısı ve etkinliği nedeniyle hala en çok tercih edilen kümeleme yöntemidir.

Kümeleme analizi içerisinde kullanımı yaygınlaşan modele dayalı kümeleme teknikleri de önemli bir kümeleme aracıdır. Karma dağılım modeliyle ilgili ilk çalışma olarak Pearson (1894) tarafından yapılan çalışma kabul edilmektedir. Pearson (1894) çalışmasında aynı coğrafik alanda yaşayan farklı yengeç türlerini modellemek için iki tek değişkenli normal dağılımın karmasını kullanmıştır. İki tek değişkenli normal dağılımın karmasındaki parametreleri, momentler yöntemini kullanarak tahmin etmiştir. Wolfe (1965, 1967), Day (1969) ve Binder (1978), çok değişkenli normal dağılımların karmasını, kümeleme analizinde istatistiksel model olarak kullanmışlardır. Celeux ve Govaert (1992), kümeleme için bir sınıflandırma EM algoritması ve iki stokastik versiyonu başlıklı çalışmalarında sınıflandırma maksimum olabilirlik yaklaşımı altında optimizasyona dayalı kümeleme yöntemleri için genel bir sınıflandırma EM algoritması tanımlamışlardır. Celeux ve Govaert (1993), karma ve sınıflandırma en çok olabilirliklerinin kümeleme analizlerinde karşılaştırılması başlıklı çalışmalarında en çok olabilirliğe dayalı karma ve sınıflandırma yaklaşımları ile yapılan kümeleme yöntemlerini farklı varsayımlar altında incelemişlerdir. Banfield ve Raftery (1993), normal dağılım modeline ve normal dağılım olmayan modellere dayalı kümeleme başlıklı çalışmalarında, çok değişkenli normal dağılımların karmasındaki bileşen kovaryans matrislerinin özdeğer

ayrışımı kullanılarak parametrik olarak ifade edilmesiyle, kümeler üzerinde bazı özelliklerin de dikkate alınmasını sağlamışlardır. McLachlan ve Peel (1996), eğitimsiz öğrenme için karma normal dağılım modellerine dayalı bir algoritma başlıklı çalışmalarında doğal küme yapısı bilinmeyen veriye bir normal karma modeli uyduran ve uydurulan bu modelin bileşenlerine gözlemleri sınıflandıran bir algoritma geliştirmişlerdir. Servi (2009), çok değişkenli karma dağılım modellerine dayalı kümeleme analizi başlıklı doktora tezinde, çok değişkenli normal dağılımların karmasındaki bileşen sayısı için bir yöntem önererek uzaktan algılanmış çok bantlı uydu görüntü verisini eğitilmiş olarak alan bazında sınıflandırmıştır.

Kümeleme analizi içerisinde çok sayıda farklı kümeleme tekniğinin olması verilerin kümelenmesinde hangi kümeleme tekniğinin daha başarılı olduğunu, elde edilen kümelemenin geçerliliğinin nasıl değerlendirileceği sorularını gündeme getirmiş ve bu konular üzerinde çok sayıda çalışma yapılmıştır. Bu çalışmalardan bazılarını aşağıda kısaca yer verilmiştir.

Sokal ve Rohlf (1962), aşamalı kümeleme tekniklerinin geçerliliği için kofenetik korelasyon katsayısını önermişlerdir. Kofenetik korelasyon katsayısı, başlangıçta hesaplanan benzerlik matrisi D ile kümeleme sonrası bağlantı uzaklıklarını içeren uzaklıklar matrisinin (R) karşılıklı elemanları arasındaki ilişkiyi ölçer. Kofenetik korelasyon katsayısı 1'e ne kadar yakınsa kümelemenin geçerliliği o kadar yüksektir.

Freidman ve Rubin (1967), küme içi değişkenliğin bir göstergesi olan küme içi kareler ve çarpımlarının toplamlarından oluşan W matrisinin determinantının, genel kareler ve çarpımlar toplamlarından oluşan T matrisinin determinantına bölünmesi ile elde edilen Wilks'in λ değerini kümelemenin geçerliliğini değerlendirmek için önermişlerdir. Önerilen bu kritere göre, λ değeri ne kadar küçük ise kümelemenin geçerliliği o kadar yüksektir.

Calinski ve Harabasz (1974), kümeler arası çarpımlar ve kareler toplamaları matrisinin izinin, kümeler içi çarpımlar ve kareler toplamaları matrisinin izine bölünmesi ile elde edilen değerin, veri setindeki birim sayısı ve küme sayısına göre hesaplanan çarpan ile çarpılmasıyla oluşan bir indeks önermişlerdir. Önerilen indeks değerini en büyük yapan kümeleme, en iyi kümeleme olarak seçilmektedir.

Marriott (1971), kümeler içi çarpımlar ve kareler toplamaları matrisinin determinantının küme sayısının karesi ile çarpılması ile elde edilen d indeksini en küçükleyen kümelemenin en başarılı kümeleme olarak kabul edilmesini önermiştir.

Rand (1971), kümeleme tekniklerinin değerlendirilmesinde objektif bir kriter isimli çalışmada ikili birim çiftlerinin iki farklı kümeleme sonucundaki durumlarına göre hesaplanan Rand indeksini önermiştir. İndeks değeri 1'e yaklaştıkça, iki farklı kümeleme sonucunun birbirine olan benzerliği artmaktadır.

Hubert ve Arabie (1985), küme sayısının artması durumunda Rand indeks değerinin de kümeleme performansına bağlı olmadan artması problemine karşın genelleştirilmiş hipergeometrik dağılım varsayımı altında düzeltilmiş Rand indeks değerini önermişlerdir.

Dunn (1974) iyi ayrılmış kümeler ve optimum bulanık parçalanma isimli çalışmada küme geçerliliği için yeni bir indeks önermiştir. Kümeleme sonuçları değerlendirilirken önerilen indeks değerini en büyükleyen kümeleme en başarılı kümeleme olarak seçilmektedir. Önerilen indeksin en önemli dezavantajı, küme sayısı artıkça indeks değerinin hesaplanmasının giderek zorlaşmasıdır.

Baker ve Hubert (1975), iki kümeleme sonucunda küme üyeliği tutarlı olan birimlerin sayısı ile tutarlı olmayan birimlerin sayısı arasındaki farkın, toplam birim sayısına oranlanması ile elde edilen Γ indeksini farklı kümeleme sonuçlarının değerlendirilmesi için önermişlerdir. Γ indeks değeri 1'e yaklaştıkça, iki farklı kümeleme sonucunun birbirine olan benzerliği artmaktadır.

Kuiper ve Fisher (1975), iki değişkenli ve çok değişkenli normal dağılımdan ürettikleri veri setlerini kullanarak tek bağlantı, tam bağlantı, ortalama, merkez, medyan ve Ward kümeleme tekniklerini karşılaştırmışlardır. Aşamalı kümeleme teknikleri içerisinde yer alan 6 farklı tekniği birimlerin gerçek küme üyeliklerinin bilindiği varsayımı altında Rand indeksi ile karşılaştırmışlardır. Çalışma sonunda küme sayısı azaldıkça kümeleme performansının yükseldiğini, tek bağlantı kümeleme tekniğinin diğer teknikler karşısında daha başarısız olduğuna, kümelerde yer alan birim sayıları yaklaşık eşit iken Ward tekniğinin daha başarılı olduğunu, kümelerde yer alan birim sayıları farklı olduğunda ise ortalama ve merkez bağlantı kümeleme tekniklerinin daha başarılı olduğunu vurgulamışlardır.

Blashfield (1976), tek bağlantı, tam bağlantı, ortalama bağlantı ve Ward kümeleme tekniklerini Kappa katsayısına göre karşılaştırmıştır. Kümeleme sonucu elde edilen küme üyelikleri ile gerçek küme üyelikleri arasındaki uyumun bir ölçüsü olan Kappa katsayısı ile karşılaştırılmasında rassal olarak üretilen 50 farklı veri seti kullanmışlardır. Karşılaştırma sonucunda Ward kümeleme tekniği en başarılı kümeleme tekniği olurken ortalama bağlantı kümeleme tekniği en başarısız kümeleme tekniği olmuştur.

Hubert ve Schultz (1976), kümelemenin geçerliliği için aynı küme içerisindeki birimler arasındaki uzaklıklar toplamına dayanan C indeksini önermişlerdir. C indeksi $[0,1]$ aralığında değer almakta ve minimum C değerine sahip kümelemenin daha başarılı olduğu kabul edilmektedir.

Davies ve Bouldin (1979), kümelemenin geçerliliğinin değerlendirilmesinde farklı iki kümedeki birimlerin kendi küme merkezleri arasındaki uzaklıklar toplamının, iki küme merkezi arasındaki uzaklığa bölünmesi ile elde edilen değerlere göre hesaplanan DB indeksini önermişlerdir. Çalışmalarında DB indeks değeri küçük olan kümelemenin, DB indeks değeri büyük olan kümelemeye göre daha geçerli olduğunu belirtmişlerdir.

Milligan (1981), kümeleme analizinde küme geçerliliği için önerilen 30 içsel (internal) kriteri Monte-Carlo simülasyonu ile karşılaştırmıştır. Çalışmasında aşamalı kümeleme tekniklerinden tek bağlantı, tam bağlantı, ortalama bağlantı ve Ward tekniklerini kullanmıştır. Çalışmasında ele aldığı 30 içsel kriteri, harici (external) kriterler olan Rand ve Jaccard indeksleri ile olan Pearson ve Spearman korelasyon değerlerine göre değerlendirmiştir. Değerlendirme sonucunda Baker ve Hubert (1975)'in önerdiği Γ indeksi en başarılı içsel kriter olarak belirlenmiştir.

Milligan ve Cooper (1985), aşamalı kümeleme tekniklerinin değerlendirilmesinde harici kriterleri ile ilgili çalışmalarında, Hubert ve Arabie (1985) tarafından önerilen düzeltilmiş Rand indeksini en başarılı kriter olarak belirlemişlerdir.

Rousseeuw (1987), kümelemenin geçerliliği için birimlerin kendi kümesi içerisindeki diğer birimlere olan uzaklığı ve diğer kümelerdeki birimlere olan uzaklığına dayanan Silhouette istatistiğini önermiştir.

Hands ve Everitt (1987), küme üyeliklerini bilerek ürettikleri rassal veri setlerini kullanarak beş farklı aşamalı kümeleme tekniğini karşılaştırmışlardır. Bu çalışmada tek bağlantı, tam bağlantı, ortalama, merkez ve Ward kümeleme teknikleri iki farklı kritere göre karşılaştırılmıştır. Çalışma sonucunda bu beş tekniğin çoğu kez benzer kümeleme performansı gösterdiğini, buna karşın bu teknikler içerisinde tek bağlantı kümeleme tekniğinin diğerlerine göre daha kötü performans sergilerken özellikle kümelerde yer alan birim sayılarının yaklaşık eşit olması durumunda Ward tekniği ile daha iyi sonuçlar elde edildiğini belirtmişlerdir.

Xie ve Beni (1991) küme geçerliliğinin değerlendirilmesinde, veri setinin geometrik yapısı ve birimlerin küme üyelik derecelerine dayanan XiE indeksini önermişlerdir. XiE indeksinin değerinin küçük olması elde edilen kümelemenin daha iyi olduğu anlamına gelmektedir.

Dinçer (1992) doktora çalışmasında, kümeleme analizinde uygun kümeleme ölçütlerini karşılaştırmıştır. Çalışmada, çok değişkenli normal dağılımdan üretilen 10, 20, 30, 40 ve 50 birimlik beş veri seti ve bu veri setlerinin birleştirilmesiyle elde edilen koşullu beş veri setini 2 ile 5 arasında küme sayısına göre k -ortalamlar tekniği ve tek bağlantı kümeleme tekniği ile kümelendirmiştir. Elde edilen kümeler, g_{1a} , g_{1b} , g_2 , g_{3a} , g_{3b} , Wilks'in Lambda, Hotelling Lawley iz ölçütü ve kofenetik katsayısı ile değerlendirilmiştir. 10 birimlik veri setinde, ölçütlerin farklı sonuçlar verdiği tespit edilmiş ve bu birim sayısının yetersizliğine bağlanmıştır.

Öztürk (1999) doktora çalışmasında, $n < p$ boyutlu biyolojik verilerde farklı kümeleme yöntemlerini karşılaştırmalı olarak incelemiştir. Erkenci fıstık tipi, orta geççi fıstık tipi ve geççi fıstık tipi şeklinde farklı 24 antep fıstığı 38 değişken bakımından incelenmiştir. Ele alınan veri setinde, $n < p$ durumu için öncelikle boyut indirgemeye gidilmiş ardından da farklı kümeleme teknikleri ile antep fıstığının türleri belirlenmeye çalışılmıştır. Çalışma sonucunda, temel bileşenler analizi ile birlikte Ward tekniğinin en iyi sonuç verdiği ifade edilmiştir. Küme sayısının belirlenmesinde ise C_{max} , Wilks'in Lambdası ve Hotelling Lawley iz ölçütünün en uygun ölçütler olduğu belirtilmiştir.

Maulik ve Bandyopadhyay (2002), bazı kümeleme algoritmalarının ve geçerlilik indekslerinin performanslarının değerlendirilmesi isimli çalışmalarında k -ortalamlar, tek bağlantı ve tavlama benzetimi yöntemine dayanan kümeleme algoritmalarının performanslarını karşılaştırmışlardır. Çalışmalarında geçerlilik indeksleri Dunn (1974), Davies ve Bouldin (1979), Calinski ve Harabasz (1974) ve son dönemlerde geliştirilen I indeksini küme sayısını belirleme amacı ile kullanmışlar ve çalışma sonucunda I indeksinin küme sayısı belirlemede daha başarılı olduğunu belirtmişlerdir.

Liao (2003), gen açıklamasında microarray verileri için kümeleme algoritmalarının karşılaştırılmasında, küme sayısının tahmini için Cureton ve D'Agostino sürecinde yeni bir düzenleme yapmışlar ve kümelemenin değerlendirilmesinde kullanılacak yeni bir yaklaşım önermişlerdir.

Erişoğlu ve ark (2006), aşamalı olmayan kümeleme analizi tekniklerinin karşılaştırılması isimli çalışmalarında k -ortalamlar tekniği ile bulanık k -ortalamlar tekniklerini karşılaştırmışlardır. Çalışmada 12 farklı durum için iki değişkenli normal dağılımın karmalarından ürettikleri 100'er birimlik veri setlerini kullanılmışlardır. Karşılaştırma kriteri olarak, her birimin atandığı küme merkezine olan Öklid uzaklıklarının toplamının örneklem hacmine bölünmesi ile elde edilen değeri göz önünde bulunduran araştırmacılar, bulanık k -ortalamlar tekniğinin k -ortalamlar tekniğine göre daha başarılı olduğunu belirtmişlerdir.

Bouguessa ve ark (2006), küme geçerliliği için objektif bir yaklaşım isimli çalışmalarında yoğunlaşma ve ayrıştırmaya dayalı fonksiyonların birleşimine dayanan yeni bir geçerlilik indeksi önermişlerdir. Önerdikleri yeni indeksin kümeleme değerlendirilmesindeki başarısını, gerçek veri setleri ve simülasyon verileri üzerinde göstermişlerdir.

Işık ve Çamurcu (2007) k -ortalamlar, k -medoids ve bulanık k -ortalamlar algoritmalarının performansın uygulamalı olarak karşılaştırmışlardır. Karşılaştırmada sentetik veri setleri kullanılmıştır. Çalışma sonunda küme sayısının uygun belirlenmediği durumlarda üç algoritmanın da kümeleme performansının kötü olduğu, zaman açısından k -ortalamlar algoritmasının, küme merkezlerini bulmada

k -medoids algoritmasının ve örtüşen kümeleri tespit etmede bulanık k -ortalamlar algoritmasının daha başarılı olduğu belirlenmiştir.

Ferreira (2007), fonksiyonel veri kümelemesindeki dört temel aşamalı kümeleme tekniğini Rand indeksini kullanarak karşılaştırmıştır. Periyodik, periyodik olmayan ve sinyal fonksiyonları ile ürettikleri veri setlerinin kümelenmesinde Ward, tek bağlantı, tam bağlantı ve ortalama bağlantı kümeleme tekniklerini kullanmışlardır. Karşılaştırma sonucunda Ward kümeleme tekniği en başarılı teknik olarak seçilmiştir. Bu çalışmada ayrıca, özellikle veri setinin periyodik bir fonksiyonla temsil edilebildiği durumlarda Ward kümeleme tekniğinin daha başarılı olduğu belirtilmiştir.

Qiu ve Tamhane (2007), veri setinin tek değişkenli olması durumunda normal dağılımların karmasıyla kümelemeyi, k -ortalamlar tekniğiyle kümelemeyle karşılaştırmıştır. Karşılaştırma sonucunda normal dağılımların karmalarının kullanımı ile gerçekleştirilen kümeleme sonucunda elde edilen hatalı sınıflandırılan birim sayısının, k -ortalamlar tekniği ile elde edilen kümelemedeki hatalı sınıflandırılan birim sayısından daha düşük olduğunu ortaya koymuşlardır.

Akın (2008), veri madenciliğinde kümeleme algoritmaları ve kümeleme analizi isimli doktora çalışmasında Türkiye İstatistik Kurumu (TÜİK) tarafından 2004 yılında yapılan “Hane halkı Bütçe Anketi” çalışmasından derlenen veriler kullanılarak hem tüketici davranış kalıplarını belirlerken uygulamada kullandığı merkeze dayalı bölümleyici kümeleme algoritması ile yoğunluk tabanlı kümeleme algoritması sonuçlarını karşılaştırmıştır.

Jain ve Koronios (2008), kümelemenin geçerliliğinin değerlendirilmesinde kullanılan Silhouette istatistiği, Calinski ve Harabaz (1974) indeksi ile Baker ve Hubert (1975) tarafından önerilen Γ indeksini karşılaştırmışlardır. Çalışmalarında cam (glass), şarap (wine), İris ve Ruspini verilerini kullanmışlardır. Veri seti içerisinde aynı gözlem vektörüne sahip birimlerin çok olması durumunda, kümelemenin geçerliliğini değerlendirmede Γ indeksinin uygun olmadığını belirtmişlerdir.

Zalik ve Zalik (2011), farklı örneklem büyüklükleri ve yoğunluktaki kümeler için geçerlilik indeksleri isimli çalışmalarında, Dunn (1974), Davies ve Bouldin

(1979), Hubert ve Schultz (1976), Baker ve Hubert (1975), Bezdek (1974a,1974b) ile Xie ve Beni (1991) tarafından önerilen geçerlilik indeksleri karşılaştırılmıştır. Çalışmalarında bu indekslerin farklı örneklem büyüklüğü ve küme yoğunluklarında tutarlı sonuçlar vermediklerini belirterek farklı örneklem büyüklüğü ve küme yoğunluklarına karşı daha tutarlı olduğunu ifade ettikleri *SV* indeksini önermişlerdir.

Doğal küme yapısı bulunmayan birimlerin kümelenmesindeki amaç her araştırmada her araştırmacı için farklı olabileceği gibi kümelenirilecek veri setinin özelliklerinin de çok farklı olacağı aşîkârdır. Bu ve bunun gibi nedenlerden dolayı ne yazık ki kümeleme analizinde genel kabul görmüş bir kümeleme yöntemi bulunmamaktadır. Farklı bilimsel alanlarda farklı amaç ve farklı veri setleri için kümeleme ile ilgili çok sayıda hem uygulama hem de teorik çalışma mevcuttur. Kümeleme analizi ile ilgili literatürde üzerinde en fazla çalışan konulardan biri de uygun küme sayısının belirlenmesi problemidir. Uygun küme sayısının belirlenmesinde küme geçerlilik kriterlerinin kullanıldığı çalışmalar kadar özellikle küme sayısının belirlenmesi için önerilen çok sayıda yaklaşım bulunmaktadır.

Mojena (1977), yaptığı çalışmada aşamalı kümeleme tekniklerinde birleşme uzaklıkları $\alpha_0, \alpha_1, \dots, \alpha_{n-1}$, birleşme uzaklıklarının ortalama ve standart sapması sırasıyla $\bar{\alpha}$ ve s_α olmak üzere $\alpha_{j+1} > \bar{\alpha} + cs_\alpha$ eşitsizliğini kümelerin belirlenmesinde kullanılacağını savunmuştur. Önerilen eşitsizlikte yer alan c sabiti için Mojena, 2.75 ve 3.50 aralığını önermiştir. Fakat Milligan ve Cooper (1985) gerçekleştirdikleri simülasyon çalışması sonucuna dayanarak bu değeri 1.25 olarak önermişlerdir.

Krzanowski ve Lai (1988), uygun küme sayısının tahmini için, k ve $(k+1)$ küme sayılarına göre elde edilen kümelemelerdeki değişken sayısı ve küme kareler toplamını göz önünde bulundurarak hesapladıkları *DIFF* değerlerinin birbirine oranı ile elde edilen *KL* indeksini önermişlerdir. Bu yaklaşıma göre önerilen indeks değerini en büyük yapan k uygun küme sayısı olarak seçilmektedir.

Krolak-Schwerdt ve Eckes (1992), en küçük ve en yüksek uzanımlı ağaç yapısı ve homomorfik fonksiyon temelli grafik model ismi verdikleri yaklaşımla küme sayını tahmin etmeye çalışmışlardır. Önerilen modelin veri seti içerisindeki

kümelerin birbirinden iyi ayrılmış olması durumunda küme sayısının tahmininde daha başarılı olduğunu belirtmişlerdir.

Tibshirani ve ark (2001), bir veri kümesinde küme sayısını tahmin etmek için Gap istatistik yöntemini önermişlerdir. Veriyi kümelemede kullanılan tüm yöntemlerde uygulanabilir olan Gap istatistik yönteminde verinin yapısına uygun bir referans sıfır dağılımı oluşturulur. Bu referans dağılımından gerçek veri ile aynı hacimli rasgele örneklem üretilir. Bu rasgele örnekleme orijinal veriye uygulanan kümeleme yöntemi uygulanarak orijinal verinin küme içi saçılımı ile rasgele örneklemden elde edilen kümeleme yapısındaki küme içi saçılımlar karşılaştırılır. Bu yaklaşımda, Gap istatistik değerini maksimum yapan k sayısı uygun küme sayısı olarak alınır.

Sun ve ark (2001) bulanık k -ortalama tekniği için küme sayısını belirlemede yeni bir doğrulama indeksi önermişlerdir. Önerdikleri indeksin özellikle örtüşen küme yapısı durumunda uygun küme sayısı tahmininde daha başarılı olduğunu belirtmişlerdir.

Vassiliou ve ark (2004) çok değişkenli veri setlerinde küme sayısının tahmini için yeni bir benzerlik ölçüsü isimli çalışmalarında, tanımladıkları benzerlik ölçüsü ile grafik model yaklaşımını kullanarak uygun küme sayısını tahmin etmeye çalışmışlardır.

Yan ve Ye (2007), küme içi homojenliğin bir göstergesi olan küme içi hata kareler toplamını GAP istatistiğinde ağırlık olarak kullanarak uygun küme sayısını tahmin etmeye çalışmışlardır. Önerilen yöntem, sürekli verilerden oluşan veri setlerinin kümelendirilmesinde kullanılacak olan her kümeleme tekniği ile kullanılabilir. Çalışmada ağırlıklandırılmış Gap istatistiği gerçek ve simülasyon verileri kullanılarak Gap istatistiği ile karşılaştırılmıştır. Karşılaştırma sonucunda uygun küme sayısının tahmininde ağırlıklandırılmış Gap istatistiğinin daha başarılı olduğunu göstermişlerdir.

Servi ve Erol (2007) modele dayalı kümeleme analizinde uygun küme sayısının belirlenmesi için, kümelemede kullanılacak her bir değişken için AIC ve BIC değerlerine göre belirledikleri küme sayılarına dayanan yeni bir aralık tahmini önermişlerdir.

Kümeleme analizi teknikleri içerisinde önemli bir yere sahip olan k -ortalamlar tekniği sonuçları başlangıç küme merkezlerinin seçiminden etkilenmektedir. Küme başlangıç merkezlerinin rassal seçilmesi durumunda her kümelemede farklı sonuçlarla karşılaşılmasına neden olabileceği gibi kümeleme sonucunda da optimum bir kümeleme elde edilemeyebilir. Bu sakıncaları ortadan kaldırmak için k -ortalamlar tekniğinde başlangıç küme merkezlerinin seçimi için birçok algoritma önerilmiştir. Önerilen algoritmalarından bazıları aşağıda belirtilmiştir.

Forgy (1965), veri setinden k tane rasgele merkez seçmek ve geriye kalan elemanları uzaklıklarına göre bu merkezlerin tanımladığı kümelere atarak oluşturulan küme merkezlerini başlangıç küme merkezleri olarak önermiştir.

MacQueen (1967), veri kümesinden k tane rasgele merkez seçmek ve geriye kalan elemanları veri sırasına ve uzaklıklarına göre bu merkezlerin tanımladığı kümelere atamayı, her atama sonrası küme merkezlerinin güncellenmesini önermiştir.

Kaufman ve Rousseeuw (1990), ilk önce tek bir merkezi bütün veri yığınının en merkezinde bulunan elemanı temsilci yapmak suretiyle seçmeyi, daha sonra geriye kalan elemanlardan her bir merkezi en çok eleman kapsayacak şekilde kümeler oluşturmayı önermiştir.

Babu ve Murty (1993), başlangıç kümelerinin seçimi problemine genetik algoritma sezgiselini kullandıkları bir çözüm önerisi getirmişlerdir.

Katsavounidis ve ark (1994), genelleştirilmiş Lloyd algoritmasında başlangıç koşullarına karşı duyarlılığı minimize edecek bir yöntem önermişlerdir.

Bradley ve Fayyad (1998) veri seti içerisinde rasgele oluşturulan küçük alt örneklemelerin k -ortalamlar tekniği ile kümelendirmesi sonucu oluşan merkezlerden başlangıç merkezlerinin oluşturulduğu arıtma algoritmasını önermişlerdir. Bu başlangıç algoritmasının kullanılması durumunda verinin k -ortalamlar algoritması ile kümeleneceği sonucu oluşan kümeleme hatasının minimum olacağını göstermişlerdir.

Pena ve ark (1999), k -ortalama tekniğinde başlangıç merkezlerinin seçilmesi ile ilgili karşılaştırma çalışması yapmışlardır. Başlangıç küme merkezlerinin Forgy

(1965), MacQueen (1967), Kaufman ve Rousseeuw (1990) tarafından önerilen yöntemlerle seçilmesi ile başlangıç küme merkezlerinin seçiminin rassal olarak seçilmesi durumunda k -ortalamlar tekniğinin kümeleme performansını incelemişlerdir. Çalışma sonunda Kaufman ve Rousseeuw (1990) tarafından önerilen yöntem yakınsama hızı bakımından diğer tekniklere göre daha başarılı bulunmuştur.

Likas ve arkadaşları (2003), $k - d$ ağaç yöntemiyle veri uzayını alt uzaylara parçalayarak başlangıç merkezlerini oluşturan global (küresel veya genel) k -ortalamlar algoritmasını önermişlerdir. Global k -ortalamlar algoritmasında, temel bileşenler analizinden elde edilen en büyük varyanslı veri eksenine dik düzlemde, hiper düzlem kesmesi kullanılır. Parçalanma birimlerin her birinin önceden belirlenen merkez sayısından az olmamak koşuluyla belirlenen noktalara bağlanmasına kadar gerçekleştirilir. Parçalanma sonucu oluşan düğümlerin merkezleri k -ortalamlar algoritmasında başlangıç küme merkezleri olarak alınır.

He ve ark (2004), k -ortalama tekniğinde başlangıç problemi ile ilgili olarak rasgele örnekleme, uzaklık eniyilemesi ve yoğunluk tahmini yaklaşımlarını ele alan bir karşılaştırma çalışması yapmışlardır.

Khan ve Ahmad (2004), veri setinde verilen her değişkeni göz önünde bulunduran küme merkezi başlangıç algoritmasını (Cluster Center Initialization Algorithm) önermişlerdir. Karmaşık bir yapıya sahip olan algoritma ile iyi küme başlangıç merkezleri elde etmek için bazı ön bilgilere ihtiyaç duyulmaktadır. Ayrıca küme sayısı arttıkça algoritmanın pratikte uygulanabilirliği azalmaktadır.

Ünler (2006) doktora çalışmasında, k -ortalamlar kümeleme algoritmasının ilk seçilen merkezlere duyarlılık ve yerel en iyi noktaya takılma gibi iki temel problemi incelemiştir. Bu çalışmada, başlangıç noktalarına olan duyarlılığa karşı önerilen k -harmonik ortalama algoritmasını tabu arama, tavlama benzetimi ve parçacık sürü eniyilemesi gibi üç temel sezgisel yaklaşımla birlikte kullanmıştır.

Deelers ve Auwatanamongkol (2007) çalışmalarında, k -ortalamlar algoritmasında küme başlangıç merkezlerinin seçimi için en büyük varyansa sahip değişkenin yayıldığı eksen boyunca veriyi parçalamaya dayalı bir yaklaşım önermişlerdir. En büyük varyansa sahip eksenin parçalanması için dikey

kesme algoritmasını kullanmışlardır. Çalışmalarında önerilen yöntemin özellikle renk ölçüm verilerinde daha başarılı olduğunu belirtmişlerdir.

Veri boyutundaki artış beraberinde yüksek boyutlu verilerin kümelenmesinde farklı yaklaşımların ortaya çıkmasına neden olmuştur. Birçok çalışmada, kümelendirilecek veri setinin çok fazla sayıda değişken içermesi durumunda kümeleme analizi öncesi değişken seçmek veya temel bileşenler analizi, faktör analizi gibi boyut indirgeme tekniklerini kullanarak veri setini daha basit bir yapıya dönüştürmenin kümeleme performansına olan etkileri incelenmiştir.

Jolliffe ve ark (1980), yaptıkları çalışmada yüksek boyutlu verilerde kümeleme analizi öncesi uygulanan temel bileşenler analizinden elde edilen ilk birkaç temel bileşen ile kümeleme analizi uygulamasıyla veri seti içerisindeki gizli kümelerin başarılı bir şekilde ortaya çıkarıldığını göstermişlerdir.

Chang (1983), küme sayısının 2 olduğu farklı ortalama ve kovaryans matrislerine sahip veri setlerinde ilk birkaç temel bileşenin küme yapısını ortaya çıkarmada her zaman etkili olamayacağını yaptığı simülasyon çalışması ile göstermiştir. Bu çalışmada sadece küme sayısının 2 olduğu durumlar için verinin iki boyutlu en iyi gösteriminin ilk ve son temel bileşen kullanılarak elde edileceğini belirtmiştir.

Steinley ve Brusco (2008), modele dayalı kümeleme ile modele dayalı olmayan kümeleme analizi öncesi boyut indirgeme amacı için önerilen sekiz farklı değişken seçim tekniğinin kümeleme performanslarını karşılaştırmışlardır. Çalışma sonucunda küme içi hata kareler toplamının düşürülmesi ile varyans-aralık ağırlıklandırılması prosedürlerinin birleştirilmesi ile elde edilen yöntem en başarılı yöntem olarak belirlenmiştir. Ayrıca sonlu karma dağılım modellerinde değişken seçimi tekniklerinin başarılı olmadığını belirtmişlerdir.

Valarmathie ve ark (2009), yüksek boyutlu veri setlerinde kümeleme analizi öncesi temel bileşenler analizinden elde edilecek en yüksek varyans açıklama oranına sahip birkaç bileşenin kümelemede kullanılmasının yakınsama hızı bakımından etkili bir yöntem olduğunu belirtmişlerdir. Bu çalışmada kümelemede kullanılacak bileşen sayısının seçiminde yamaç eğim grafiği kullanılmıştır.

Kümeleme en temel anlamda benzer birimlerin belirlenmesi problemidir. Dolayısıyla kümeleme tekniği ne olursa olsun benzerliğin belirlenmesi oldukça önemlidir. Benzerliğin belirlenmesinde kullanılan uzaklık ölçülerinin seçimi kümelemenin başarısında önemli bir etkindir. Uzaklık ölçüleri ve uzaklık ölçülerinin seçimi ile ilgili literatürde yer alan çalışmalardan bazıları aşağıda verilmiştir.

Cronbach ve Gleser (1953), karesel Öklit uzaklığı, değişkenler arasında ilişki olması durumunda analiz öncesi uygulanan temel bileşenler analizinden elde edilen birim varyanslı skorlar arasındaki karesel Öklit uzaklığı, Mahalonobis uzaklığı, Kendall'ın parametrik olmayan uzaklığı, Manhattan City-Block uzaklığının karesi ve Gower'ın logaritmik uzaklık ölçüsü arasındaki ilişkileri teorik olarak ortaya koymuşlardır.

Gower (1966), çok değişkenli analizler içerisinde yer alan diskriminant analizi, temel bileşenler analizi ve faktör analizinde bazı uzaklık ölçülerinin özelliklerini teorik olarak incelemiştir. Öklid uzayının kullanılması için gerekli ve yeter koşulların verildiği çalışmada, analizlerden optimum sonuç almada uygun uzaklık ölçüsü seçiminin önemli olduğu belirtilmiştir.

Green ve Rao (1969), kümeleme analizinde kullanılan on farklı uzaklık ölçüsünü teorik ve uygulamalı olarak karşılaştırmışlardır. Çalışmalarında 15 farklı marka bilgisayarın 6 özelliğine göre oluşturulan veri setinde farklı uzaklık ölçüleri ile hesaplanan uzaklıklar arasındaki korelasyonları incelemişlerdir. Korelasyon değerleri incelendiğinde diğer uzaklık ölçülerinden en farklı uzaklık ölçüsünün parametrik olmayan Kendall uzaklık ölçüsü olduğu görülmüştür.

Tombak (1996), çok değişkenli analizlerde kullanılan uzaklık ve benzerlik ölçütlerinin kullanım yerleri üzerine uygulamalı bir doktora çalışması yapmıştır. Çalışmada UNICEF'in 1995 yılı "Dünya Çocuklarının Durumu" yayınından Avrupa, Asya ve Afrika ülkeleri arasından rasgele seçilen 5'er ülkenin 1993 yılına ait 9 değişken bakımından incelenmesiyle oluşturulan veri seti kullanılmıştır. Ele alınan veri seti, farklı uzaklık ölçütleri ile tek bağlantı kümeleme tekniğine göre kümelendi. Kümeleme sonuçları değerlendirilirken uzaklık ölçütüne göre birbirine en yakın birimler belirlenmiştir.

Roussos ve ark (1998) çok boyutluluğun belirlenmesi için aşamalı kümeleme analizinde yeni bir benzerlik ölçüsünün kullanılması isimli çalışmalarında iki ve üç boyutlu olarak ürettikleri veri setlerinde koşullu kovaryans değerini temel alarak geliştirdikleri uzaklık ölçüsünün tek bağlantı, tam bağlantı, ortalama bağlantı ve ağırlıklı ortalama bağlantı aşamalı kümeleme tekniklerindeki performansı incelemişlerdir.

Finch ve Huynh (2000), iki değer alabilen değişkenlerle kümeleme analizi yapıldığında kullanılan Dice, Jaccard, eşleştirme ve Russel/Rao benzerlik ölçülerini karşılaştırmışlardır. Karşılaştırma sonucunda iki değer alan değişkenlerden oluşan veri setlerinde verinin iki kümeye ayrılması durumunda ele aldıkları benzerlik ölçülerinin benzer kümeleme performansı gösterdiğini ifade etmişlerdir.

Gunjaca ve ark (2000), bitki sınıflamasında kullanılan kümeleme analizinde farklı uzaklık ölçülerinin kullanılması durumunda farklı sonuçlar elde edileceğini göstererek analiz öncesi uygun uzaklık ölçüsünün seçiminin önemli olduğunu belirtmişlerdir. Çalışmalarında üç farklı türden 123 fasulyenin 17 nitel ve 9 nicel değişken bakımından gözlemlenmesi ile oluşan veri setini kullanmışlardır. Nitel ve nicel değişkenlerin bir arada bulunduğu bu veri seti için Gower (1971), Peeters-Martinelli (1989) ile Cole-Rodgers ve ark (1997) önerdikleri uzaklık ölçülerini kullanarak kümeleme analizi gerçekleştirmişlerdir. Uzaklık ölçülerine bağlı olarak kümeleme sonucunda farklı büyüklük ve yapıda kümelerin oluştuğunu göstererek, karma veri setleri için Cole-Rodgers ve ark (1997) tarafından önerilen uzaklık ölçüsünün en uygun uzaklık ölçüsü olduğunu vurgulamışlardır.

Mimmack ve ark (2001), iklim biliminde k -ortalamlar kümeleme algoritmasında yaygın bir kullanıma sahip olan Öklid uzaklığına alternatif olarak Mahalonobis uzaklığını önermişlerdir. Çalışmalarında 1961-1990 yılları arasında Güney Afrika ve Lesotho'da bulunan 517 istasyondan elde edilen metre kareye düşen aylık yağış miktarlarıyla ilgili veri setinin k -ortalamlar tekniği ile kümelenmesinde Mahalonobis uzaklığının Öklid uzaklığı kadar iyi sonuç verdiğini göstermişlerdir.

Jie ve ark (2006), k -ortalamlar tekniğinde birimlerin küme merkezlerine olan Öklid ve Manhattan uzaklıklarının hesaplanmasında aritmetik ortalama yerine

geometrik ortalama, medyan, harmonik ortalama gibi farklı ortalamaların kullanılması durumunda bu uzaklıkların kümeleme performansına olan etkilerini incelemişlerdir. Çalışma sonunda aritmetik ortalama yerine farklı ortalamalar kullanılarak elde edilen uzaklık ölçülerinin de klasik uzaklık ölçüleri kadar iyi performans gösterdiklerini vurgulamışlardır.

Pun ve Ali (2007), k -ortalamalar kümeleme algoritması için tek uzaklık ölçüsü yaklaşımı isimli çalışmalarında k -ortalamalar kümeleme algoritmasında “Hangi uzaklık ölçütü kullanılmalıdır ?” sorusuna cevap aramışlardır. Çalışmalarında farklı özelliklere sahip 112 veri seti bilinen gerçek küme sayılarına göre sırasıyla City-Block, Öklid ve karesel Öklid uzaklığı kullanılarak k -ortalamalar tekniğiyle kümelendi ve doğru sınıflandırma yüzdeleri hesaplanmıştır. Çalışma sonucunda entropiye dayanarak hesaplanan eşik değerinin 8.4595’den büyük olması durumunda k -ortalamalar algoritmasında uzaklık ölçüsü olarak City-Block uzaklığını, eşik değerinin 8.4595’e eşit veya daha küçük olması durumunda ise karesel Öklid uzaklığını önermişlerdir.

Vimal ve ark (2008), kümeleme analizinde en yaygın kullanılan Öklid uzaklık ölçüsünü, bilgisayar programlama algoritmalarına kullanılan bit-vektör uzaklığı, karşılaştırmalı kümeleme uzaklığı ve Huffman kod uzaklığı ile karşılaştırmışlardır. İlgili uzaklık ölçülerinin k -ortalamalar, yoğunluk tabanlı ve üstünlük-tabanlı kümeleme algoritmalarındaki etkinliklerini, sentetik veri setleri ve gerçek kriket verisini kullanarak incelemişlerdir. Çalışma sonucunda, tüm algoritmalarda Öklid uzaklığını en iyi kümeleme performansı gösteren uzaklık ölçüsü olarak belirlenmiştir.

3. KÜMELEME ANALİZİ

Benzer birimlerin bir araya getirilmesini amaçlayan kümeleme analizi gelişimini, birbirlerinden bağımsız olarak farklı disiplinlerde farklı amaçlar doğrultusunda sürdürmüştür. Kullanım alanları ve amaçlardaki farklılıklar nedeniyle kümeleme analizi için tek bir tanımın verilmesi oldukça zordur. Bunun yanında günümüzde kümeyi oluşturan birimleri bir araya getiren ortak özelliklerin farklı olmasından dolayı küme kavramı ile ilgilide çok değişik şekillerde tanımlamalar bulunmaktadır. Everitt ve ark (2001), küme (cluster) terimi için genel olarak ifade edilebilecek tek bir tanımın olmadığını, böyle bir tanımın yapılmasının zor olmasının yanında yanlış anlamlar çıkarılabileceğini belirtmiştir. Küme terimi için yapılan tanımlardan birkaçı

- Küme (cluster) birbirine benzer bireylerin bir topluluğu olup, bu topluluktaki bireylerin diğer topluluktaki bireylerden farklı olması durumudur.
- Bir küme (test uzayında noktalar toplamı olup) içerisindeki herhangi iki nokta çifti arasındaki uzaklığın, içerisindeki bir nokta ile dışındaki bir nokta arasındaki uzaklıktan küçük olması durumudur.
- Kümeler p boyutlu özellikler uzayında oldukça yüksek yoğunluklu noktaları içeren ve nispeten daha düşük yoğunlukta noktaları içeren diğer bölgelerden ayrılan sürekli bölgelerdir.

şeklinde. Aşağıda kümeleme analizine ilişkin bazı tanımlara yer verilmiştir.

- Mac Quenn (1967), kümeleme analizini bireylere ait çok boyutlu gözlem değerlerinin analizi ve sınıflandırılması olarak tanımlamaktadır.
- Romesburg (1984) birimler üzerinden alınan gözlem değerlerinden oluşan veri matrisi $\mathbf{X}$ olmak üzere kümeleme analizini, $\mathbf{X}$ matrisini oluşturan ve doğal gruplamaları kesin olarak bilinmeyen birimlerinden, birbirleri ile benzer olanları alt kümelere ayırmaya yardımcı olan yöntemler topluluğu olarak tanımlamaktadır.
- Kaufman ve Rousseeuw (1990) kümeleme analizini bir araştırmada incelenen birimleri aralarındaki benzerliklerine göre belirli gruplar

içinde toplayarak sınıflandırma yapmayı, birimlerin ortak özelliklerini ortaya koymayı ve bu sınıflar ile ilgili genel tanımlar yapmayı sağlayan yöntemler olarak tanımlamaktadır.

- Hair ve Black (2000) kümeleme analizini, benzer birimleri belirleyerek kümeleyen ve gözlem kümelerinde yer alan birimlerin yapısal karakteristiklerini inceleyen objektif yöntemler topluluğu olarak tanımlamaktadırlar.
- Sharma (1996), doğal sınıflamaları hakkında açık bilgi bulunmayan durumlarda üzerinde çalışılan gözlemler topluluğunu, kendi içinde türdeş fakat diğerlerinden farklı olacak biçimde kümelere ayırma olanağı tanıyarak ana kitleye ilişkin tahminlerin yapılmasında yararlanılan çok değişkenli istatistiksel bir teknik olarak tanımlamaktadır
- Webb (2002), kümeleme analizini veri seti içerisindeki yapıları keşfetmek için benzer birimlerin bir araya getirildiği yöntemler topluluğu olarak tanımlamaktadır.
- Ye (2003), kümeleme analizini X veri matrisinde yer alan ve doğal gruplamaları kesin olarak bilinmeyen birimleri, değişkenleri ya da birim ve değişkenleri birbirleri ile benzer olan alt kümeler ayırmaya yardımcı olan yöntemler topluluğu olarak tanımlamaktadır.
- Kantardzic (2003), kümeleme analizini elde bulunan verinin belirlenen yöntemlerle analiz edilerek daha önceden etiketleri belli olmayan gruplara ayrılması işlemi olarak tanımlamaktadır.
- Mirkin (2005) kümeleme analizini veri içerisindeki kümeleri bulma ve özelliklerini belirleme arayışı olarak tanımlamaktadır.
- Gan ve ark (2007) kümeleme analizini, benzer birimlerin aynı kümede, farklı birimlerin farklı kümelerde olacağı şekilde birimlerin gruplanmasındaki yöntemler olarak tanımlamaktadırlar.

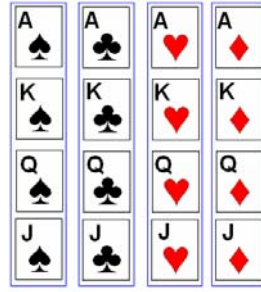
Bu tanımlamalar altında istatistik alanında kümeleme analizinin en genel anlamda tanımı, “Küme yapısı hakkında bir ön bilgiye sahip olunmayan birimlerin

p değişken bakımından ölçülmesi ile oluşan $n \times p$ boyutlu X veri matrisindeki farklı birimleri farklı kümelerde, benzer birimleri aynı kümede bir araya getirerek veri seti içerisindeki gizli kümeleri ortaya çıkaran ve bu kümelerin özelliklerini belirleyerek kitle hakkında tahminlerin yapılmasına olanak sağlayan yöntemlerden oluşan çok değişkenli istatistiksel bir analizdir” şeklinde yapılabilir.

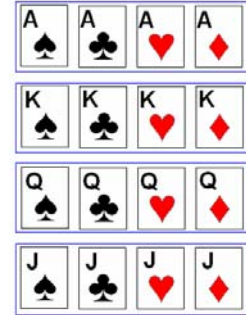
Birimlerin kümelenmesinin zorluğunu anlamak için 16 oyun kartını aynı nesneler kümesi haline getirmeye çalışalım. Bu kartların kümelenmesinde elde edilen bazı kümeler Şekil 3.1’de verilmiştir.



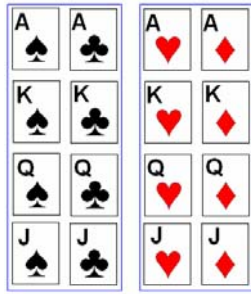
(a) Bireysel kartlar



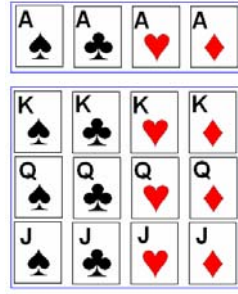
(b) Aynı şekle sahip kartlar



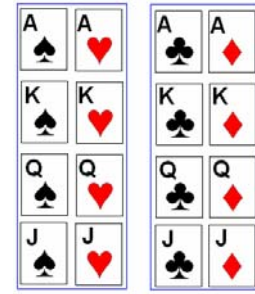
(c) Aynı harfe sahip kartlar



(d) Aynı renkteki kartlar



(e) Aslar ve diğer kartlar



(f) Major ve minor kartlar

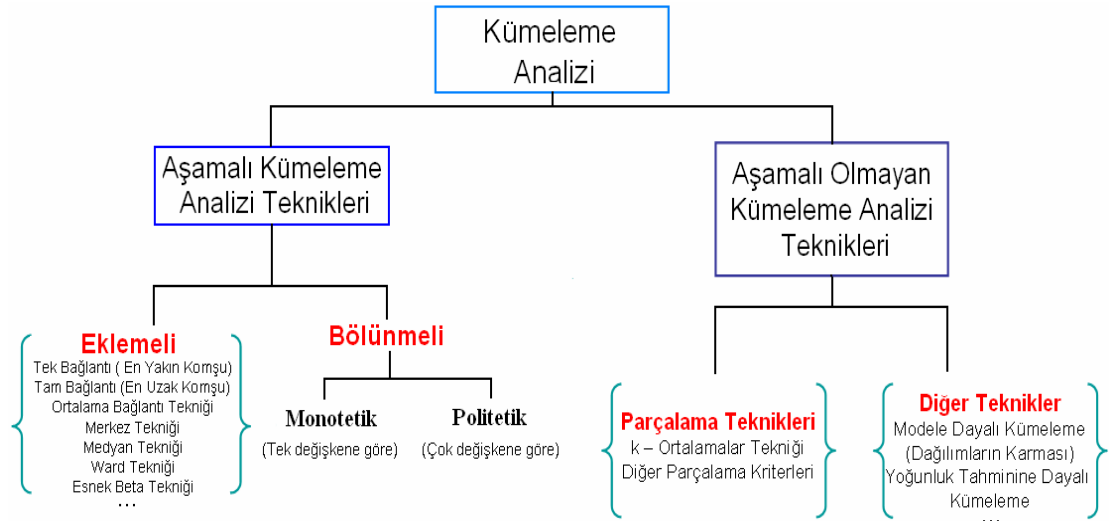
Şekil 3.1. 16 kartın kümelenmesinden elde edilebilecek farklı bazı kümeler

Veri yığını oluşturulan n birimin g kümeye ayrılmasında

$$N(n, g) = \frac{1}{g!} \sum_{k=1}^g \binom{g}{k} (-1)^{g-k} k^n \quad (3.1)$$

eşitliği ile verilen sayıda farklı durum mevcuttur (Duran ve Odell 1974, Jensen 1969, Seber 1984). Örneğimizdeki 16 kartı bir kümede toplamının 1 yolu varken iki

kümele ayırmak için 32767 farklı yolu vardır. 16 kartı 3 kümele ayırmak için ise 7141686 farklı yol mevcuttur. Sadece 16 birimi 3 kümele ayırmada bile mümkün tüm kümelemeleri gerçekleştirip içerisinde en iyi olan kümelemeyi belirlemek neredeyse imkânsız gibi görülmektedir. Bunun için birimlerin en geçerli şekilde kümelemesi için algoritmalara ihtiyaç vardır. Kümeleme analizi içerisinde benzer birimleri bir araya getirmede kullanılan çok sayıda algoritma vardır. Bu algoritmaların farklı sınıflandırmaları olmasına karşın genel olarak benzer birimleri bir araya getirmede izledikleri yöntemle göre aşamalı (hierarchical) ve aşamalı olmayan (nonhierarchical) kümeleme yöntemleri olarak ikiye ayrılır (Aldenferder ve Blashfield 1984, Jain ve Dubes 1988, Öztürk 1999, Everitt ve ark 2001, Ünler 2006, Servi 2009). Kümeleme analizi içerisindeki tekniklerin sınıflandırılmasına ilişkin diyagram Şekil 3.2’de verilmiştir.



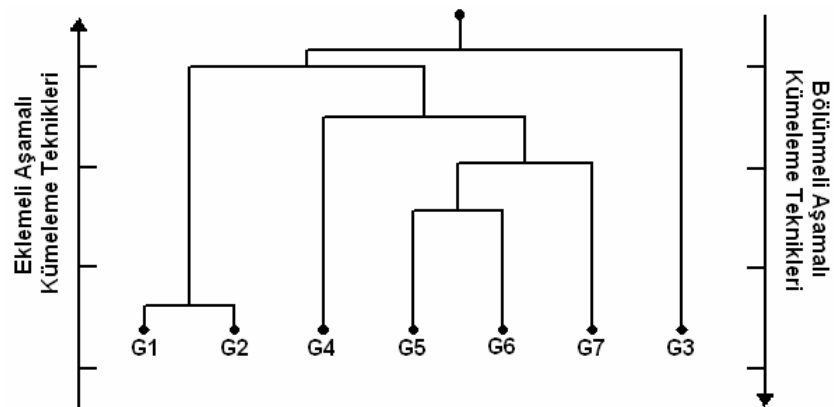
Şekil 3.2. Kümeleme analizi tekniklerinin sınıflandırılması

Bu bölümde, kümeleme analizi içerisinde yer alan bazı önemli teknikler aşamalı ve aşamalı olmayan kümeleme analizi teknikleri başlıkları altında anlatılacaktır. Ayrıca kümeleme ile ilgili önemli kavramlardan küme geçerliliği için önerilen kriterler, içsel (internal), dışsal (external) ve oransal (relative) geçerlilik kriterleri başlıkları altında verilecektir.

3.1. Aşamalı Kümeleme Analizi Teknikleri

Aşamalı kümeleme algoritmaları, ardışık bir işlem süreci içerir. Elde edilen kümeleme sonuçları, dendrogram olarak bilinen bir ağaç diyagramı ile grafiksel olarak gösterilebilir. Bu grafikte birleştirilen küme çiftleri ve bağlanma uzaklıkları görülebilir. Anderberg (1973), aşamalı kümeleme yöntemlerinin, araştırmacının incelediği veri setinde kaç küme bulunduğunu başlangıçta bilmediği durumlarda çok uygun bir yöntem olduğunu ve yöntemin araştırmacıya incelediği veri setinde daha önce gözlemlenmemiş ilişkileri ve prensipleri keşfetme olanağı vermesi nedeniyle faydalı olduğunu belirtmektedir.

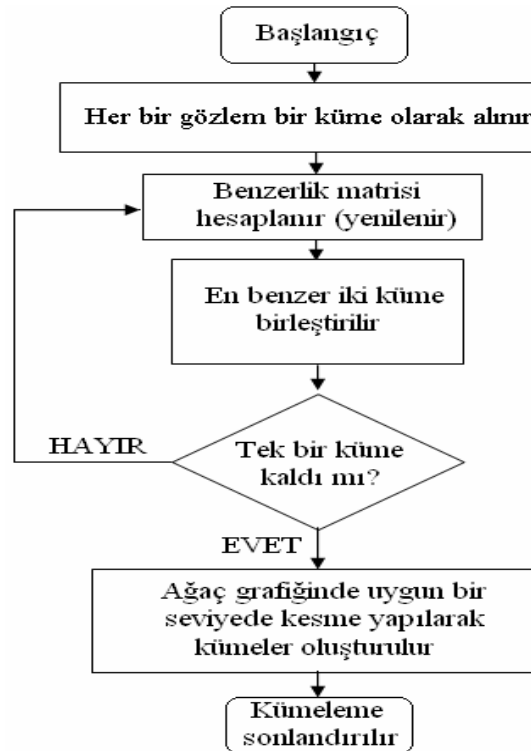
Aşamalı kümeleme yöntemleri izledikleri yola göre eklenmeli ve bölünmeli aşamalı kümeleme teknikleri olarak iki alt sınıfta incelenmektedir. Eğer başlangıçta n sayıda gözlemin her biri bir küme olarak düşünülüp, bu kümeler her aşamada en benzer iki kümenin bağlanması ile tek küme kalıncaya kadar birleştiriliyorsa teknik eklemeli aşamalı kümeleme tekniği olarak isimlendirilir (Hansen ve ark. 1997). Başlangıçta n sayıda gözlemi tek bir küme içerisinde kabul edip, n birimlik bu kümeden, küme bölme işlemlerinin uygulanmasıyla her bir gözlem farklı kümede olacak şekilde n sayıda küme oluşturuluyorsa teknik bölünmeli aşamalı kümeleme tekniği olarak isimlendirilir (Gordon 1999).



Şekil 3.3. Aşamalı kümelenin ağaç diyagramı(dendrogram) ile grafiksel gösterimi

Bölünmeli bir aşamalı kümeleme algoritması n sayıda gözlemden oluşan bir veriye uygulandığında daha ilk adımda, $2^{n-1} - 1$ sayıda mümkün yolla boş olmayan iki alt kümeye bölünebilir. Bölünmeli aşamalı kümeleme algoritmaları hesaplama maliyetleri oldukça yüksek olduğunu için pek tercih edilmezler. Bölünmeli hiyerarşik kümeleme genel olarak iki kategoriye ayrılır. Bunlar monothetic ve polythetic bölünmeli hiyerarşik kümeleme yöntemleridir (Everitt 1993, Willett 1988). Monotetik aşamalı kümelemede bir küme, iki alt kümeye yalnız tek bir değişken dikkate alınarak bölünürken, Politetik hiyerarşik kümelemede ise bölünme, tüm p değişkeninde dikkate alınmasıyla yapılır (Rencher 2002).

Aşamalı kümeleme teknikleri içerisinde kullanımı daha yaygın olan eklemeli algoritmalarda, her aşamada benzer kümeler birbirine bağlanır ve sonrasında benzerlik matrisi yenilenir. Benzerlik matrisinin yenilenmesinde başka bir ifadeyle mevcut kümelerin oluşan yeni kümeye olan benzerliklerinin hesaplanmasındaki farklılıklara göre farklı algoritmalar oluşturulmuştur. Genel olarak eklemeli aşamalı kümeleme algoritmalarındaki akış şeması Şekil 3.4’de verilmiştir.



Şekil 3.4. Eklemeli aşamalı kümeleme algoritmalarının akış şeması

Aşamalı kümeleme teknikleri başlığında daha yaygın bir kullanıma sahip olan eklemeli aşamalı kümeleme algoritmalarından tek bağlantı, tam bağlantı, ortalama bağlantı, ağırlıklı ortalama bağlantı, merkez bağlantı, medyan bağlantı ve Ward kümeleme teknikleri ile bu teknikler arasındaki bağıntıyı ortaya koyan esnek Beta tekniği anlatılacaktır. Çalışmamızda uzaklık ölçülerinin kümelemeye olan etkisi inceleneceğinden benzerliğin ölçümünde uzaklık türü ölçüler kullanılacaktır. Bu nedenle eklemeli aşamalı kümeleme teknikleri benzerlik matrisinin uzaklık türü ölçülerden elde edilmesi durumuna göre açıklanacaktır. Tekniklerin gösteriminde kullanılacak 6 birimin iki değişkene göre aldığı değerlerden oluşan örnek veri seti ve birimler arasındaki uzaklık matrisi aşağıda verilmiştir.

Tablo 3.1. Eklemeli aşamalı kümeleme tekniklerinin gösteriminde kullanılacak örnek veri seti

Birim No	x	y
1	0.40	0.53
2	0.18	0.38
3	0.35	0.32
4	0.26	0.20
5	0.08	0.41
6	0.45	0.30

Tablo 3.2. Birimler arasındaki Öklid uzaklığına göre elde edilen uzaklık matrisi

	1	2	3	4	5	6
1	0	0.266	0.216	0.358	0.342	0.235
2	0.266	0	0.180	0.197	0.104	0.282
3	0.216	0.180	0	0.150	0.285	0.102
4	0.358	0.197	0.150	0	0.277	0.215
5	0.342	0.104	0.285	0.277	0	0.386
6	0.235	0.282	0.102	0.215	0.386	0

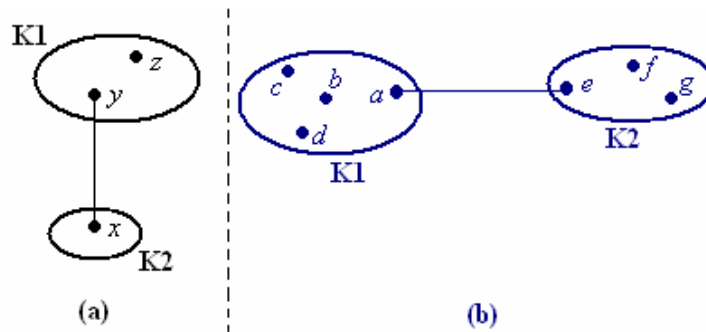
3.1.1. Tek Bağlantı Kümeleme Tekniği (Single Linkage Method)

Florek ve ark (1951) tarafından önerilen tek bağlantı kümeleme tekniği en fazla bilinen aşamalı kümeleme tekniğidir. Bu teknik Florek ve ark (1951)'nin çalışmasından bağımsız olarak McQuitty (1957) ve Sneath (1957) tarafından da önerildiği için farklı kaynaklarda bu tekniği ilk olarak önerenler olarak bu araştırmacıların ismi yer alabilmektedir (Gan ve ark 2007). “En yakın komşuluk”

olarak da bilinen bu teknikte en benzer birimlerin bir araya getirilmesinden sonra uzaklık matrisi hesaplanırken(yenilenirken) iki farklı kümedeki en yakın gözlemlere göre belirlenen uzaklık, iki küme arasındaki yenilenmiş uzaklık değeri olarak alınır. Tek bağlantı kümeleme tekniği zincirleme etkisine sahip olduğu için uzamış kümeleri ortaya çıkarma eğilimindedir (Everitt ve ark 2001). Kümeler birbirinden uzak ayrılmış ise tek bağlantı kümeleme tekniği iyi sonuç vermektedir (Xu ve Wunsch 2009). Tek bağlantı kümeleme tekniğinde x biriminin y ve z birimlerinden oluşan yeni kümeye olan uzaklığı,

$$d(x, \{y, z\}) = \text{enküçük}(d(x, y), d(x, z)) \quad (3.2)$$

eşitliği ile hesaplanır. Şekil 3.5’de kümeler arasındaki uzaklıkların tek bağlantı kümeleme tekniğine göre belirlenmesi grafiksel olarak gösterilmiştir.



Şekil 3.5. Tek bağlantı kümeleme tekniğine göre iki küme arasındaki uzaklığın belirlenmesi

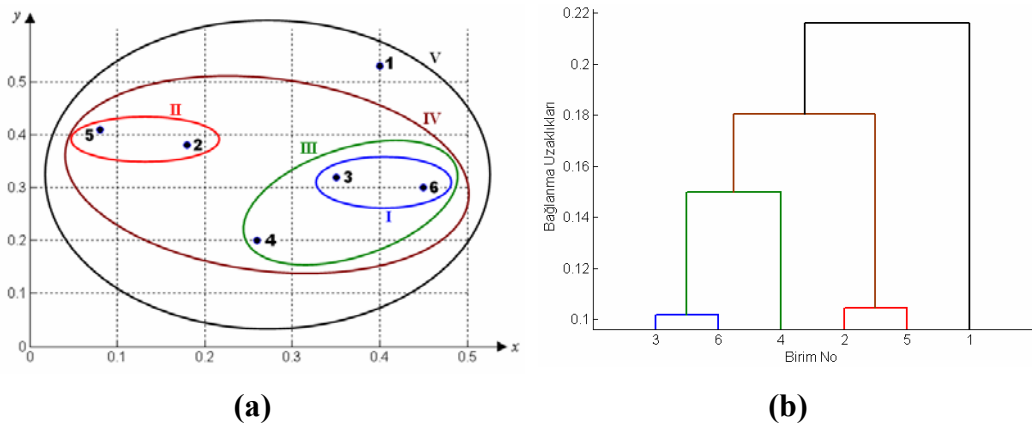
Şekil 3.5 (a)’da x birimden oluşan K2 kümesinin y ve z birimlerinden oluşan K1 kümesine olan uzaklığı tek bağlantı kümeleme tekniğine göre x ve y birimleri arasındaki uzaklığa eşittir. Şekil 3.5 (b)’de a, b, c ve d birimlerinden oluşan K1 kümesinin e, f ve g birimlerinden oluşan K2 kümesine olan uzaklığı tek bağlantı kümeleme tekniğine göre a ve e birimleri arasındaki uzaklığa eşittir. Tablo 3.1’de verilen örnek veri setindeki birimlerin tek bağlantı kümeleme tekniğine göre bağlanma(birleştirilme) aşamalarında elde edilen uzaklık matrisleri Tablo 3.3’de verilmiştir.

Tablo 3.3. Tek bağlantı kümeleme tekniğine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri

I. AŞAMA							II. AŞAMA					
	1	2	3	4	5	6		1	2	{3,6}	4	5
1	0	0.266	0.216	0.358	0.342	0.235	1	0	0.266	0.216	0.358	0.342
2		0	0.180	0.197	0.104	0.282	2		0	0.180	0.197	0.104
3			0	0.150	0.285	0.102	{3,6}			0	0.150	0.285
4				0	0.277	0.215	4				0	0.277
5					0	0.386	5					0
6						0						

III. AŞAMA					IV. AŞAMA				V. AŞAMA		
	1	{2,5}	{3,6}	4		1	{2,5}	{3,6,4}		1	{2,5,3,6,4}
1	0	0.266	0.216	0.358	1	0	0.266	0.216		1	0
{2,5}		0	0.180	0.197	{2,5}		0	0.180	{2,5,3,6,4}		0
{3,6}			0	0.150	{3,6,4}			0			
4				0							

Tablo 3.3 incelendiğinde I. aşamada en küçük uzaklık değerine sahip 3. ve 6 birimlerin bir küme oluşturduğu görülmektedir. Yeni oluşturulan kümeye olan uzaklıklar tek bağlantı kümeleme tekniğine göre hesaplanarak II. aşamada elde edilen uzaklık matrisindeki en küçük uzaklığa sahip 2. ve 5. birimler birleştirilmiştir. Örnek veri setindeki birimlerin tek bağlantı kümeleme tekniğine göre bağlanma(birleştirilme) aşamaları Şekil 3.6 (a)'da iki boyutlu düzlemde gösterilirken, Şekil 3.6 (b)'de Öklid uzaklığı kullanılarak gerçekleştirilen tek bağlantı kümeleme tekniğinden elde edilen ağaç grafiğine yer verilmiştir.



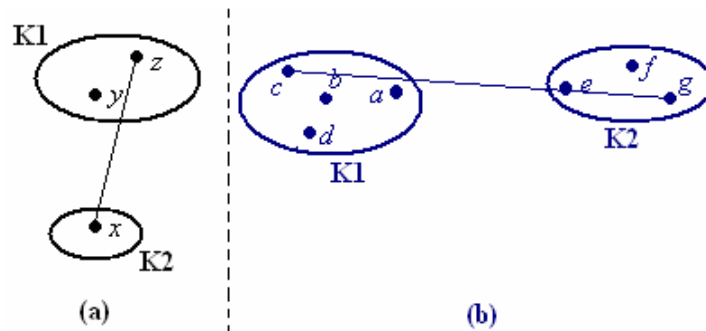
Şekil 3.6. Tek bağlantı kümeleme tekniği ile birimlerin bağlanmasının (a) iki boyutlu düzlemde ve (b) ağaç grafiği ile gösterimi

3.1.2. Tam Bağlantı Kümeleme Tekniği (Complete Linkage Method)

İlk olarak Sorensen (1948) tarafından önerilen tam bağlantı kümeleme tekniği aynı zamanda “en uzak komşuluk” olarak bilinir. Tam bağlantı kümeleme tekniğinde, iki küme arasındaki uzaklık ölçüsü olarak küme elemanları arasındaki en büyük uzaklık alınmaktadır. Tam bağlantı kümeleme tekniğinde x biriminin y ve z birimlerinden oluşan yeni kümeye olan uzaklığı,

$$d(x, \{y, z\}) = \text{enbüyük}(d(x, y), d(x, z)) \quad (3.3)$$

eşitliği ile hesaplanır. Şekil 3.7’de kümeler arasındaki uzaklıkların tam bağlantı kümeleme tekniğine göre belirlenmesi grafiksel olarak gösterilmiştir.



Şekil 3.7. Tam bağlantı kümeleme tekniğine göre iki küme arasındaki uzaklığın belirlenmesi

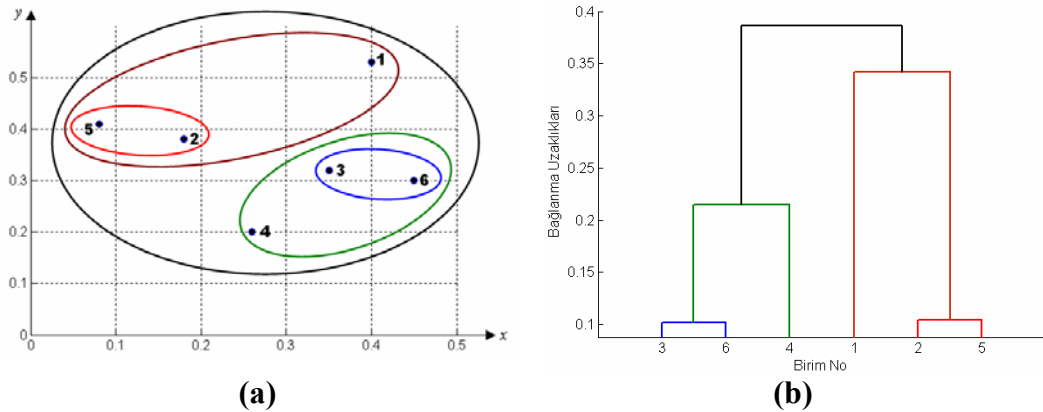
Şekil 3.7 (a)’da x birimden oluşan K2 kümesinin y ve z birimlerinden oluşan K1 kümesine olan uzaklığı tam bağlantı kümeleme tekniğine göre x ve z birimleri arasındaki uzaklığa eşittir. Şekil 3.7 (b)’de a, b, c ve d birimlerinden oluşan K1 kümesinin e, f ve g birimlerinden oluşan K2 kümesine olan uzaklığı tam bağlantı kümeleme tekniğine göre c ve g birimleri arasındaki uzaklığa eşittir. Görüldüğü gibi tam bağlantı kümeleme tekniğinde yeni uzaklıklar belirlenirken tek bağlantı tekniğinin tam tersi bir düşünce ile küme elemanları arasındaki en büyük uzaklık dikkate alınmaktadır. Bu nedenle teknik özellikle belirgin olmayan, küçük ve yoğun küme yapısına sahip verilerde etkin bir kümeleme tekniğidir (Martinez ve

Martinez 2005). Tablo 3.1’de verilen örnek veri setindeki birimlerin tam bağlantı kümeleme tekniğine göre bağlanma aşamalarında elde edilen uzaklık matrisleri Tablo 3.4’de verilmiştir.

Tablo 3.4. Tam bağlantı kümeleme tekniğine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri

I. AŞAMA							II. AŞAMA						
	1	2	3	4	5	6		1	2	{3,6}	4	5	
1	0	0.266	0.216	0.358	0.342	0.235	1	0	0.266	0.235	0.358	0.342	
2		0	0.180	0.197	0.104	0.282	2		0	0.282	0.197	0.104	
3			0	0.150	0.285	0.102	{3,6}			0	0.215	0.386	
4				0	0.277	0.215	4				0	0.277	
5					0	0.386	5					0	
6						0							
III. AŞAMA							IV. AŞAMA				V. AŞAMA		
	1	{2,5}	{3,6}	4			1	{2,5}	{3,6,4}		{2,5,1}	{3,6,4}	
1	0	0.342	0.235	0.358	1	0	0.342	0.358	{2,5,1}	0	0.386		
{2,5}		0	0.386	0.277	{2,5}		0	0.386	{3,6,4}		0		
{3,6}			0	0.215	{3,6,4}			0					
4				0									

Tablo 3.4 incelendiğinde ilk aşamada 3 ile 6. birimler, ikinci aşamada 2 ile 5. birimler, üçüncü aşamada {3,6} kümesi ile 4. birim, dördüncü aşamada {2,5} kümesi ile 1. birim ve son aşamada {2,5,1} ile {3,6,4} kümelerin bağlandığı görülmektedir. Örnek veri setindeki birimlerin tam bağlantı kümeleme tekniğine göre bağlanma aşamaları Şekil 3.8 (a)’da iki boyutlu düzlemde gösterilirken, Şekil 3.8 (b)’de Öklid uzaklığı kullanılarak gerçekleştirilen tam bağlantı kümeleme tekniğinden elde edilen ağaç grafiğine yer verilmiştir.



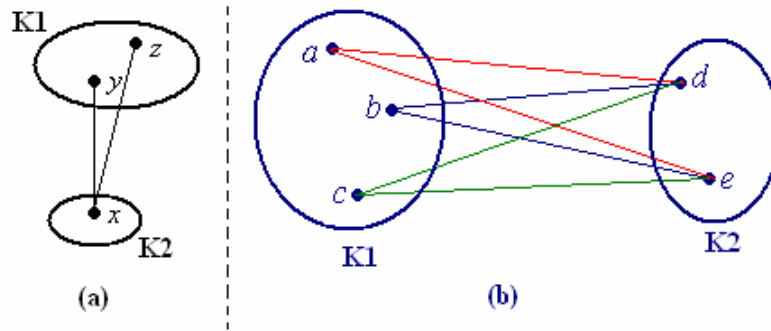
Şekil 3.8. Tam bağlantı kümeleme tekniği ile birimlerin bağlanmasının (a) iki boyutlu düzlemde ve (b) ağaç grafiği ile gösterimi

3.1.3. Ortalama Bağlantı Kümeleme Tekniği (Average Linkage Method)

Ortalama bağlantı kümeleme tekniğinde, iki küme arasındaki uzaklık bir kümedeki birimlerin diğer kümedeki birimlere olan uzaklıklarının ortalaması ile elde edilir. Bu yöntem aynı zamanda “ağırlıksız grup çiftleri yöntemi” olarak da adlandırılır (Everitt ve ark 2001, Jain ve Dubes 1988). Her biri sırasıyla n ve m adet birimden oluşan K_i ve K_j kümeleri arasındaki uzaklık

$$d(K_i, K_j) = \frac{1}{n \times m} \sum_{x_i \in K_i} \sum_{x_j \in K_j} d(x_i, x_j) \quad (3.4)$$

eşitliği ile tanımlanır. Bu teknik küçük ve yaklaşık olarak eşit varyansa sahip kümeler oluşturma eğilimindedir (Martinez ve Martinez 2005). Şekil 3.9’da kümeler arasındaki uzaklıkların ortalama bağlantı kümeleme tekniğine göre belirlenmesi grafiksel olarak gösterilmiştir.



Şekil 3.9. Ortalama bağlantı kümeleme tekniğine göre iki küme arasındaki uzaklığın belirlenmesi

Şekil 3.9 (a)’da x birimden oluşan K2 kümesinin y ve z birimlerinden oluşan K1 kümesine olan uzaklığı ortalama bağlantı kümeleme tekniğine göre x ve y birimleri arasındaki uzaklık ile x ve z birimleri arasındaki uzaklığın ortalamasına eşittir. Şekil 3.9 (b)’deki K1 kümesinin K2 kümesine olan uzaklığı a , b ve c birimlerinin sırasıyla d ve e birimlerine olan uzaklıklarının ortalaması ile elde edilir. Tablo 3.1’de verilen örnek veri setindeki birimlerin ortalama bağlantı

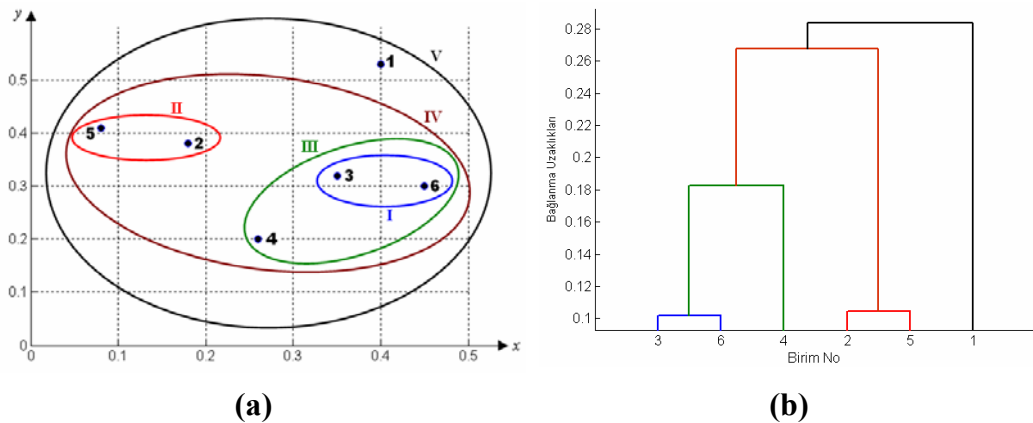
kümeleme tekniğine göre bağlanma aşamalarında elde edilen uzaklık matrisleri Tablo 3.5’de verilmiştir.

Tablo 3.5. Ortalama bağlantı kümeleme tekniğine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri

I. AŞAMA							II. AŞAMA					
	1	2	3	4	5	6		1	2	{3,6}	4	5
1	0	0.266	0.216	0.358	0.342	0.235	1	0	0.266	0.226	0.358	0.342
2		0	0.180	0.197	0.104	0.282	2		0	0.231	0.197	0.104
3			0	0.150	0.285	0.102	{3,6}			0	0.182	0.335
4				0	0.277	0.215	4				0	0.277
5					0	0.386	5					0
6						0						

III. AŞAMA					IV. AŞAMA				V. AŞAMA		
	1	{2,5}	{3,6}	4		1	{2,5}	{3,6,4}		1	{2,5,3,6,4}
1	0	0.304	0.226	0.358	1	0	0.304	0.270	1	0	0.284
{2,5}		0	0.283	0.237	{2,5}		0	0.268	{2,5,3,6,4}		0
{3,6}			0	0.182	{3,6,4}			0			
4				0							

Örnek veri setinin ortalama bağlantı tekniği ile kümelenmesinde III. aşamada {2,5} kümesinin {3,6} kümesine olan uzaklığı $\frac{d(2,3) + d(2,6) + d(5,3) + d(5,6)}{2 \times 2}$ ile hesaplanarak 0.283 olarak bulunmuştur. Örnek veri setindeki birimlerin ortalama bağlantı kümeleme tekniğine göre bağlanma aşamaları Şekil 3.10 (a)’da iki boyutlu düzlemde gösterilirken, Şekil 3.10 (b)’de Öklid uzaklığı kullanılarak gerçekleştirilen ortalama bağlantı kümeleme tekniğinden elde edilen ağaç grafiğine yer verilmiştir.



Şekil 3.10. Ortalama bağlantı kümeleme tekniği ile birimlerin bağlanmasının (a) iki boyutlu düzlemde ve (b) ağaç grafiği ile gösterimi

3.1.4. Ağırlıklı Ortalama Bağlantı Kümeleme Tekniği (Weighted Average Linkage Method)

Ağırlıklı ortalama bağlantı kümeleme tekniği, aritmetik ortalama ile ağırlıklı çift grup yöntemi olarak da adlandırılır (Jain ve Dubes 1988). K_i kümesinin, K_i ve K_j kümelerinin bağlanması ile elde edilen $\{K_i \cup K_j\}$ kümesine olan uzaklığı ağırlıklı ortalama bağlantı tekniğinde,

$$d(K_i, \{K_i \cup K_j\}) = \frac{1}{2}d(K_i, K_i) + \frac{1}{2}d(K_i, K_j) \quad (3.5)$$

eşitliği ile hesaplanır. Tablo 3.1’de verilen örnek veri setindeki birimlerin ağırlıklı ortalama bağlantı kümeleme tekniğine göre bağlanma aşamalarında elde edilen uzaklık matrisleri Tablo 3.6’de verilmiştir.

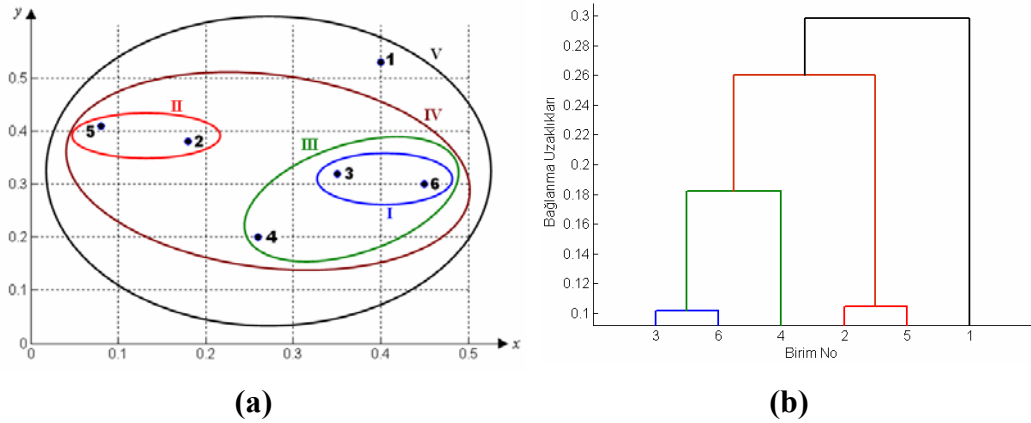
Tablo 3.6. Ağırlıklı ortalama bağlantı kümeleme tekniğine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri

I. AŞAMA							II. AŞAMA					
	1	2	3	4	5	6		1	2	{3,6}	4	5
1	0	0.266	0.216	0.358	0.342	0.235	1	0	0.266	0.226	0.358	0.342
2		0	0.180	0.197	0.104	0.282	2		0	0.231	0.197	0.104
3			0	0.150	0.285	0.102	{3,6}			0	0.182	0.335
4				0	0.277	0.215	4				0	0.277
5					0	0.386	5					0
6						0						

III. AŞAMA					IV. AŞAMA				V. AŞAMA			
	1	{2,5}	{3,6}	4		1	{2,5}	{3,6,4}		1	{2,5,3,6,4}	
1	0	0.304	0.226	0.358	1	0	0.304	0.292	1	0	0.298	
{2,5}		0	0.283	0.237	{2,5}		0	0.260	{2,5,3,6,4}		0	
{3,6}			0	0.182	{3,6,4}			0				
4				0								

Örnek veri setinin ağırlıklı ortalama bağlantı tekniği ile kümelenmesinde III. aşamada $\{3,6\}$ kümesi ile $\{4\}$ kümesi 0.182 bağlanma uzaklığı ile bir araya getirilmiştir. Kümelemenin bir sonraki aşamasında $\{2,5\}$ kümesinin, yeni oluşan $\{3,6,4\}$ kümesine olan uzaklığı $\frac{1}{2}d(\{2,5\}, \{3,6\}) + \frac{1}{2}d(\{2,5\}, \{4\})$ ile hesaplanarak 0.260

olarak bulunmuştur. Örnek veri setindeki birimlerin ağırlıklı ortalama bağlantı kümeleme tekniğine göre bağlanma aşamaları Şekil 3.11 (a)'da iki boyutlu düzlemde gösterilirken, Şekil 3.11 (b)'de Öklid uzaklığı kullanılarak gerçekleştirilen ağırlıklı ortalama bağlantı kümeleme tekniğinden elde edilen ağaç grafiğine yer verilmiştir.



Şekil 3.11. Ağırlıklı ortalama bağlantı kümeleme tekniği ile birimlerin bağlanmasının (a) iki boyutlu düzlemde ve (b) ağaç grafiği ile gösterimi

3.1.5. Merkezi Bağlantı Kümeleme Tekniği (Centroid Linkage Method)

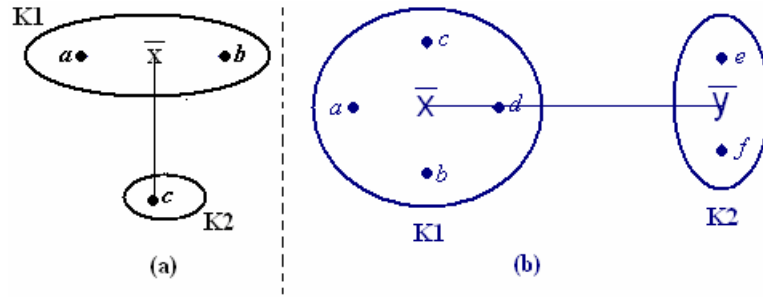
Merkezi bağlantı kümeleme tekniğinde iki küme arasındaki uzaklık, iki küme merkezi arasındaki uzaklık olarak tanımlanır. Küme merkezi olarak kümede yer alan birimler için hesaplanan aritmetik ortalama kullanılır. K_i kümesinin merkezi $\bar{x}$ ve K_j kümesinin merkezi $\bar{y}$ olarak alındığında, K_i kümesinin K_j kümesine olan uzaklığı merkez bağlantı tekniğine göre

$$d(K_i, K_j) = d(\bar{x}, \bar{y}) \quad (3.6)$$

şeklinde hesaplanır. Kümelemede dikkate alınan değişken sayısı p , K_i kümesini oluşturan n birimin oluşturduğu veri matrisi $\mathbf{X}_{n \times p}$ ve $\mathbf{j}$, $n \times 1$ boyutlu 1'lerden oluşan vektör olmak üzere K_i kümesinin küme merkezi $\bar{x}$,

$$\bar{x} = \frac{1}{n} \mathbf{X}' \mathbf{j} \quad (3.7)$$

eşitliği ile hesaplanır. Tek bağlantı, tam bağlantı, ortalama bağlantı ve ağırlıklı ortalama bağlantı kümeleme tekniklerinde küme bağlantı uzaklıkları monoton artan özelliktedir. Ancak merkez bağlantı tekniğinde küme bağlantı uzaklıkları monoton artan özellikte değildir. Yani herhangi bir aşamada elde edilen küme bağlantı uzaklığı önceki aşamada gerçekleşen küme bağlantı uzaklığından daha küçük olabilir. Bu da kümelerin yorumlanmasını güçleştirir. Şekil 3.12’de kümeler arasındaki uzaklıkların merkezi bağlantı kümeleme tekniğine göre belirlenmesi grafiksel olarak gösterilmiştir.



Şekil 3.12. Merkezi bağlantı kümeleme tekniğine göre iki küme arasındaki uzaklığın belirlenmesi

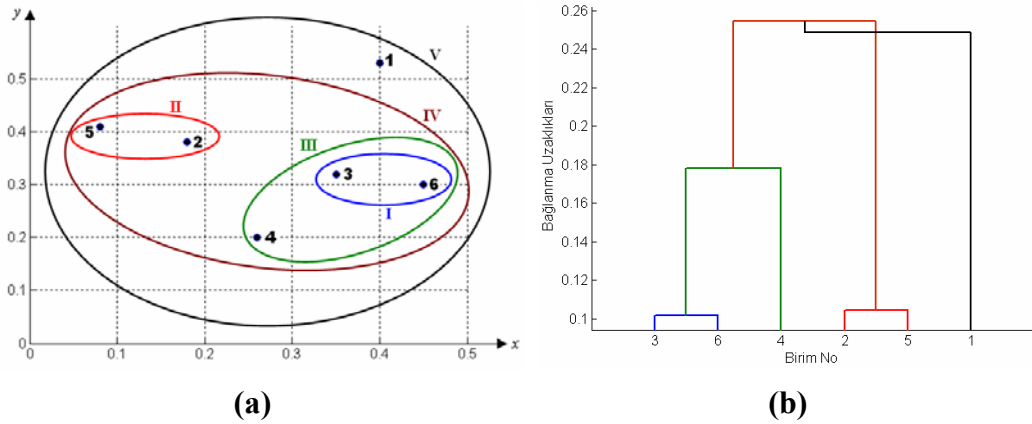
Tablo 3.1’de verilen örnek veri setindeki birimlerin merkezi bağlantı kümeleme tekniğine göre bağlanma aşamalarında elde edilen uzaklık matrisleri Tablo 3.7’de verilmiştir. Kümelemenin IV. aşamasında {2,5} kümesinin {3,6,4} kümesine olan uzaklığı, {2,5} kümesinin merkezi (0.130, 0.395) noktası ile {3,6,4} kümesinin merkezi (0.353, 0.273) noktaları arasındaki Öklid uzaklığı olarak $\left((0.130 - 0.353)^2 + (0.395 - 0.273)^2\right)^{0.5} = 0.254$ hesaplanmıştır. Tablo 3.6 incelendiğinde küme bağlanma uzaklıklarının merkezi bağlantı tekniğinde monoton artmayabileceği bu örnekte görülmektedir. IV. aşamada birleştirilen iki küme arasındaki uzaklık 0.254 iken V. aşamada birleştirilen iki küme arasındaki uzaklık 0.249’dur.

Tablo 3.7. Merkezi bağlantı kümeleme tekniğine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri

I. AŞAMA							II. AŞAMA					
	1	2	3	4	5	6		1	2	{3,6}	4	5
1	0	0.266	0.216	0.358	0.342	0.235	1	0	0.266	0.220	0.358	0.342
2		0	0.180	0.197	0.104	0.282	2		0	0.231	0.197	0.104
3			0	0.150	0.285	0.102	{3,6}			0	0.178	0.335
4				0	0.277	0.215	4				0	0.277
5					0	0.386	5					0
6						0						

III. AŞAMA					IV. AŞAMA				V. AŞAMA		
	1	{2,5}	{3,6}	4		1	{2,5}	{3,6,4}		1	{2,5,3,6,4}
1	0	0.302	0.220	0.358	1	0	0.302	0.261	1	0	0.249
{2,5}		0	0.283	0.234	{2,5}		0	0.254	{2,5,3,6,4}		0
{3,6}			0	0.178	{3,6,4}			0			
4				0							

Örnek veri setindeki birimlerin merkezi bağlantı kümeleme tekniğine göre bağlanma aşamaları Şekil 3.13 (a)'da iki boyutlu düzlemde gösterilirken, Şekil 3.13 (b)'de Öklid uzaklığı kullanılarak gerçekleştirilen merkezi bağlantı kümeleme tekniğinden elde edilen ağaç grafiğine yer verilmiştir.



Şekil 3.13. Merkezi bağlantı kümeleme tekniği ile birimlerin bağlanmasının (a) iki boyutlu düzlemde ve (b) ağaç grafiği ile gösterimi

Şekil 13. (b) incelendiğinde merkezi bağlantı tekniği ile elde edilen ağaç grafiğindeki bağlanma uzaklıklarının monoton artmadığı açıkça görülmektedir.

3.1.6. Medyan Bağlantı Kümeleme Tekniği (Median Linkage Method)

Merkez bağlantı kümeleme tekniğinde iki küme bağlandığında yeni oluşan kümenin merkezi, bağlanan kümelerden birim sayısı fazla olan kümenin merkezine yakın olmaktadır. Gower (1967) iki kümenin birleşmesi ile oluşan kümelerin merkezlerinin belirlenmesinde birim sayısı fazla olan kümenin etkisini ortadan kaldırmak için medyan bağlantı kümeleme tekniğini önermiştir. İki kümenin bağlanması ile oluşan yeni küme merkezi medyan tekniğine göre

$$\bar{x}_{\{K_i, K_j\}} = \frac{1}{2}(\bar{x}_{\{K_i\}} + \bar{x}_{\{K_j\}}) \quad (3.8)$$

eşitliği ile belirlenir. Eşitlikte yer alan $\bar{x}_{\{K_i\}}$ gösterimi ile K_i küme merkezi ifade edilirken diğer gösterimlerde benzer şekilde tanımlanmaktadır. Teknik her ne kadar medyan bağlantı tekniği olarak isimlendirilirse de burada medyan istatistik biliminde kullanılan “küçükten büyüğe doğru sıralanmış dizideki ortanca terim” tanımına uymamaktadır. Medyan bağlantı kümeleme tekniğine göre K_l kümesinin, K_i ve K_j kümelerinin bağlanması ile elde edilen $\{K_i \cup K_j\}$ kümesine olan uzaklığı

$$d(K_l, (K_i \cup K_j)) = \frac{1}{2}d(K_l, K_i) + \frac{1}{2}d(K_l, K_j) - \frac{1}{4}d(K_i, K_j) \quad (3.9)$$

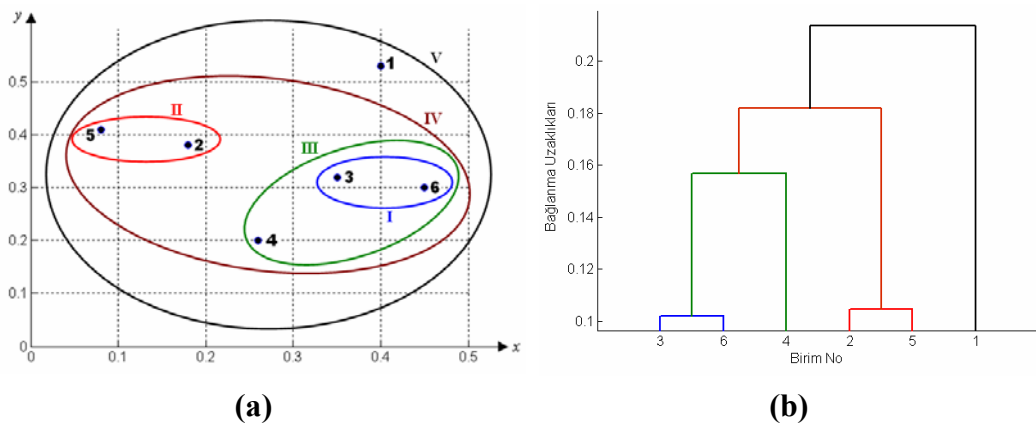
eşitliği ile hesaplanır. Medyan bağlantı kümeleme tekniğinde de merkez bağlantı kümeleme tekniğine göre bağlanma uzaklıkları monoton artan olmayabilir (Martinez ve Martinez 2005). Tablo 3.1’de verilen örnek veri setindeki birimlerin medyan bağlantı kümeleme tekniğine göre bağlanma aşamalarında elde edilen uzaklık matrisleri Tablo 3.8’de verilmiştir.

Tablo 3.8. Medyan bağlantı kümeleme tekniğine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri

I. AŞAMA							II. AŞAMA					
	1	2	3	4	5	6		1	2	{3,6}	4	5
1	0	0.266	0.216	0.358	0.342	0.235	1	0	0.266	0.200	0.358	0.342
2		0	0.180	0.197	0.104	0.282	2		0	0.205	0.197	0.104
3			0	0.150	0.285	0.102	{3,6}			0	0.157	0.310
4				0	0.277	0.215	4				0	0.277
5					0	0.386	5					0
6						0						

III. AŞAMA					IV. AŞAMA			V. AŞAMA			
	1	{2,5}	{3,6}	4		1	{2,5}	{3,6,4}		1	{2,5,3,6,4}
1	0	0.278	0.200	0.358	1	0	0.278	0.240	1	0	0.214
{2,5}		0	0.232	0.211	{2,5}		0	0.182	{2,5,3,6,4}		0
{3,6}			0	0.157	{3,6,4}			0			
4				0							

Kümelemenin IV. aşamasında $\{2,5\}$ kümesinin $\{3,6,4\}$ kümesine olan uzaklığı, $\frac{1}{2}(d(\{2,5\}, \{3,6\}) + d(\{2,5\}, \{4\})) - \frac{1}{4}d(\{3,6\}, \{4\})$ eşitliği ile hesaplanmıştır. Örnek veri setindeki birimlerin medyan bağlantı kümeleme tekniğine göre bağlanma aşamaları Şekil 3.14 (a)'da iki boyutlu düzlemde gösterilirken, Şekil 3.14 (b)'de Öklid uzaklığı kullanılarak gerçekleştirilen medyan bağlantı kümeleme tekniğinden elde edilen ağaç grafiğine yer verilmiştir.



Şekil 3.14. Medyan bağlantı kümeleme tekniği ile birimlerin bağlanmasının (a) iki boyutlu düzlemde ve (b) ağaç grafiği ile gösterimi

3.1.7. Ward Bağlantı Kümeleme Tekniği (Ward Linkage Method)

Ward (1963), iki kümenin birleşmesinde oluşacak bilgi kaybını minimize etmeyi amaçlayan genel bir aşamalı kümeleme tekniği önermiştir. Literatürde genelde Ward bağlantı kümeleme tekniği olarak isimlendirilen bu teknik en küçük varyanslı kümeleri birleştirdiği için “minimum varyans bağlantı” tekniği olarak da bilinir (Gan ve ark 2007). K_i ve K_j kümeleri birleştirilerek oluşturulan yeni küme $\{K_i \cup K_j\}$ olsun. Bu durumda bilgi kaybındaki artış,

$$I_{\{K_i \cup K_j\}} = \frac{n_i n_j}{n_i + n_j} (\bar{\mathbf{x}}_{K_i} - \bar{\mathbf{x}}_{K_j})' (\bar{\mathbf{x}}_{K_i} - \bar{\mathbf{x}}_{K_j}) \quad (3.10)$$

eşitliği ile hesaplanır (Rencher 2002). Eşitlikte yer alan $\bar{\mathbf{x}}_{K_i}$ ve $\bar{\mathbf{x}}_{K_j}$ sırasıyla K_i ve K_j küme merkezlerini gösterirken n_i ve n_j gösterimleri K_i ve K_j kümelerindeki birim sayılarını ifade etmektedir. Ward bağlantı kümeleme tekniğine göre K_i kümesinin, K_i ve K_j kümelerinin bağlanması ile elde edilen $\{K_i \cup K_j\}$ kümesine olan uzaklığı

$$d(K_i, \{K_i \cup K_j\}) = \frac{(n_i + n_j)d(K_i, K_i) + (n_i + n_j)d(K_i, K_j) - n_i d(K_i, K_j)}{n_i + n_i + n_j} \quad (3.11)$$

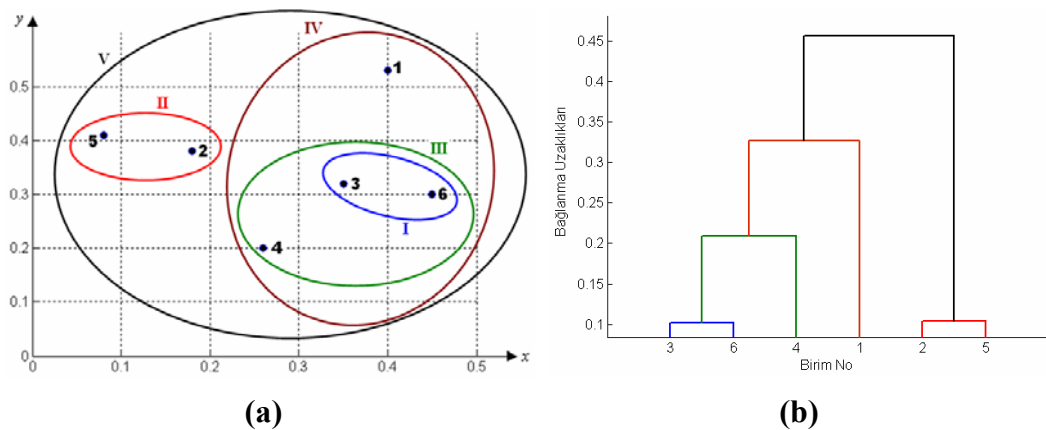
şeklinde hesaplanır. Birim sayısı az olan kümeleri birleştirme eğilimde olan Ward bağlantı kümeleme tekniğinde birim sayıları yaklaşık olarak eşit olan küresel kümeler oluşturmaktadır (Martinez ve Martinez 2005). Tablo 3.1’de verilen örnek veri setindeki birimlerin Ward bağlantı kümeleme tekniğine göre bağlanma aşamalarında elde edilen uzaklık matrisleri Tablo 3.9’da verilmiştir.

Tablo 3.9. Ward bağlantı kümeleme tekniğine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri

I. AŞAMA							II. AŞAMA					
	1	2	3	4	5	6		1	2	{3,6}	4	5
1	0	0.266	0.216	0.358	0.342	0.235	1	0	0.266	0.267	0.358	0.342
2		0	0.180	0.197	0.104	0.282	2		0	0.274	0.197	0.104
3			0	0.150	0.285	0.102	{3,6}			0	0.209	0.413
4				0	0.277	0.215	4				0	0.277
5					0	0.386	5					0
6						0						

III. AŞAMA				IV. AŞAMA				V. AŞAMA		
	1	{2,5}	{3,6}	4		1	{2,5}	{3,6,4}	{1,3,6,4}	{2,5}
1	0	0.371	0.267	0.358	1	0	0.371	0.327		
{2,5}		0	0.463	0.281	{2,5}		0	0.455		0.456
{3,6}			0	0.209	{3,6,4}			0		0
4				0						

Kümelemenin V. aşamasında $\{2,5\}$ kümesinin $\{1,3,6,4\}$ kümesine olan uzaklığı, $\frac{3}{6}d(\{2,5\}, \{1\}) + \frac{5}{6}d(\{2,5\}, \{3,6,4\}) - \frac{2}{6}d(\{1\}, \{3,6,4\})$ eşitliği ile hesaplanmıştır. Örnek veri setindeki birimlerin Ward bağlantı kümeleme tekniğine göre bağlanma aşamaları Şekil 3.15 (a)'da iki boyutlu düzlemde gösterilirken, Şekil 3.15 (b)'de Öklid uzaklığı kullanılarak gerçekleştirilen Ward bağlantı kümeleme tekniğinden elde edilen ağaç grafiğine yer verilmiştir.



Şekil 3.15. Ward bağlantı kümeleme tekniği ile birimlerin bağlanmasının (a) iki boyutlu düzlemde ve (b) ağaç grafiği ile gösterimi

3.1.8. Esnek Beta Tekniği

K_i kümesinin $\{K_i \cup K_j\}$ kümesine olan uzaklığının belirlenmesi için önerilen çok sayıda aşamalı kümeleme tekniği bulunmaktadır. Lance ve Williams (1967), K_i kümesinin $\{K_i \cup K_j\}$ kümesine olan uzaklığını belirlemek için önerilen aşamalı kümeleme tekniklerine,

$$d(K_i, \{K_i \cup K_j\}) = \alpha_i d(K_i, K_i) + \alpha_j d(K_i, K_j) + \beta d(K_i, K_j) + \gamma |d(K_i, K_i) - d(K_i, K_j)| \quad (3.12)$$

eşitliğindeki α_i , α_j , β ve γ parametrelerine uygun değerler verilerek ulaşılabileceğini göstermişlerdir. Lance ve Williams (1967), iyi bir kümeleme elde edebilmek için eşitlik 3.12'deki parametreler için aşağıdaki kısıtlamaları önermişlerdir.

$$\alpha_i + \alpha_j + \beta = 1$$

$$\alpha_i = \alpha_j$$

$$\gamma = 0$$

$$\beta < 1$$

Parametrelerin değeri bu kısıtlamalar ile sadece β parametresine bağlı kaldığından tanımlanan bu aşamalı kümeleme tekniği esnek beta tekniği olarak isimlendirilmektedir. Lance ve Williams (1967), β parametresi için $\beta = -0.25$ gibi küçük negatif değerlerin kullanılmasını önermişlerdir. Eğer veri seti içerisinde aykırı değerler varsa bu aykırı değerlerin kümelemeye olan etkisini azaltmak için β parametresinin $\beta = -0.5$ gibi daha küçük değerlerinin kullanılmasının uygun olacağını belirtmişlerdir.

Eşitlik (3.12)'deki α_i , α_j , β ve γ parametrelerinin kullanılan tekniğe göre aldığı değerler n_i , n_j ve n_l sırasıyla K_i , K_j ve K_l kümelerindeki birim sayıları olmak üzere Tablo 3.10'da verilmiştir.

Tablo 3.10. Lance ve Williams'ın önerdiği genel formüldeki parametrelerin bağlantı tekniğine göre değerleri

Aşamalı Kümeleme Tekniği	α_i	α_j	β	γ
Tek Bağlantı	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Tam Bağlantı	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Ortalama Bağlantı	$\frac{n_i}{n_i + n_j}$	$\frac{n_i}{n_i + n_j}$	0	0
Ağırlıklı Ortalama Bağlantı	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Merkezi Bağlantı	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	$\frac{-n_i n_j}{(n_i + n_j)^2}$	0
Medyan Bağlantı	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward Bağlantı	$\frac{n_i + n_l}{n_i + n_j + n_l}$	$\frac{n_j + n_l}{n_i + n_j + n_l}$	$\frac{-n_l}{n_i + n_j + n_l}$	0
Esnek Beta Tekniği	$\frac{1 - \beta}{2}$	$\frac{1 - \beta}{2}$	$\beta(<1)$	0

3.2. Aşamalı Olmayan Kümeleme Teknikleri

Birim sayısının fazla olması durumunda, birimleri adım adım birbirine bağlayarak küme yapısını ortaya çıkarmak oldukça zaman almaktadır. Kümelenmesi gereken çok sayıda birimin bulunduğu uygulamalarda, ağaç grafiğinin oluşturulması gibi çok ağır bir iş yükü getiren algoritmalar yerine aşamalı olmayan kümeleme teknikleri kullanılmaktadır.

Birimlerin birbirleri ile olan benzerliklerini benzer grup ortalama vektörleri tanımlayarak kendi içinde homojen ve aralarında heterojen gruplamaları belirlemeyi amaçlayan kümeleme tekniklerine aşamalı olmayan kümeleme teknikleri

denilmektedir (Hartigan 1975). Aşamalı olmayan kümeleme tekniklerinde küme sayısı k önceden tanımlanarak, k adet küme merkezi belirlenir ve n birimden oluşan veri seti k kümeye(parçaya) bölünür. Parçalanmaya dayanan kümeleme teknikleri olarak da isimlendirilen bu tekniklerin tamamı küme merkezlerinin kümeyi temsil etmesi esasına dayanır.

Aşamalı olmayan kümeleme teknikleri veriyi önceden tanımlanan küme sayısına göre kümelendirmede oldukça etkili ve basit teknikler olmasına karşın küme sayısının önceden belirlenmesi ve küme başlangıç merkezlerinin seçiminin keyfi olması gibi dezavantajlara sahiptir (Aldenferder ve Blashfield 1984).

Aşamalı olmayan kümeleme teknikleri kümeleri oluştururken çoğunlukla bir amaç fonksiyonunu optimize etmeye çalışır. Bu fonksiyonun muhtemel en iyi değerini bulmak için tüm birimler üzerinde tüm kombinasyonları denemek mümkün değildir. Bunun yerine amaç fonksiyonunu optimize etmek için farklı algoritmalar geliştirilmiştir. Aşamalı olmayan kümeleme yaklaşımında üzerinde en çok çalışılan merkez tabanlı kümeleme yaklaşımıdır (Ünler 2006). Merkez tabanlı kümeleme yaklaşımı yanında küme içi kareler ve çarpımlar matrisi $\mathbf{W}$ ile kümeler arası kareler ve çarpımlar matrisi $\mathbf{B}$ olmak üzere;

- $\min(\text{iz}(\mathbf{W}))$
- $\min|\mathbf{W}|$
- $\max(\text{iz}(\mathbf{W}^{-1}\mathbf{B}))$

kriterlerini sağlamaya yönelik geliştirilen parçalama teknikleri bulunmaktadır (Rencher 2002).

Merkez tabanlı kümelemede en temel kümeleme algoritması k - ortalamalar kümeleme algoritmasıdır. Bunun yanında bulanık k -ortalamalar ve k -harmonik ortalama, k-medoids algoritmaları gibi k - ortalamalar algoritmasından türetilmiş çok sayıda esnek kümeleme algoritmaları vardır. Bu çalışma kapsamında literatürde en fazla bilinen ve kullanılan k -ortalamalar algoritması üzerinde çalışılacağından aşamalı olmayan kümele tekniklerinden k -ortalamalar algoritması anlatılacaktır.

3.2.1. k -Ortalamalar Kümeleme Algoritması

k -ortalamalar algoritması kümeleme analizinde en yaygın olarak kullanılan bir kümeleme algoritmasıdır (Duda ve ark. 2001, Theodoridis ve Koutroumbas 2006). Farklı bilimsel alanlarda birbirlerinden bağımsız bir şekilde Steinhaus (1955), Lloyd (1957), Ball ve Hall (1965) ile Mac Queen (1967)'in çalışmalarında k -ortalamalar algoritması ilk olarak kullanılmıştır. Fakat Mac Queen (1967) çalışmasında “ k -ortalamalar” kavramını kullandığı için çoğu kaynakta algoritmayı ilk öneren olarak onun ismi geçmektedir (Jain 2008). Ayrıca Hartigan ve Wong (1979)'un k -ortalamalar tekniği ile ilgili çalışması yöntemin gelişimine önemli katkılar sağlamıştır.

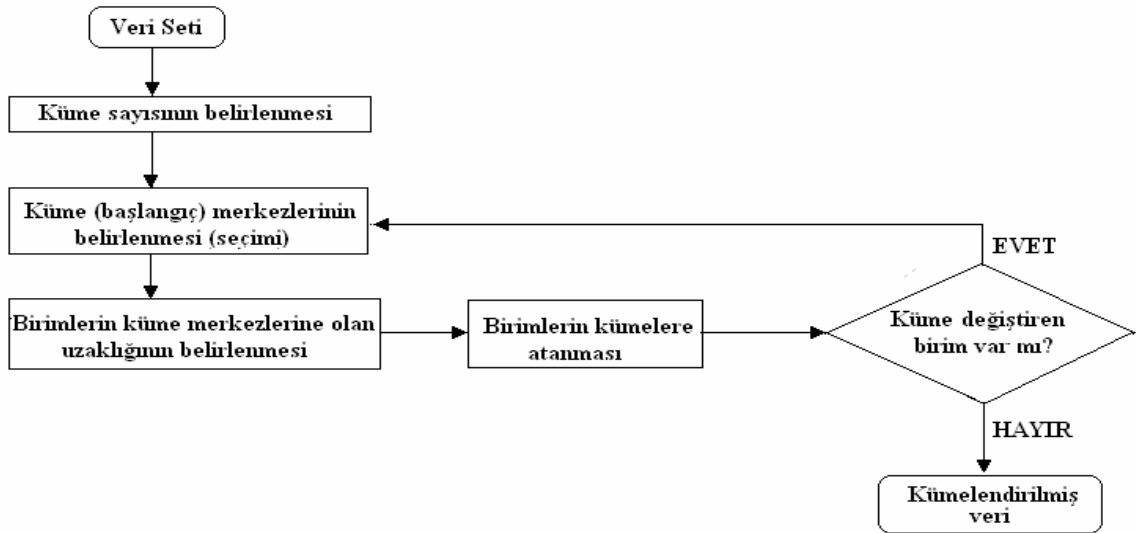
k -ortalamalar algoritması, hata fonksiyonunu minimum yapılmasına dayalı olarak verideki en uygun parçalanmayı bulmayı amaçlayan adımsal optimizasyon yöntemidir. Teknik ile verinin kümelenmesi küme sayısı k 'nın belirlenmesi ile başlar. Algoritmanın ikinci adımında küme başlangıç merkezleri seçilir ve birimlerin küme merkezlerine olan uzaklıkları hesaplanır. Küme merkezlerine yakınlığa göre birimler kümeler atanır ve yeni oluşan kümelerin merkezleri belirlenir. Belirlenen küme merkezlerine olan uzaklıklar hesaplanarak tekrar atamalar yapılır. Bu süreç küme üyelikleri sabitleninceye kadar veya hata fonksiyonunun değeri değişmeyinceye kadar devam eder (Hartigan 1975, Hartigan ve Wong 1979).

X , $n \times p$ boyutlu veri matrisi ve $K_1, K_2, \dots, K_k$ bu veri matrisindeki ayrık kümeler olmak üzere hata fonksiyonu,

$$E = \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} d(\mathbf{x}, \mu(K_i)) \quad (3.13)$$

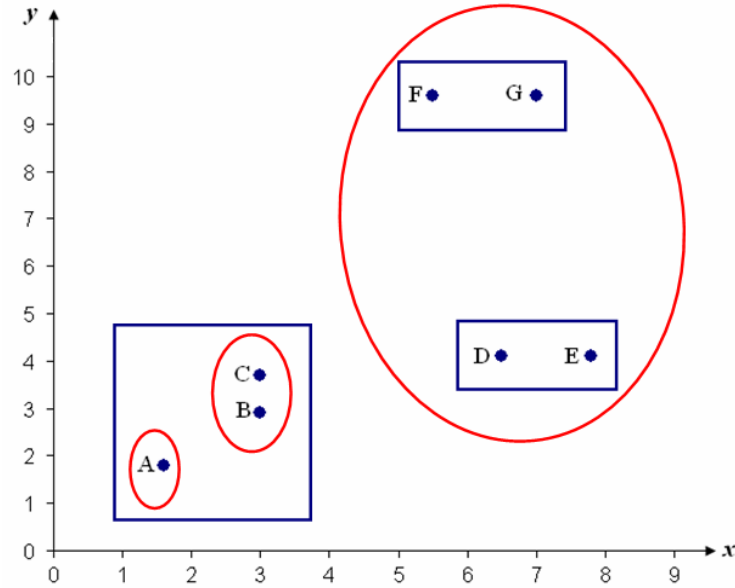
eşitliği ile ifade edilir. Eşitlikte yer alan $\mu(K_i)$ gösterimi ile K_i küme merkezinin hesaplanılmasını sağlayan ortalama fonksiyonu, $d(\mathbf{x}, \mu(K_i))$ gösterimi ile de K_i kümesinde yer alan ve $\mathbf{x}$ gözlem vektörüne sahip birimin K_i küme merkezine olan uzaklığının hesaplanıldığı uzaklık fonksiyonu ifade edilmektedir. Genellikle küme

merkezlerinin belirlenmesinde aritmetik ortalama, birimlerin küme merkezlerine olan uzaklıklarının hesaplanılmasında Öklid uzaklığı kullanılmaktadır (Gan ve ark. 2007). k -ortalamlar algoritmasına ait akış şeması Şekil 3.16’da verilmiştir.



Şekil 3.16 . k -ortalamlar algoritmasının akış şeması

k -ortalamlar algoritmasının 1. adımında küme sayısının belirlenmesi ve küme başlangıç merkezlerinin seçimi kümeleme sonucunu doğrudan etkileyen faktörlerdendir. Küme başlangıç merkezlerinin seçiminin rasgele olması aynı verinin kümelenmesinde farklı kümeleme sonuçları ile karşılaşılmasına neden olabileceği gibi bulunan çözüm yerel minimum olabilir. Örnek vermek gerekirse A(1.6,1.8), B(3.0,2.9), C(3.0,3.7), D(6.5,4.1), E(7.8,4.1), F(5.5,9.6) ve G(7.0,9.6) noktalarından oluşan ve Şekil 3.17’de gösterilen veri setinin kümelenmesinde başlangıç noktaları olarak A, B, ve C seçilirse Öklid uzaklığıyla k -ortalama algoritmasının ürettiği kümeler elips şekilleri ile gösterilmiştir. Bu durumda oluşan kümeler {A}, {B,C} ve {D, E, F,G} şeklindedir.



Şekil 3.17. k - ortalamlar algoritmasında küme başlangıç merkezlerinin seçiminin kümeleme etkisi (Tantuğ 2002)

Oysa bu veri setinin en iyi kümelenmesi $\{A,B,C\}$, $\{D,E\}$ ve $\{F,G\}$ biçimindedir. Bu bölümlenme Şekil 3.17’de dikdörtgenler ile gösterilmiştir ve göz ile de rahatlıkla yapılabilir. İlk durumda üretilen kümelerin hata fonksiyonunun değeri 12.29, ikinci bölümlenmede elde edilen kümelerin hata fonksiyonunun değeri 5.66’dır. Yani ilk kümeleme işlemi sonucunda bu fonksiyonun yerel bir en küçük değeri elde edilmiştir ancak bu değer en küçük değildir. En küçük değerini bulmak için seçilmesi gereken başlangıç noktaları A, D ve F olmaktadır. Görüldüğü gibi başlangıç küme merkezlerinin seçimi, algoritmanın en iyi kümelemeyi sağlayacak şekilde hata fonksiyonunu minimize etmede doğrudan etkilidir.

k -ortalamlar algoritmasının pratikte işleyişini göstermek için oluşturulan iki boyutlu örnek veri seti Tablo 3.11’de verilmiştir.

Tablo 3.11. k -ortalamalar algoritmasının uygulaması için oluşturulan örnek veri seti

Gözlem No (i)	x_i	y_i
1	0.375	1.667
2	0.750	0.500
3	1.000	0.500
4	1.125	0.667
5	1.000	1.000
6	0.750	1.333
7	0.750	2.333
8	0.750	4.167
9	2.125	2.000
10	2.750	1.500
11	3.125	4.500
12	3.375	0.833
13	3.750	0.501
14	3.375	2.167
15	3.420	2.835
16	4.060	1.375
17	3.875	3.750
18	4.000	4.500
19	4.125	4.000
20	4.150	1.800
21	4.250	0.525
22	4.375	1.000
23	4.875	0.175
24	4.250	4.167
25	4.375	3.750
26	4.400	4.350
27	4.625	3.325

Örnek veri setini k -ortalamalar algoritması ile üç kümeye ayırmak için seçilen küme merkezleri $\mathbf{m}_1 = [3.0 \ 0.5]$, $\mathbf{m}_2 = [3.5 \ 3.5]$ ve $\mathbf{m}_3 = [1.5 \ 2.5]$ olsun. Verinin k -ortalamalar algoritması ile kümelenmesinde küme başlangıç merkezlerinin seçiminden sonra gözlemlerin seçilen küme merkezlerine olan Öklid uzaklıkları hesaplanır ve gözlemler en yakın oldukları kümelere atanır. Gözlemlerin seçilen küme başlangıç merkezlerine göre kümelere atanması ile ilgili hesaplamalar Tablo 3.12’de verilmiştir.

Tablo 3.12. Örnek veri setinin k -ortalamalar algoritması ile kümeleneşinde I. adımda gerçekleştirilen hesaplamalar ve gözlemlerin kümelere atanması

Gözlem No (i)	x_i	y_i	$d(i, K_1)$	$d(i, K_2)$	$d(i, K_3)$	Küme Üyelikleri
1	0.375	1.667	2.873	3.623	1.400	3
2	0.750	0.500	2.250	4.070	2.136	3
3	1.000	0.500	2.000	3.905	2.062	1
4	1.125	0.667	1.882	3.697	1.871	3
5	1.000	1.000	2.062	3.536	1.581	3
6	0.750	1.333	2.399	3.501	1.387	3
7	0.750	2.333	2.902	2.987	0.768	3
8	0.750	4.167	4.302	2.830	1.828	3
9	2.125	2.000	1.737	2.035	0.800	3
10	2.750	1.500	1.031	2.136	1.601	1
11	3.125	4.500	4.002	1.068	2.577	2
12	3.375	0.833	0.502	2.670	2.509	1
13	3.750	0.501	0.750	3.009	3.010	1
14	3.375	2.167	1.709	1.339	1.904	2
15	3.420	2.835	2.372	0.670	1.949	2
16	4.060	1.375	1.374	2.198	2.796	1
17	3.875	3.750	3.366	0.451	2.684	2
18	4.000	4.500	4.123	1.118	3.202	2
19	4.125	4.000	3.676	0.800	3.023	2
20	4.150	1.800	1.736	1.820	2.741	1
21	4.250	0.525	1.250	3.068	3.386	1
22	4.375	1.000	1.463	2.649	3.243	1
23	4.875	0.175	1.903	3.598	4.098	1
24	4.250	4.167	3.874	1.004	3.216	2
25	4.375	3.750	3.529	0.910	3.135	2
26	4.400	4.350	4.097	1.238	3.440	2
27	4.625	3.325	3.259	1.139	3.232	2

Gözlemlerin kümelere atanmasından sonra oluşan yeni küme merkezleri $\mathbf{m}_1 = [3.621 \ 0.912]$, $\mathbf{m}_2 = [3.666 \ 3.440]$ ve $\mathbf{m}_3 = [0.982 \ 1.881]$ olarak belirlenir. Kümelemenin I. adımında hata fonksiyonunun değeri yani gözlemlerin kendi küme merkezlerine olan Öklid uzaklıkları toplamı 33.517'dir. Kümeleme gözlemlerin yeni oluşan küme merkezlerine olan Öklid uzaklığının hesaplanması ve kümelere tekrar atanması ile devam eder. Kümelemenin II. adımda gerçekleştirilen hesaplamalar Tablo 3.13'de verilmiştir.

Tablo 3.13. Örnek veri setinin k -ortalamalar algoritması ile kümeleneğinde II. adımda gerçekleştirilen hesaplamalar ve gözlemlerin kümelere atanması

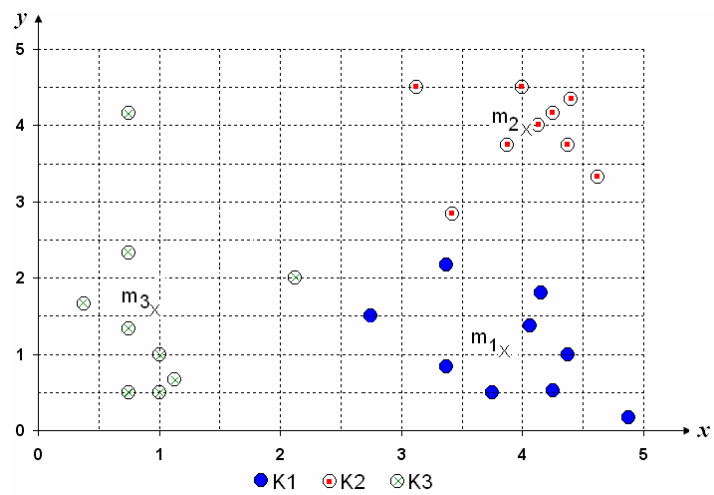
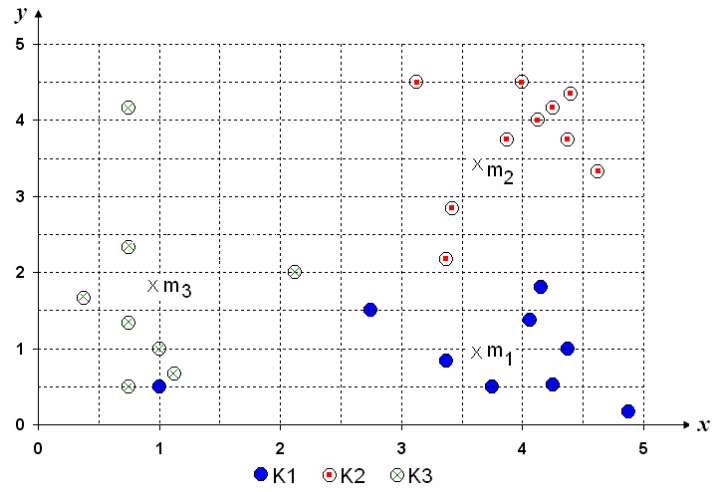
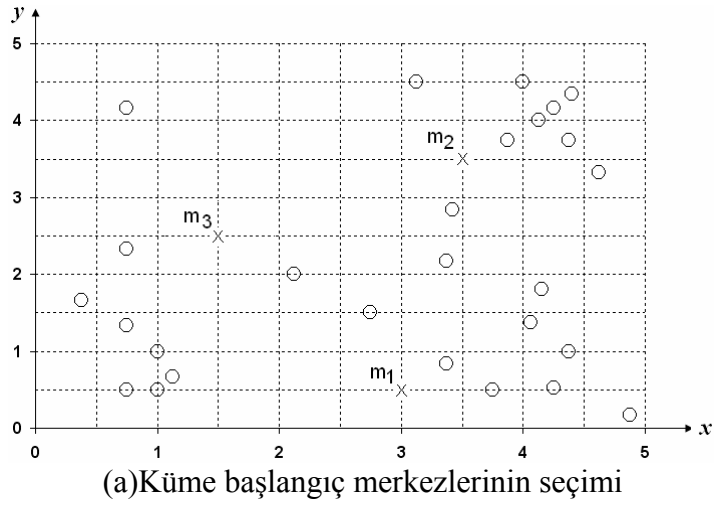
Gözlem No (i)	x_i	y_i	$d(i, K_1)$	$d(i, K_2)$	$d(i, K_3)$	Küme Üyelikleri	Önceki Küme Üyelikleri
1	0.375	1.667	3.332	3.738	0.644	3	3
2	0.750	0.500	2.900	4.141	1.400	3	3
3	1.000	0.500	2.653	3.969	1.381	3	1
4	1.125	0.667	2.508	3.761	1.222	3	3
5	1.000	1.000	2.622	3.614	0.881	3	3
6	0.750	1.333	2.901	3.597	0.595	3	3
7	0.750	2.333	3.203	3.119	0.508	3	3
8	0.750	4.167	4.340	3.005	2.298	3	3
9	2.125	2.000	1.849	2.109	1.149	3	3
10	2.750	1.500	1.050	2.145	1.808	1	1
11	3.125	4.500	3.622	1.190	3.384	2	2
12	3.375	0.833	0.258	2.623	2.612	1	1
13	3.750	0.501	0.431	2.941	3.093	1	1
14	3.375	2.167	1.279	1.306	2.410	1	2
15	3.420	2.835	1.933	0.653	2.618	2	2
16	4.060	1.375	0.638	2.103	3.119	1	1
17	3.875	3.750	2.849	0.374	3.444	2	2
18	4.000	4.500	3.608	1.111	3.996	2	2
19	4.125	4.000	3.129	0.724	3.790	2	2
20	4.150	1.800	1.034	1.710	3.169	1	1
21	4.250	0.525	0.739	2.973	3.538	1	1
22	4.375	1.000	0.760	2.541	3.505	1	1
23	4.875	0.175	1.455	3.482	4.250	1	1
24	4.250	4.167	3.315	0.933	3.988	2	2
25	4.375	3.750	2.936	0.774	3.874	2	2
26	4.400	4.350	3.525	1.169	4.216	2	2
27	4.625	3.325	2.614	0.966	3.919	2	2

II. adımdaki kümeleme sonrası küme değiştiren gözlem veya gözlemler olup olmadığı incelendiğinde 3. ve 14. gözlemlerin küme değiştirdiği görülmektedir. Bu aşamada küme değiştiren birimler olduğu için yeni oluşan küme merkezleri hesaplanır. Hesaplanan yeni küme merkezleri $\mathbf{m}_1 = [3.884 \ 1.097]$, $\mathbf{m}_2 = [4.022 \ 3.909]$ ve $\mathbf{m}_3 = [0.958 \ 1.574]$ olarak belirlenir. Kümelemenin II. adımında hata fonksiyonunun değeri 25.617'dir. Kümeleme gözlemlerin yeni oluşan küme merkezlerine olan Öklid uzaklığının hesaplanması ve kümelere tekrar atanması ile devam eder. Kümelemenin III. adımda gerçekleştirilen hesaplamalar Tablo 3.14'de verilmiştir.

Tablo 3.14. Örnek veri setinin k -ortalamalar algoritması ile kümelenmesinde III. adımda gerçekleştirilen hesaplamalar ve gözlemlerin kümelere atanması

Gözlem No (i)	x_i	y_i	$d(i, K_1)$	$d(i, K_2)$	$d(i, K_3)$	Küme Üyelikleri	Önceki Küme Üyelikleri
1	0.375	1.667	3.555	4.281	0.591	3	3
2	0.750	0.500	3.191	4.725	1.094	3	3
3	1.000	0.500	2.946	4.555	1.075	3	3
4	1.125	0.667	2.793	4.347	0.922	3	3
5	1.000	1.000	2.886	4.194	0.576	3	3
6	0.750	1.333	3.143	4.164	0.319	3	3
7	0.750	2.333	3.369	3.631	0.787	3	3
8	0.750	4.167	4.387	3.282	2.601	3	3
9	2.125	2.000	1.977	2.691	1.242	3	3
10	2.750	1.500	1.204	2.724	1.793	1	1
11	3.125	4.500	3.486	1.074	3.641	2	2
12	3.375	0.833	0.574	3.143	2.528	1	1
13	3.750	0.501	0.611	3.418	2.991	1	1
14	3.375	2.167	1.185	1.858	2.488	1	1
15	3.420	2.835	1.799	1.231	2.766	2	2
16	4.060	1.375	0.329	2.534	3.108	1	1
17	3.875	3.750	2.653	0.216	3.639	2	2
18	4.000	4.500	3.405	0.592	4.220	2	2
19	4.125	4.000	2.913	0.138	3.989	2	2
20	4.150	1.800	0.751	2.112	3.200	1	1
21	4.250	0.525	0.679	3.391	3.455	1	1
22	4.375	1.000	0.500	2.930	3.465	1	1
23	4.875	0.175	1.353	3.830	4.159	1	1
24	4.250	4.167	3.091	0.345	4.190	2	2
25	4.375	3.750	2.698	0.387	4.051	2	2
26	4.400	4.350	3.293	0.581	4.422	2	2
27	4.625	3.325	2.348	0.839	4.063	2	2

III. adımdaki kümeleme sonrası küme değiştiren gözlem veya gözlemler olmadığından algoritma durdurulur ve nihai kümeler oluşturulur. Kümelemenin III. adımında hata fonksiyonunun değeri 21.796'dır. Örnek veri setinin k -ortalamalar algoritması ile kümelenmesinde adım adım küme merkezleri ve küme üyeliklerinin değişimi Şekil 3.18'de verilmiştir.



Şekil 3.18. Örnek veri setinin k -ortalamalar algoritması ile kümelенmesinde küme merkezleri ve küme üyeliklerinin değişiminin gösterimi

3.3. Kümelerin Değerlendirilmesi

Kümeleme analizinde kullanılan farklı algoritmalarla veya aynı algorithmada alınan farklı parametre başlangıç değerleri ile aynı verinin kümelenmesinde farklı birçok kümeleme yapısı elde edilebilir. Aynı veri kümesinden elde edilen farklı kümeleme yapılarından hangisinin geçerli olup olmadığı kümeleme geçerliliği olarak adlandırılır (Gordon 1998, Halkidi ve ark. 2002a, Halkidi ve ark. 2002b).

Genel anlamda, kümelemenin geçerliliğini incelemek için 3 yaklaşım mevcuttur. Bunlar dışsal (external) kriterler, içsel (internal) kriterler ve oransal (relative) kriterlerdir (Jain ve Dubes 1988, Theodoridis ve Koutroumbas 2006, Halkidi ve ark 2002a). Dışsal (external) kriterler, veri hakkında sahip olunan ön bilgi ile kümeleme algoritması sonunda elde edilen (C) kümeleme yapısı karşılaştırır. İçsel (internal) kriterler ise veri hakkında bir ön bilgi olmaksızın kümelemenin geçerliliğini sadece veri setinin özellikleri ve nicelikleri kullanılarak değerlendirirler. Oransal (relative) kriterlerde, farklı kümeleme algoritmaları ya da aynı kümeleme algoritmasından farklı parametre değerleri ile elde edilen kümeleme sonuçları karşılaştırılır.

3.3.1. Dışsal (External) Kriterler

Bir X veri matrisinde n gözlem vektörü bulunsun. Bu X veri matrisi hakkında bilgi sahibi olunan kümelenme P ve bu veri kümesine uygulanan bir kümeleme algoritması sonu elde edilen kümeleme C ile gösterilsin. X veri matrisinde x_i ve x_j gözlem çiftinin C ve P kümelenmesindeki yerlerine göre dört durum vardır.

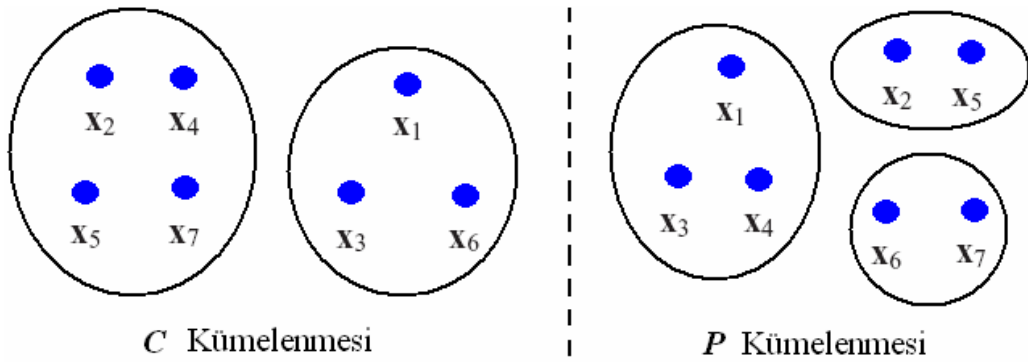
a : x_i ve x_j , C kümelenmesinde aynı küme içerisinde ve P kümelenmesinde de aynı küme içerisinde yer almasıdır.

b : x_i ve x_j , C kümelenmesinde aynı küme içerisinde ve P kümelenmesinde farklı küme içerisinde yer almasıdır.

c : x_i ve x_j , C kümelenmesinde farklı küme içerisinde ve P kümelenmesinde aynı küme içerisinde yer almasıdır.

d : x_i ve x_j , C kümelenmesinde farklı küme içerisinde ve P kümelenmesinde de farklı küme içerisinde yer almasıdır.

Muhtemel dört durumu göstermek için C ve P kümelenmesi Şekil 3.19'daki gibi olsun.



Şekil 3.19. C ve P kümelenmesi için gözlem çiftlerinin muhtemel dört durumu için örnek gösterim

Şekil 3.19'daki örnek C ve P kümelenmesi için hesaplanan a, b, c ve d değerleri Tablo 3.15'de verilmiştir.

Tablo 3.15. Şekil 3.19'daki örnek C ve P kümelenmesi için hesaplanan a, b, c ve d değerleri ve açıklamaları

Durum	Veri setindeki gözlem çiftleri	Toplam
a	(x_1, x_3) ve (x_2, x_5)	2
b	$(x_1, x_6); (x_2, x_4); (x_2, x_7); (x_3, x_6); (x_4, x_5); (x_4, x_7)$ ve (x_5, x_7)	7
c	$(x_1, x_4); (x_3, x_4)$ ve (x_6, x_7)	3
d	$(x_1, x_2); (x_1, x_5); (x_1, x_7); (x_2, x_3); (x_2, x_6);$ $(x_3, x_5); (x_3, x_7); (x_4, x_6)$ ve (x_5, x_6)	9

Her iki kümeden de aynı küme içersinde yer alan birimler (x_1, x_3) ve (x_2, x_5) olduğundan $a = 2$, her iki kümeden de farklı kümelerde yer alan gözlem çiftleri (x_1, x_2) ; (x_1, x_5) ; (x_1, x_7) ; (x_2, x_3) ; (x_2, x_6) ; (x_3, x_5) ; (x_3, x_7) ; (x_4, x_6) ve (x_5, x_6) olduğundan $d = 9$ olarak bulunmuştur. Muhtemel dört durum içersinden a ve d durumları iki kümeden benzerliği ile ilgili iken b ve c durumları iki kümeden farklılığı ile ilgilidir. **C** ve **P** gibi iki kümeden eşleştirilmesi ile küme geçerliliğini değerlendiren dışsal (external) kriterler arasından literatürde yaygın kullanılan indeksler aşağıda verilmiştir (Martinez ve Martinez 2005, Gan ve ark 2007, Xu ve Wunsch 2009, Servi 2009).

3.3.1.1. Rand İndeksi

Rand (1971), tarafından önerilen Rand indeksi

$$R = \frac{a + d}{a + b + c + d} \quad (3.15)$$

olarak tanımlanır ve $[0,1]$ aralığında değerler alır. İndeks değerleri bire yaklaştıkça **C** ve **P** kümedenmelerinin birbirine olan benzerliği artmaktadır.

3.3.1.2. Jaccard Katsayısı

Küme geçerliliğinin değerlendirilmesinde kullanılan Jaccard katsayısı (Jaccard 1908),

$$J = \frac{a}{a + b + c} \quad (3.16)$$

olarak tanımlanır ve $[0,1]$ aralığında değerler alır. Katsayı bire yaklaştıkça **C** ve **P** kümedenmelerinin birbirine olan benzerliği artmaktadır

3.3.1.3. Fowlkes ve Mallows İndeksi

Fowlkes ve Mallows (1983), tarafından küme geçerliliğini değerlendirilmesi için önerilen indeks,

$$FM = \sqrt{\frac{a}{a+b} \frac{a}{a+c}} \quad (3.17)$$

olarak tanımlanır. İndeks değerleri bire yaklaştıkça **C** ve **P** kümelenmelerinin birbirine olan benzerliği artmaktadır

3.3.1.4. Γ İstatistiği

Xu ve Wunsch (2009), tarafından önerilen Γ istatistiği,

$$\Gamma = \frac{Ma - m_1 m_2}{\sqrt{m_1 m_2 (M - m_1)(M - m_2)}} \quad (3.18)$$

olarak tanımlanır ve $[-1, 1]$ aralığında değerler alır. Burada $M = a + b + c + d$, $m_1 = a + b$ ve $m_2 = a + c$ olarak tanımlanır. İndeks değerleri 1'e yaklaştıkça **C** ve **P** kümelenmelerinin birbirine olan benzerliği, -1'e yaklaştıkça farklılığı artmaktadır.

3.3.1.5. Düzeltilmiş Rand İndeksi

Rand indeks değerinde küme sayısının artması durumunda kümeleme performansına bağlı olmadan Rand indeks değeri artmaktadır. Bu probleme karşın Hubert ve Arabie (1985), genelleştirilmiş hipergeometrik dağılım varsayımı altında düzeltilmiş Rand indeks değerini önermişlerdir. **X** veri matrisinin **C** ve **P** gibi elde edilen kümeler $\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, \dots, \mathbf{C}_c\}$ ve $\mathbf{P} = \{\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_p\}$ olmak üzere $\mathbf{C}_i$ ve $\mathbf{P}_j$

kümelerindeki ortak birimlerin sayısı $n_{ij} (n_{ij} = |\mathbf{C}_i \cap \mathbf{P}_j|)$ olsun. Bu tanımlamalar altında iki kümelenme sonucunun çapraz sınıflandırılması

	$\mathbf{P}_1$	$\mathbf{P}_2$	...	$\mathbf{P}_P$	Toplam
$\mathbf{C}_1$	n_{11}	n_{12}	...	n_{1P}	$n_{1.}$
$\mathbf{C}_2$	n_{21}	n_{22}	...	n_{2P}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\mathbf{C}_C$	n_{C1}	n_{C2}	...	n_{CP}	$n_{C.}$
Toplam	$n_{.1}$	$n_{.2}$	...	$n_{.P}$	n

şeklinde olur. Düzeltilmiş Rand indeks değeri,

$$R_D = \frac{\binom{n}{2} \sum_{i=1}^C \sum_{j=1}^P \binom{n_{ij}}{2} - \left[\sum_{i=1}^C \binom{n_{i.}}{2} \sum_{j=1}^P \binom{n_{.j}}{2} \right]}{\binom{n}{2} \frac{1}{2} \left[\sum_{i=1}^C \binom{n_{i.}}{2} + \sum_{j=1}^P \binom{n_{.j}}{2} \right] - \left[\sum_{i=1}^C \binom{n_{i.}}{2} \sum_{j=1}^P \binom{n_{.j}}{2} \right]} \quad (3.19)$$

eşitliği ile hesaplanır. İndeks değeri ne kadar büyükse iki parçalanmanın benzerliği o kadar fazladır.

3.3.2. İçsel (Internal) Kriterler

İçsel kriterler, veri setinin sahip olduğu özellikler ile verinin kümelenmesi ile elde edilen özellikleri göz önünde bulundurarak küme geçerliliğini değerlendiren ölçütlerdir (Gan ve ark 2007). İçsel kriterler tipik olarak veri setinin genel yapısı ile veri setinin parçalanması ile oluşan küme yapısı arasındaki uyumun bir göstergesidir (Miligan 1981). İçsel kriterler başlığı altında özellikle aşamalı kümeleme algoritmalarında küme geçerliliğinin değerlendirilmesinde yaygın kullanılan kofenetik korelasyon katsayısı ile Freidman ve Rubin (1967) tarafından küme geçerliliğinin değerlendirilmesi için önerilen Wilks'in Lambda test istatistiği incelenecektir.

3.3.2.1. Kofenetik (Cophenetic) Korelasyon Katsayısı

Kofenetik korelasyon katsayısı, başlangıçta hesaplanan benzerlik matrisi D ile ağaç grafiği oluşturulduktan sonra bağlantı uzaklıklarını içeren uzaklıklar matrisinin R , karşılıklı elemanları arasındaki ilişkiyi ölçen bir ilişki katsayısıdır ve CCC ile gösterilir (Sneath ve Sokal 1973, Anderberg 1973, Hawkins ve ark 1982). D matrisinde i . ve j . birimler arasındaki uzaklık d_{ij} , D matrisinde yer alan d_{ij} uzaklıklarının ortalaması $\bar{d}$ ve benzer şekilde r_{ij} , R matrisinde i . ve j . birimler arasındaki uzaklık ve $\bar{r}$, R matrisinde yer alan r_{ij} uzaklıklarının ortalaması olmak üzere kofenetik korelasyon katsayısı

$$CCC = \frac{\sum_{i < j} (d_{ij} - \bar{d})(r_{ij} - \bar{r})}{\sum_{i < j} (d_{ij} - \bar{d})^2 \sum_{i < j} (r_{ij} - \bar{r})^2} \quad (3.20)$$

şeklinde hesaplanır. Kofenetik korelasyon katsayısı $[-1,1]$ aralığında değer alır ve 1'e ne kadar yakın değer alırsa kümeleme de o kadar başarılıdır.

3.3.2.2. Wilks'in Lambda Test İstatistiği

Çok değişkenli varyans analizinde gruplar arasındaki farklılığının testinde kullanılan Wilks'in lambda test istatistiği, Friedman ve Rubin (1967) tarafından küme geçerliliğinin değerlendirilmesi için önerilmiştir. Wilks'inin lambda test istatistiği

$$\lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} \quad (3.21)$$

eşitliği ile tanımlanır. Eşitlikte yer alan $|\mathbf{W}|$, kümeler içi kareler ve çarpımlar toplamı matrisinin determinantını ve $|\mathbf{T}|$ ise genel kareler ve çarpımlar toplamı matrisinin determinantını göstermektedir. Wilks'in lambda test istatistiği aynı zamanda benzerlik oran kriteri olarak da adlandırılmaktadır (Dinçer, 1992). Wilks'in lambda test istatistiği yeterince küçük ise küme merkezlerine ait ortalama vektörlerinin eşit olduğunu öne süren yokluk (sıfır) hipotezi red edilir (Overall ve Klett 1972). Dolayısıyla λ değeri ne kadar küçük değer alırsa gerçekleştirilen kümeleme o kadar başarılıdır.

3.3.3. Oransal (Relative) Kriterler

Dışsal (external) ve içsel (internal) kriterler istatistiksel testlerin kullanımını gerektirdiğinden hesaplamalara oldukça duyarlıdır. Oransal (relative) kriterler de bu gereklilik yoktur ve farklı kümeleme algoritmalarının ya da aynı kümeleme algoritmasının farklı başlangıç parametreleri ile elde edilen karşılaştırılması için önerilmişlerdir.

Bir kümeleme algoritmasında verideki küme sayısı k 'nın bilinmesi gerekir. Küme sayısı k bilinmiyorsa bu değer için mümkün minimum değer olan $k_{\min}$ ve maksimum değer olan $k_{\max}$ arasında algoritma uygulanarak kümeleme yapıları elde edilir. Elde edilen bu farklı kümeleme yapıları belirlenen indekse göre değerlendirilir ve en iyi indeks değerine sahip kümeleme yapısı seçilir. Kümeleme analizi içerisinde, kümelerin geometrik ve istatistiksel yapılarını da dikkate alarak belirli bir standartta kümelerin elde edilmesinde bu şekilde birçok indeks önerilmiştir. Önerilen bu indekslerden önemli olan bazıları Calinski ve Harabasz indeksi, Davies-Bouldin indeksi ve Dunn indeksi bu başlık altında incelenecektir.

3.3.3.1. Calinski ve Harabasz İndeksi

Calinsky ve Harabasz (1974) yaptıkları çalışmada, $\mathbf{B}$ kümeler arası çarpımlar ve kareler matrisi, $\mathbf{W}$ kümeler içi çarpımlar ve kareler matrisi, k küme sayısı olmak üzere,

$$CH(k) = \frac{iz(\mathbf{B})(n-k)}{iz(\mathbf{W})(k-1)} \quad (3.22)$$

nolu eşitlikte verilen $CH(k)$ değerini en büyük yapan k değerini uygun küme sayısı olarak önermişlerdir. Küme sayısı sabit alınarak farklı iki algoritmadan elde edilen kümelemenin değerlendirilmesinde, $CH(k)$ değerini en büyük yapan algoritmadan elde edilen kümeleme daha geçerli kümeleme olarak kabul edilmektedir.

3.3.3.2. Davies-Bouldin İndeksi

Davies-Bouldin (1979), tarafından önerilen indeks, küme içindeki gözlemlerin küme merkezine olan uzaklıklarını minimum yapmayı ve kümeler arası uzaklıkları maksimum yapmayı amaçlar. $i = 1, 2, \dots, k$ ve $j = 1, 2, \dots, k$ olmak üzere i . ve diğer kümeler arasındaki maksimum karşılaştırma oranı her bir küme için R_i ile gösterilen küme indeksi,

$$R_i = \max_{i \neq j} \left(\frac{e_i + e_j}{d_{ij}} \right) \quad (3.23)$$

şeklinde hesaplanır. Burada d_{ij} i . ve j . küme merkezleri arasındaki uzaklığı, e_i ve e_j sırasıyla i . ve j . kümedeki gözlemlerin kendi küme merkezlerine olan ortalama uzaklıklarını gösterir. Davies-Bouldin indeksi,

$$DB(k) = \frac{1}{k} \sum_{i=1}^k R_i \quad (3.24)$$

şeklinde tanımlanır. Davies-Bouldin indeks değerini minimum yapan k uygun küme sayısı olarak belirlenebilir.

3.3.3.3. Dunn İndeksi

Dunn (1974), yüksek yoğunluklu iyi ayrımlı kümelerin belirlenmesinde oldukça etkin bir kriter önermiştir. İndeks değeri hesaplanırken, K_i ve K_j ile gösterilen iki küme arasındaki uzaklığı $D(K_i, K_j)$, sırasıyla K_i ve K_j kümelerinde yer alan $\mathbf{x}$ ve $\mathbf{y}$ gözlem çiftleri arasındaki minimum uzaklık olarak,

$$D(K_i, K_j) = \min_{\mathbf{x} \in K_i, \mathbf{y} \in K_j} d(\mathbf{x}, \mathbf{y}) \quad (3.25)$$

tanımlanır. K_i kümesinin çapı ise bu küme içerisinde yer alan gözlem çiftleri arasındaki maksimum uzaklık olarak,

$$\text{çap}(K_i) = \max_{\mathbf{x}, \mathbf{y} \in K_i} d(\mathbf{x}, \mathbf{y}) \quad (3.26)$$

şeklinde tanımlanır. Bu tanımlamalar altında Dunn indeksi,

$$DN(k) = \min_{i=1,2,\dots,k} \left(\min_{j=i+1,\dots,k} \left(\frac{D(K_i, K_j)}{\max_{l=1,2,\dots,k} \text{çap}(K_l)} \right) \right) \quad (3.27)$$

şeklinde ifade edilir. Dunn indeks değeri daha büyük olan kümeleme, geçerli kümeleme olarak kabul edilir (Xu ve Wunsch 2009).

4. UZAKLIK ÖLÇÜLERİ

Kümeleme analizi literatüründe, iki küme ya da iki gözlem arasındaki benzerlik(veya benzemezlik); uzaklık ölçüleri veya benzerlik ölçüleri ile tanımlanmıştır. Her kümeleme algoritması iki nokta arasındaki uzaklık veya benzerliği temel alır (Jain ve Dubes 1988). Eğer iki nokta arasındaki uzaklık veya benzerlik ölçülmemiş ise geçerli bir kümeleme yapmak mümkün değildir. Kümeleme analizinde özellikle veri setini oluşturan değişkenlerin sürekli olması durumunda benzerliğin belirlenmesinde uzaklık ölçüleri kullanılır.

Kümeleme analizi içerisindeki hangi teknik kullanılırsa kullanılsın amaç benzer birimleri bir araya getirerek küme içi değişimi minimum, kümeler arası değişimi ise maksimum yapmak olduğundan kümeleme performansı seçilen uzaklık ölçüsü ile doğrudan ilişkilidir. Kümeleme için en uygun uzaklık ölçüsünün seçimi deneyim, beceri, bilgi ve şanslı bir kombinasyonla mümkündür (Gan ve ark 2007). Bu bölümde kümeleme analizinde yaygın olarak kullanılan uzaklık ölçüleri incelenecektir.

4.1. Uzaklık ve Özellikleri

Çok değişkenli istatistiksel analizlerde genellikle n birimin p değişken bakımından gözlemlenmesi sonucu oluşan veri seti

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1(p-1)} & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2(p-1)} & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{i(p-1)} & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{n(p-1)} & x_{np} \end{bmatrix} \quad (4.1)$$

biçiminde $n \times p$ tipinde bir $\mathbf{X}$ matrisi ile ifade edilir. Burada x_{ij} , i . birim üzerinden gözlenen j . değişken değerini ifade eder. Bu tanımlama altında

$$\mathbf{x}_i = [x_{i1} \quad x_{i2} \quad \dots \quad x_{ij} \quad \dots \quad x_{i(p-1)} \quad x_{ip}] \quad (4.2)$$

olmak üzere $\mathbf{x}_i$, $1 \times p$ boyutlu gözlem vektörü olarak isimlendirilir.

$\mathbf{X}$ $n \times p$ boyutlu bir veri matrisi olmak üzere $d : \mathbf{X} \times \mathbf{X} \rightarrow \mathfrak{R}$ şeklinde tanımlanan bir fonksiyon eğer tüm $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ için

$$\text{i. } d(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \quad (\text{Pozitiflik}) \quad (4.3)$$

$$\text{ii. } d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i) \quad (\text{Simetri}) \quad (4.4)$$

$$\text{iii. } d(\mathbf{x}_i, \mathbf{x}_i) = 0 \quad (\text{Özdeşlik}) \quad (4.5)$$

koşullarını sağlıyorsa d , $\mathbf{X}$ üzerinde bir **uzaklık fonksiyonu** olarak adlandırılır (Deza ve Deza 2006). Yukarıda belirtilen pozitiflik, simetri ve özdeşlik koşulları yanında eğer tüm $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k \in \mathbf{X}$ için

$$\text{iv. } d(\mathbf{x}_i, \mathbf{x}_j) = 0 \text{ sadece ve sadece } \mathbf{x}_i = \mathbf{x}_j \quad (\text{Belirlilik}) \quad (4.6)$$

$$\text{iv. } d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_l) + d(\mathbf{x}_j, \mathbf{x}_l) \quad (\text{Üçgen Eşitsizliği}) \quad (4.7)$$

koşulları da sağlanıyorsa d uzaklık fonksiyonu **metrik** olarak isimlendirilir (Webb 2002). Kümeleme analizinde $i, j = 1, 2, \dots, n$ olmak üzere $(\mathbf{x}_i, \mathbf{x}_j)$ gözlem çiftleri arasındaki uzaklıklar d_{ij} olmak üzere n birimden oluşan veri setinden elde edilecek uzaklık matrisi $\mathbf{D}$,

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & d_{13} & \dots & d_{1n} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2n} \\ d_{31} & d_{32} & d_{33} & \dots & d_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & d_{n3} & \dots & d_{nn} \end{bmatrix} \quad (4.8)$$

şeklindedir. Uzaklık tanımındaki özdeşlik koşulu gereğince $i = j$ durumunda $d_{ii} = 0$ ve simetri koşulu gereğince $d_{ij} = d_{ji}$ olduğundan uzaklık matrisi $\mathbf{D}$

$$\mathbf{D} = \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1n} \\ d_{12} & 0 & d_{23} & \cdots & d_{2n} \\ d_{13} & d_{23} & 0 & \cdots & d_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{1n} & d_{2n} & d_{3n} & \cdots & 0 \end{bmatrix} \quad (4.9)$$

şeklinde gösterilir. Genel olarak kümeleme analizi içersinde yer alan aşamalı kümeleme teknikleri birimler arasındaki uzaklıkları gösteren $\mathbf{D}$ uzaklık matrisini temel alırken, aşamalı olmayan teknikler ise birimlerin seçilen küme merkezlerine olan uzaklığını temel alır. Kümeleme tekniği ne olursa olsun benzerliğin ölçümünde kullanılan uzaklık fonksiyonunun seçimi kümeleme performansını doğrudan etkileyen önemli bir faktördür (Jain ve ark 1999).

4.2. Nicel Değişkenler İçin Uzaklık Ölçüleri

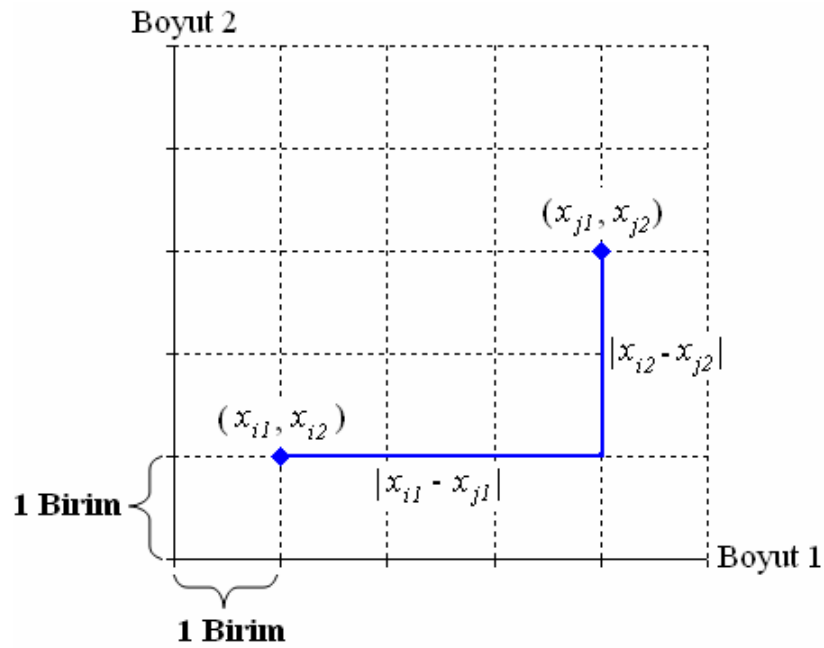
Nicel değişkenlerden oluşan veri setlerinde gözlem vektörleri arasındaki benzerlik uzaklık ölçüleri ile belirlendiğinde iki gözlem vektörü arasındaki minimum uzaklığa sahip olan gözlem çiftleri daha benzer olarak tanımlanmaktadır. Bu kısımda, nicel değişkenlerden oluşan veri setlerinde benzerliğin belirlenmesinde yaygın kullanılan uzaklık ölçüleri ve bu uzaklık ölçülerinin özellikleri incelenecektir.

4.2.1. Manhattan (City-Block) Uzaklığı

Manhattan uzaklığı birimlerin, değişkenleri arasındaki mutlak farkların toplamına eşittir. Manhattan uzaklığı City-Block uzaklığı olarak da bilinir. Bu ölçüt kullanıldığında farkın karesi alınmadığı için farklılıkların etkisi azalır. İki gözlem vektörü arasındaki Manhattan uzaklığı $i, j = 1, 2, \dots, n$ ve $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ olmak üzere

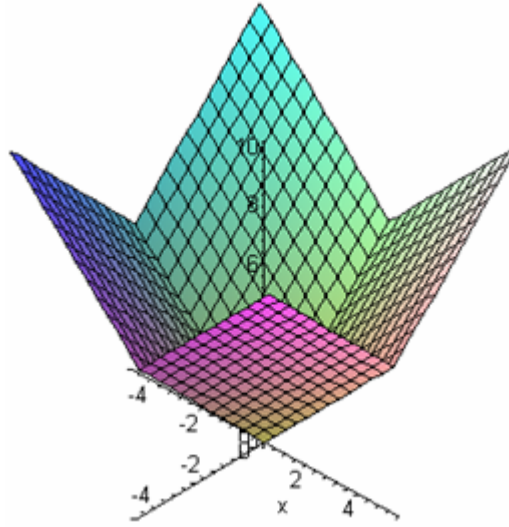
$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^p |x_{ik} - x_{jk}| \quad (4.10)$$

eşitliği ile verilir. Değişken sayısının iki olması durumunda $p = 2$ için $\mathbf{x}_i$ ve $\mathbf{x}_j$ gözlem vektörleri arasındaki Manhattan uzaklığının hesaplanması Şekil 4.1’de grafiksel olarak verilmiştir.



Şekil 4.1. Manhattan uzaklığının iki boyutlu düzlemde hesaplanması

Şekil 4.1’de iki boyutlu düzlemde gösterilen $\mathbf{x}_i$ ve $\mathbf{x}_j$ gözlemleri arasındaki Manhattan uzaklığı $3+2=5$ birimdir. Manhattan uzaklığı aynı zamanda metriktir ve L_1 metriği olarak da isimlendirilir. Değişken sayısının iki ve değişkenlerin alacağı değerler $(-\infty, \infty)$ aralığında olmak üzere (x_{i1}, x_{i2}) noktasının orijin noktasına olan Manhattan uzaklık fonksiyonunun grafiği Şekil 4.2’de verilmiştir.



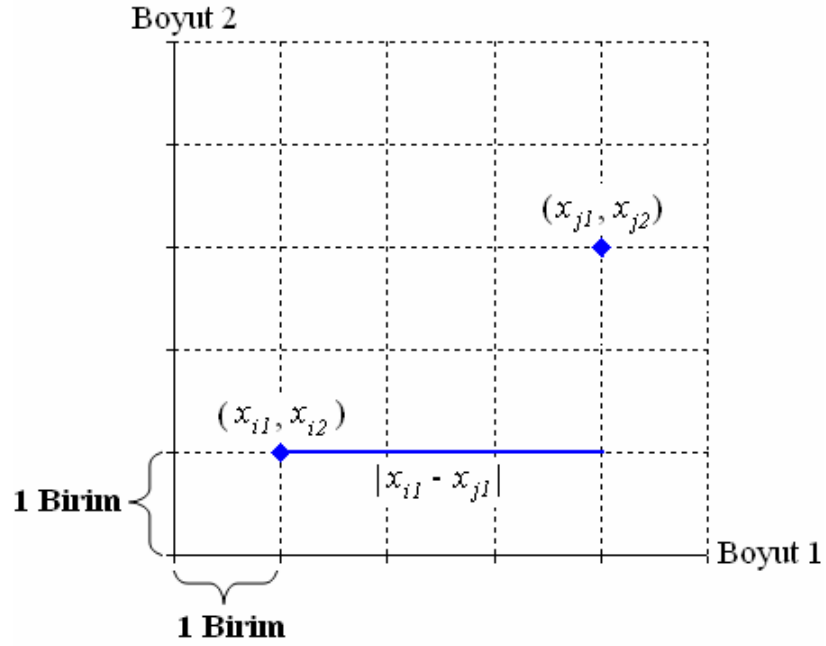
Şekil 4.2. Manhattan uzaklık fonksiyonunun üç boyutlu uzayda grafiği

4.2.2. Chebyshev Uzaklığı

Chebyshev uzaklığı iki gözlem vektörü arasındaki maksimum farklılığa sahip boyuttaki farkı, iki gözlem vektörü arasındaki uzaklık olarak alan metrik bir uzaklık ölçüsüdür. Chebyshev uzaklığı, maksimum metrik veya L_∞ metriği olarak da isimlendirilir. İki gözlem vektörü arasındaki Chebyshev uzaklığı $i, j = 1, 2, \dots, n$ ve $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ olmak üzere

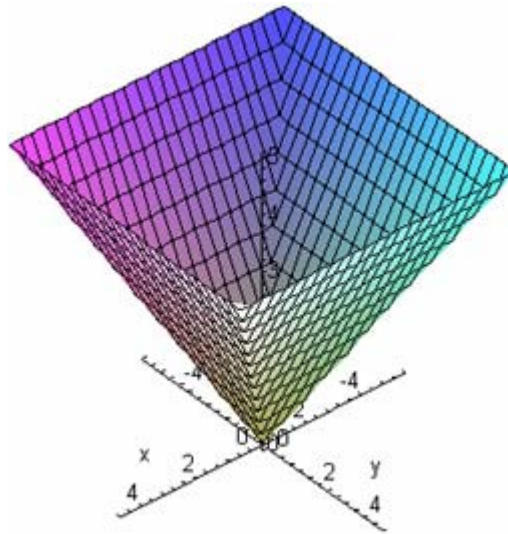
$$d(\mathbf{x}_i, \mathbf{x}_j) = \max_{k=1, \dots, p} \{ |x_{ik} - x_{jk}| \} \quad (4.11)$$

eşitliği ile verilir. Veri setinde yer alan değişkenlerin ölçü birimleri farklı ise başka bir ifadeyle terim büyüklükleri farklı ise Chebyshev uzaklığında terim büyüklüğü fazla olan değişken diğer değişkenlere baskın gelecektir. Terim büyüklüklerinin farklı olduğu veri setlerinde, benzerliğin hesaplanmasında terim büyüklüğü fazla olan boyut dışındaki diğer boyutlardaki farklı özellikler göz önüne alınamayacağından Chebyshev uzaklığının kullanılması önerilmez. Değişken sayısının iki olması durumunda ($p = 2$ için) $\mathbf{x}_i$ ve $\mathbf{x}_j$ gözlem vektörleri arasındaki Chebyshev uzaklığının hesaplanması Şekil 4.3’de grafiksel olarak verilmiştir.



Şekil 4.3. Chebsyhev uzaklığının iki boyutlu düzlemde hesaplanması

Şekil 4.3’de iki boyutlu düzlemde gösterilen $\mathbf{x}_i$ ve $\mathbf{x}_j$ gözlemleri arasındaki Chebsyhev uzaklığı 3 birim olarak elde edilmiştir. Değişken sayısının iki olması durumunda (x_{i1}, x_{i2}) noktasının orijin noktasına olan Chebsyhev uzaklık fonksiyonunun grafiği Şekil 4.4’de verilmiştir.



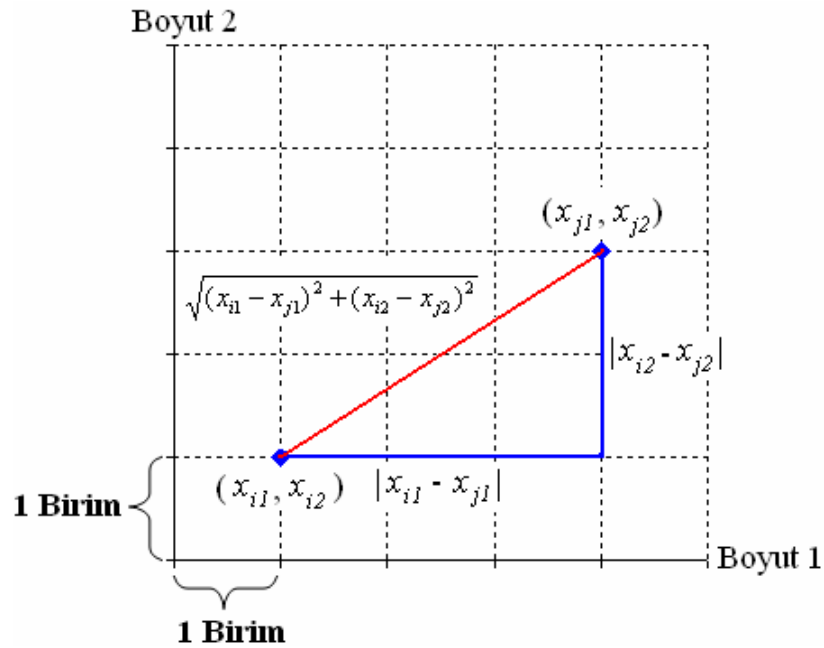
Şekil 4.4. Chebsyhev uzaklık fonksiyonunun üç boyutlu uzayda gösterimi

4.2.3. Öklid (Euclidean) Uzaklığı

Kümeleme analizinde muhtemelen en çok bilinen ve en yaygın kullanılan uzaklık ölçüsü olan Öklid uzaklığı, gözlem vektörleri arasındaki farkların kareleri toplamının karekökünün alınması ile hesaplanır. Metrik bir uzaklık ölçüsü olan Öklid uzaklığı L_2 metriği olarak da isimlendirilir. İki gözlem vektörü arasındaki Öklid uzaklığı $i, j = 1, 2, \dots, n$ ve $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ olmak üzere

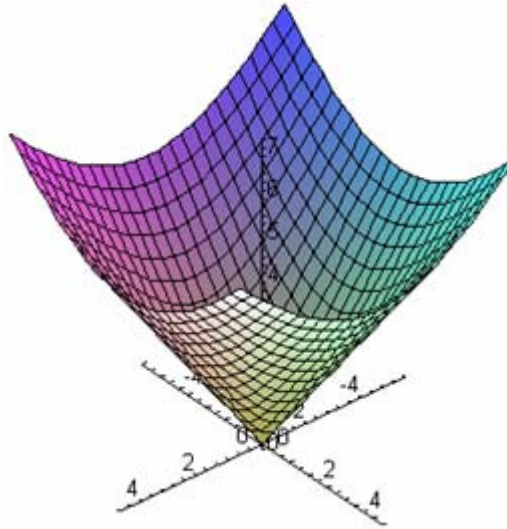
$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^p (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}} = \left[(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)' \right]^{\frac{1}{2}} \quad (4.12)$$

eşitliği ile verilir. Değişken sayısının iki olması durumunda ($p = 2$ için) Öklid uzaklığı Pisagor teoremine göre dik üçgende hipotenüs olarak isimlendirilen kenarın uzunluğuna karşılık gelir. Gözlem vektörleri $\mathbf{x}_i$ ve $\mathbf{x}_j$ arasındaki Öklid uzaklığının hesaplanması Şekil 4.5’de grafiksel olarak verilmiştir.



Şekil 4.5. Öklid uzaklığının iki boyutlu düzlemde hesaplanması

Şekil 4.5’de iki boyutlu düzlemde gösterilen $\mathbf{x}_i$ ve $\mathbf{x}_j$ gözlemleri arasındaki Öklid uzaklığı $\sqrt{4^2 + 2^2} = \sqrt{20}$ birim olarak elde edilmiştir. Yapılan çalışmalarda Öklid uzaklığının hiperküresel kümeler oluşturma eğiliminde olduğu görülmüştür (Xu ve Wunsch 2009). Veri setini oluşturan değişkenlerin ölçü birimlerinin farklı olması durumunda Öklid uzaklığı hesaplanırken terim büyüklüğü ve varyansı büyük olan değişkenin katkısı diğer değişkenlerden fazladır. Değişken sayısının iki olması durumunda (x_{i1}, x_{i2}) noktasının orijin noktasına olan Öklid uzaklık fonksiyonunun grafiği Şekil 4.6’da verilmiştir.



Şekil 4.6. Öklid uzaklık fonksiyonunun üç boyutlu uzayda gösterimi

4.2.4. Standartlaştırılmış Öklid Uzaklığı

Standartlaştırılmış Öklid uzaklığı Karl Pearson uzaklık ölçütü olarak da adlandırılır. Öklid uzaklığında her değişkenin varyansının ağırlık olarak kullanılmasıyla elde edilir. İki gözlem vektörü arasındaki standartlaştırılmış Öklid uzaklığı $i, j = 1, 2, \dots, n$ ve $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ olmak üzere

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^p \frac{1}{s_k^2} (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}} \quad (4.13)$$

eşitliği ile tanımlanır. Eşitlikte yer alan s_k^2 , k . değişkene ait örneklem varyansını göstermektedir ve $\bar{x}_{.k}$, k . değişkene ait aritmetik ortalama olmak üzere

$$\bar{x}_{.k} = \frac{\sum_{i=1}^n x_{ik}}{n}, \quad k = 1, \dots, p \quad (4.14)$$

$$s_k^2 = \frac{\sum_{i=1}^n (x_{ik} - \bar{x}_{.k})^2}{n-1}, \quad k = 1, \dots, p \quad (4.15)$$

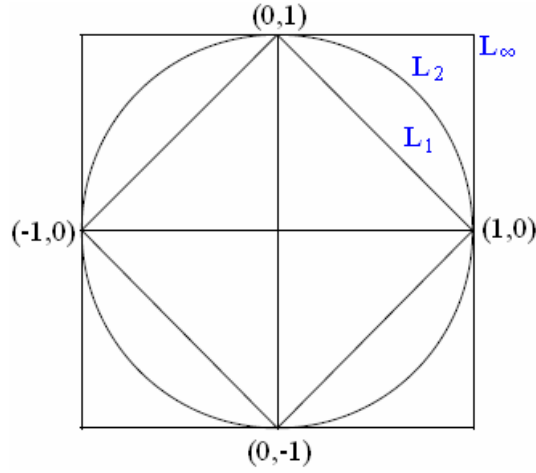
şeklinde hesaplanır. Standartlaştırılmış Öklid uzaklığı değişkenlerin ölçülmesinde kullanılan ölçü biriminden ve terim büyüklüklerinden etkilenmez.

4.2.5. Minkowski Uzaklığı

Kümeleme analizinde nicel veriler için kullanılan uzaklık ölçütlerinden biri de Minkowski uzaklığıdır. İki gözlem vektörü arasındaki Minkowski uzaklığı, $\lambda \geq 1$ ve $i, j = 1, 2, \dots, n$ için $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ olmak üzere

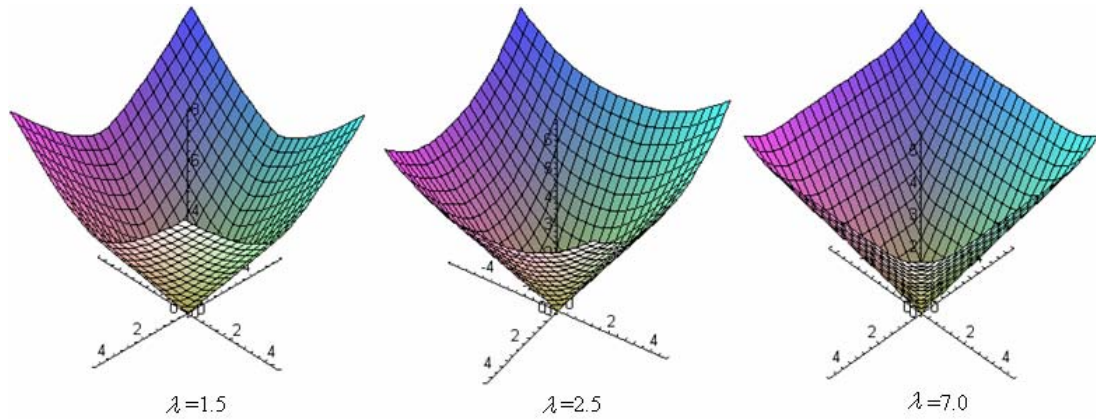
$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^p |x_{ik} - x_{jk}|^\lambda \right)^{1/\lambda} \quad (4.16)$$

eşitliği ile hesaplanır. Aynı zamanda Minkowski uzaklığı bir metrik uzaklık ölçüsüdür ve L_λ metriği olarak da isimlendirilir. Minkowski uzaklığı genel bir uzaklık ölçüsüdür. Minkowski uzaklığında $\lambda = 1$ alındığında Manhattan uzaklığı, $\lambda = 2$ alındığında Öklid uzaklığı ve $\lambda \rightarrow \infty$ için Chebyshev uzaklığı elde edilir. L_1, L_2 ve L_∞ metrikleri arasındaki geçişler birim çember üzerinde Şekil 4.7’de verilmiştir.



Şekil 4.7. İki boyutlu düzlemde orijinden bir birim uzaklıktaki noktalardan oluşan yüzeyde L_1 , L_2 ve L_∞ metrikleri

Veri seti içerisindeki kompakt ve birbirlerinden iyi ayrılmış kümeleri belirlemede Minkowski uzaklığı iyi sonuçlar vermektedir (Mao ve Jain 1996). Aksi takdirde, değişimi fazla olan değişkenin etkisi Minkowski uzaklığının belirlenmesinde baskın hale gelmektedir. Bu durumdan kaçınmak için standartlaştırma veya ağırlıklandırma kullanılabilir (Jain ve ark 1999). Farklı λ değerleri, değişken sayısının iki ve değişkenlerin alacağı değerler $(-\infty, \infty)$ aralığında olmak üzere (x_{i1}, x_{i2}) noktasının orijin noktasına olan Minkowski uzaklık fonksiyonunun grafikleri Şekil 4.8’de verilmiştir.



Şekil 4.8. Farklı λ değerleri Minkowski uzaklık fonksiyonunun üç boyutlu uzayda gösterimi

4.2.6. Mahalanobis Uzaklığı

Mahalanobis uzaklığı değişkenlerin varyanslarına ve karşılıklı (ikili) korelasyonlarına göre farklı özniteliklere farklı ağırlıklar verir. Diğer bir deyişle bu ölçü küme elemanlarının çok değişkenli Gauss dağılımdan geldiğini varsayar. $\mathbf{S}$ örneklem kovaryans matrisi ve $i, j = 1, 2, \dots, n$ için $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ olmak üzere Mahalanobis uzaklığı

$$d(\mathbf{x}_i, \mathbf{x}_j) = [(\mathbf{x}_i - \mathbf{x}_j)\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j)']^{\frac{1}{2}} \quad (4.17)$$

eşitliği ile tanımlanır. Eşitlik de yer alan $\mathbf{S}$ örneklem kovaryans matrisi $\mathbf{X}$ veri matrisi için

$$\mathbf{S} = \text{cov}(\mathbf{X}) = \frac{1}{n-1} \mathbf{X}'(\mathbf{I} - \frac{1}{n} \mathbf{J})\mathbf{X} \quad (4.18)$$

eşitliği ile hesaplanır. Kovaryans matrisi

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \quad (4.19)$$

şeklindedir. $\mathbf{S}$ matrisinin esas köşegen üzerindeki terimleri örneklem varyansını gösterirken ($s_1^2 = s_{11}$), diğer terimler değişkenler arasındaki kovaryansı gösterdiğinden $\mathbf{S}$ matrisi varyans-kovaryans matrisi olarak da isimlendirilmektedir. $\mathbf{S}$ örneklem kovaryans matrisinde $i, j = 1, 2, \dots, p$ olmak üzere $s_{ij} = s_{ji}$ olduğundan örneklem kovaryans matrisi simetrik bir matristir. $\mathbf{X}$ veri matrisindeki değişkenler ilişkisiz ise yani $i, j = 1, 2, \dots, p$ ve $i \neq j$ olmak üzere $s_{ij} = 0$ ise Mahalanobis

uzaklığı standartlaştırılmış Öklid uzaklığına dönüşür. $\mathbf{S} = \mathbf{I}$ olduğunda yani değişkenler arasında ilişkinin olmadığı ve değişkenlere ait varyansların birim varyansa eşit olması durumunda Mahalanobis uzaklığı karesel Öklid uzaklığına dönüşür. İki gözlem vektörü arasındaki karesel Öklid uzaklığı

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^p (x_{ik} - x_{jk})^2 = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)' \quad (4.20)$$

eşitliği ile elde edilir. Mahalanobis uzaklığı ölçü birimi ve terim büyüklüklerinden etkilenmez ve lineer dönüşümler altında değişmezdir (Duda ve ark. 2001). Mahalanobis uzaklığı hiper-elips yapıda kümeler oluşturma eğilimindedir. Değişken sayısının artması durumunda Mahalanobis uzaklığının hesaplanmasında kullanılan $\mathbf{S}$ kovaryans matrisinin tersinin bulunmasının güçleşmesi uzaklığın önemli bir dezavantajıdır (Xu ve Wunsch 2009).

Morrison (1967), Mahalanobis uzaklığında değişkenlerin ağırlıklandırarak genelleştirilmiş Mahalanobis uzaklığını önermiştir. Genelleştirilmiş Mahalanobis uzaklığı w_k , k . değişkene ait ağırlık ve $\mathbf{W} = \text{diagonal}(w_k)$, $k = 1, \dots, p$ olmak üzere

$$d(\mathbf{x}_i, \mathbf{x}_j) = [(\mathbf{x}_i - \mathbf{x}_j)\mathbf{W}\mathbf{S}^{-1}\mathbf{W}(\mathbf{x}_i - \mathbf{x}_j)']^{\frac{1}{2}} \quad (4.21)$$

şeklinde tanımlanır.

4.2.7. Nicel Veriler İçin Diğer Uzaklıklar

Öklid uzaklığının değişkenlerin ölçü biriminin farklı olması durumundan etkilenmesine karşın Orloci (1967), bir yarıçaplı hiper küre içindeki normalleştirilmiş iki noktayı birleştiren kiriş uzunluğunu önermiştir. İki gözlem vektörü arasındaki Chord uzaklığı $i, j = 1, 2, \dots, n$ için $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ olmak üzere

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left[2 - 2 \frac{\sum_{k=1}^p x_{ik} x_{jk}}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2} \right]^{\frac{1}{2}} \quad (4.22)$$

eşitliği ile hesaplanır. Eşitlikte yer alan $\|\cdot\|_2$ gösterimi $\mathbf{x}_i$ gözlem vektörünün orijine olan Öklid uzaklığıdır ve

$$\|\mathbf{x}_i\|_2 = \sqrt{\sum_{k=1}^p x_{ik}^2} \quad (4.23)$$

eşitliği ile elde edilir. Chord uzaklığı normalleştirilmemiş veriler içinde (4.22) nolu eşitlik kullanılarak doğrudan hesaplanabilir.

Whittaker (1952) iki gözlem arasındaki benzerliğin hesaplanması için birleşme indeksi adı verdiği bir uzaklık önermiştir. $i, j = 1, 2, \dots, n$ ve $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ olmak üzere iki gözlem vektörü arasındaki birleşme indeksi

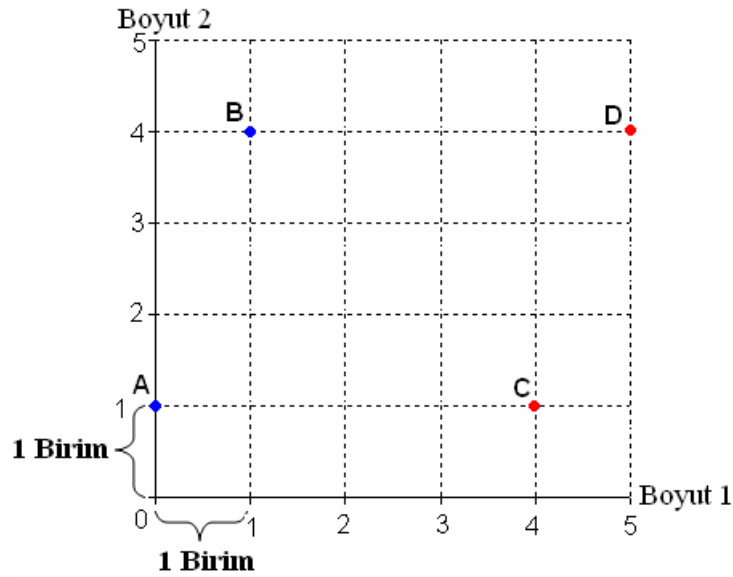
$$d(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{2} \sum_{k=1}^p \left| \frac{x_{ik}}{\sum_{l=1}^p x_{il}} - \frac{x_{jk}}{\sum_{l=1}^p x_{jl}} \right| \quad (4.24)$$

eşitliği ile elde edilir.

Legendere ve Legendere (1983) iki gözlem vektörü arasındaki benzerliğin belirlenmesinde Canberra uzaklığını önermişlerdir. İki gözlem vektörü arasındaki Canberra uzaklığı $i, j = 1, 2, \dots, n$ ve $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ olmak üzere

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^p \frac{|x_{ik} - x_{jk}|}{(x_{ik} + x_{jk})} \quad (4.25)$$

şeklinde hesaplanır. Canberra uzaklığı $\mathbf{X}$ veri matrisindeki tüm elemanların pozitif olması durumunda uygun bir uzaklık ölçüsüdür. Canberra uzaklığı gözlemlerin konumlarına ve sıfıra yakın değerlere sahip değişkendeki farklılıklara karşı oldukça duyarlıdır (Ziviani ve ark 2004). Canberra uzaklığının gözlemlerin konumlarına olan duyarlılığı Şekil 4.9’da verilmiştir.



Şekil 4.9. Gözlemlerin konumlarının Canberra uzaklığındaki etkisine ait gösterimi

Şekil 4.9’da verilen A ve B noktaları arasındaki Öklid uzaklığı ile C ve D noktaları arasındaki Öklid uzaklığı eşit iken ($d(A, B) = d(C, D) = \sqrt{10}$), A ve B noktaları arasındaki Canberra uzaklığı ($d(A, B) = 1.6$) ile C ve D noktaları arasındaki Canberra uzaklığı ($d(C, D) = 0.711$) eşit değildir.

$\mathbf{X}$ veri matrisinde $i = 1, \dots, n$ ve $j = 1, \dots, p$ olmak üzere $x_{ij} \geq 0$ ise iki gözlem vektörü arasındaki benzerliğin ölçümünde kullanılabilecek bir diğer uzaklık ölçüsü Hellinger uzaklığıdır. $i, j = 1, 2, \dots, n$ ve $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ olmak üzere iki gözlem vektörü arasındaki Hellinger uzaklığı

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left[\sum_{k=1}^p \left(\sqrt{x_{ik}} + \sqrt{x_{jk}} \right)^2 \right]^{\frac{1}{2}} \quad (4.26)$$

eşitliği ile elde edilir (Abdi 2007). Hellinger uzaklığı değişkenliğe karşı daha az duyarlı bir uzaklık ölçüsüdür.

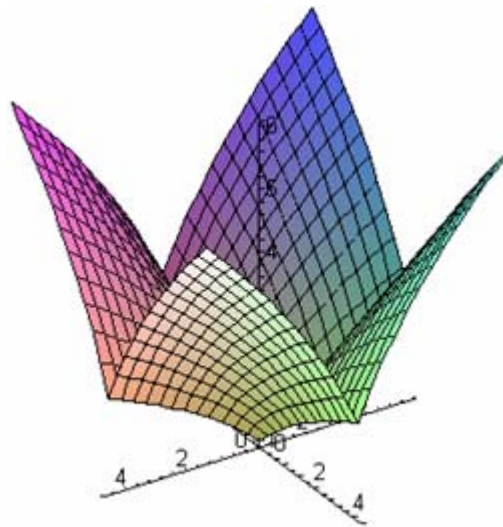
4.3. Önerilen Uzaklık Ölçüleri

4.3.1. Geometrik Uzaklık

$\mathbf{X}$ veri matrisindeki aykırı değerlerin gözlem vektörleri arasındaki uzaklığa olan etkisini azaltmak için temeli geometrik ortalamaya dayanan uzaklık ölçüsü tanımlanabilir. $i, j = 1, 2, \dots, n$ ve $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ olmak üzere iki gözlem vektörü arasındaki geometrik uzaklık

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\prod_{k=1}^p (|x_{ik} - x_{jk}| + 1) \right)^{1/p} - 1 \quad (4.27)$$

eşitliği ile elde edilir. Geometrik uzaklık ölçüsünde pozitiflik, simetri ve özdeşlik koşulları sağlanmaktadır. Değişken sayısının iki olması durumunda (x_{i1}, x_{i2}) noktasının orijin noktasına olan geometrik uzaklık fonksiyonunun grafiği Şekil 4.10'da verilmiştir.



Şekil 4.10. Geometrik uzaklık fonksiyonunun üç boyutlu uzayda grafiği

4.3.2. Karma Uzaklık

Öklid uzaklığında her değişkenin ortalamasının ağırlık olarak kullanılmasıyla elde edilen ağırlıklı Öklid uzaklığı

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^p \frac{(x_{ik} - x_{jk})^2}{\bar{x}_{.k}^2} \right)^{1/2} \quad (4.28)$$

eşitliği ile tanımlanır. Eşitlikte yer alan $\bar{x}_{.k}$ $k=1, \dots, p$ olmak üzere $\mathbf{X}$ veri matrisindeki k . değişkene (sütuna) ait aritmetik ortalamayı göstermektedir.

$$\bar{x}_{.k} = \frac{\sum_{i=1}^n x_{ik}}{n}, \quad k = 1, \dots, p \quad (4.29)$$

İlişki türü benzerlik ölçüsü olan Pearson korelasyon katsayısı benzerliğin pozitif olması koşulundan dolayı kümeleme analizinde istatistiki uygulamalardan farklı olarak mutlak değerce ele alınır. İki gözlem vektörü arasındaki mutlak Pearson korelasyon katsayısı

$$|\rho(\mathbf{x}_i, \mathbf{x}_j)| = \frac{\left| \sum_{k=1}^p (x_{ik} - \bar{x}_{i.})(x_{jk} - \bar{x}_{j.}) \right|}{\left[\sum_{k=1}^p (x_{ik} - \bar{x}_{i.})^2 \right]^{1/2} \left[\sum_{k=1}^p (x_{jk} - \bar{x}_{j.})^2 \right]^{1/2}} \quad (4.30)$$

eşitliği ile hesaplanır ve $[0,1]$ aralığında değer alır. Eşitlikte yer alan $\bar{x}_{i.}$, $i = 1, \dots, n$ olmak üzere i . gözleme ait ortalama olup

$$\bar{x}_i = \frac{\sum_{k=1}^p x_{ik}}{p} \quad (4.31)$$

eşitliği ile hesaplanır. İlişki türü benzerlik ölçüsü için s gösterimi, uzaklık türü benzerlik ölçüsü için d gösterimi kullanıldığında aralarındaki bağıntı

$$d = 1 - s \quad (4.32)$$

eşitliği ile verilir. İlişki türü benzerlik ölçüsü olan mutlak Pearson korelasyon katsayısı, 1'den çıkartılarak uzaklık türü ölçüye dönüştürülmüş olur. Elde edilen $(1 - |\rho(\mathbf{x}_i, \mathbf{x}_j)|)$ uzaklık ölçüsü ile eşitlik (4.28)'de verilen ağırlıklı Öklid uzaklığı bir arada kullanılarak karma bir uzaklık tanımlanabilir. Kümeleme analizinde $\mathbf{x}_i$ ve $\mathbf{x}_j$ gözlem vektörleri arasındaki benzerliği belirlenmesinde iki farklı uzaklığın karması ile elde edilen bu yeni uzaklık

$$d(\mathbf{x}_i, \mathbf{x}_j) = w \left(\sum_{k=1}^p \frac{(x_{ik} - x_{jk})^2}{\bar{x}_{.k}^2} \right)^{1/2} + (1 - w) (1 - |\rho(\mathbf{x}_i, \mathbf{x}_j)|) \quad (4.33)$$

eşitliği ile tanımlanır. Eşitlikte yer alan w ise ölçüt ağırlığıdır ve $0 \leq w \leq 1$ aralığındadır. Ölçüt ağırlığı w 'nin değeri 1'e yaklaştıkça uzaklık fonksiyonunda ağırlıklandırılmış Öklid uzaklığının etkisi artar.

Karma uzaklık ölçüsünde pozitiflik, simetri ve özdeşlik koşulları sağlanmaktadır. Karma uzaklıkta kullanılan her iki ölçütte ölçü birimindeki farklılıklardan etkilenmediği için yeni tanımlanan uzaklık ölçüsü de değişkenlerin ölçümünde kullanılan ölçü birimi farklılıklarından etkilenmez.

5. BOYUT İNDİRGEMENİN KÜMELEMeye ETKİSİ

Kümeleme analizi ile ilgili literatür incelendiğinde, yüksek boyutlu veri (High-Dimensional Data) matrisi için bazen kümeleme analizi öncesi boyut indirmek amacıyla temel bileşenler analizinin kullanıldığı görülmektedir (Yeung ve Ruzzo 2001). Bu çalışmaların genelinde orijinal değişkenler yerine temel bileşenler analizinden elde edilen ilk birkaç temel bileşenle kümeleme analizi uygulanmaktadır.

Boyut indirgeme işlemi, değişken dönüştürme veya değişken seçimi ile iki farklı şekilde gerçekleştirilebilmektedir. Değişken dönüştürme teknikleri içerisinde temel bileşenler analizi kullanımı en yaygın olan tekniktir (Bouveyron ve ark 2007). Orijinal değişkenlerin birleşimlerinden yararlanarak elde edilen daha az sayıda bileşenlerin veri matrisi içerisindeki kümeleri ortaya çıkarmada oldukça başarılı olduğu ifade edilmektedir (Lance ve ark 2004). Jolliffe ve arkadaşları 1980 yılında yaptıkları çalışmada yüksek boyutlu verilerde kümeleme analizi öncesi uygulanan temel bileşenler analizinden elde edilen ilk birkaç temel bileşen ile kümeleme analizi uygulanması ile veri seti içerisindeki gizli kümelerin başarılı bir şekilde ortaya çıkarıldığını göstermişlerdir. Ayrıca küme sayısının 2 olması durumunda kural olmamakla birlikte ilk bileşenin kümelemede iyi sonuçlar verdiğini ifade etmişlerdir.

Diğer taraftan Chang (1983), küme sayısının 2 olduğu farklı ortalama ve kovaryans matrislerine sahip veri setlerinde ilk birkaç temel bileşenin küme yapısını ortaya çıkarmada her zaman etkili olamayacağını yaptığı simülasyon çalışması ile göstermiştir. Bu çalışmada sadece küme sayısının 2 olduğu durumlar için verinin iki boyutlu en iyi gösteriminin ilk ve son temel bileşen kullanılarak elde edileceğini belirtmiştir.

Yeung ve Ruzzo (2001), kümelemede orijinal değişkenler yerine ilk birkaç temel bileşeni kullanmanın kümeleme kalitesini bozduğunu ortaya koymuşlardır. Yaptıkları çalışmada farklı kümeleme algoritmaları ve farklı benzerlik matrislerinin kullanımında bu yaklaşımının farklı etkilere sahip olduğunu göstermişlerdir.

Bu bölümde yüksek boyutlu veri matrislerinde kümeleme analizi öncesi uygulanacak temel bileşenler analizinden elde edilen bileşenler içerisinde tercih

edilen bileşen sayısının kümelemeye etkisi incelenecektir. Orijinal değişkenler yerine temel bileşenlerin kümelemeye etkisi incelenirken küme sayısının farklı olması ve kümelerin kovaryans matrislerinin eşit olup olmaması göz önünde bulundurulacaktır. Ayrıca kümeleme analizi öncesi verinin standartlaştırılarak kümeleme de ham veri yerine standartlaştırılmış verinin kullanılmasının kümelemeye etkisi incelenecektir.

5.1. Veri Setlerinin Üretimi

$X_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), X_2 \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \dots, X_k \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ olmak üzere, her birinden örneklem hacmi sırasıyla $n_1, n_2, \dots, n_k$ olan gözlemler üretilip,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_k \end{bmatrix}_{n \times p}, \quad n = \sum_{i=1}^k n_i \quad (5.1)$$

şeklinde yazılırsa, içerisinde k küme barındıran veri matrisi oluşturulur. Ortalama vektörü $\boldsymbol{\mu}_i$ ve kovaryans matrisi $\boldsymbol{\Sigma}_i$ olan çok değişkenli normal dağılıma sahip p değişkenli X_i rassal değişkeninin olasılık yoğunluk fonksiyonu,

$$f(\underline{x}_i) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2} (\underline{x}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\underline{x}_i - \boldsymbol{\mu}_i)\right) \quad (5.2)$$

şeklindedir. Çok değişkenli normal dağılımdan pozitif tanımlı $\boldsymbol{\Sigma}$ matrisi kullanılarak Cholesky ayrışımı veya spektral ayrışım ile iki farklı yoldan rassal sayılar üretilmektedir.

$\boldsymbol{\Sigma}$ kovaryans matrisini, Cholesky ayrıştırması ile

$$\boldsymbol{\Sigma} = \mathbf{A}'\mathbf{A} \quad (5.3)$$

şeklinde yazmak mümkündür. Eşitlikte yer alan $\mathbf{A}$ matrisi üst üçgen bir matristir. $\mathbf{Z}' = [\mathbf{Z}'_1 \ \mathbf{Z}'_2 \ \dots \ \mathbf{Z}'_p]$ matrisi her bir elemanı standart normal dağılıma sahip ve birbirinden bağımsız rassal değişkenler olmak üzere, çok değişkenli normal dağılıma sahip $\mathbf{X}$ rassal değişkeni,

$$\mathbf{X} = \mathbf{A}'\mathbf{Z} + \boldsymbol{\mu} \quad (5.4)$$

dönüşümüyle elde edilir. $\boldsymbol{\Sigma}$ kovaryans matrisini, spektral ayrışım ile

$$\boldsymbol{\Sigma}^{-\frac{1}{2}} = \mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}' \quad (5.5)$$

şeklinde yazmak mümkündür. Eşitlikte yer alan $\mathbf{V}$, kovaryans matrisinin özvektörlerinden oluşan bir matris ve $\mathbf{D}$ esas köşegen elemanları kovaryans matrisinin özdeğerlerinden oluşan diyagonal bir matristir. $\mathbf{Z}' = [\mathbf{Z}'_1 \ \mathbf{Z}'_2 \ \dots \ \mathbf{Z}'_p]$ matrisi her bir elemanı standart normal dağılıma sahip ve birbirinden bağımsız rassal değişkenler olmak üzere, çok değişkenli normal dağılıma sahip $\mathbf{X}$ rassal değişkeni,

$$\mathbf{X} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{Z} + \boldsymbol{\mu} \quad (5.6)$$

dönüşümüyle elde edilir. Görüldüğü gibi çok değişkenli normal dağılımdan rassal sayı üretmek için öncelikle standart normal dağılımdan rassal sayılar üretmek gerekmektedir. Standart normal dağılımdan rassal sayı üretmek için, R_1 ve R_2 bağımsız Uniform değişkenler olmak üzere,

$$z = \sqrt{-2 \ln(R_1)} \sin(2\pi R_2) \quad (5.7)$$

formülü ile standart normal dağılımdan rassal sayılar üretilir. $\mathbf{X}$ $n \times p$ boyutlu veri matrisi, $\boldsymbol{\mu}$ ortalama vektörü ve $\boldsymbol{\Sigma}$ kovaryans matrisi olmak üzere, $\mathbf{X}$ veri matrisi,

$$\mathbf{Z}_{n \times p} = (\mathbf{X}_{n \times p} - \mathbf{1}_{n \times 1} \boldsymbol{\mu}_{1 \times p}) \left(\sqrt{\text{diagonal}(\boldsymbol{\Sigma})} \right)_{p \times p}^{-1} \quad (5.8)$$

formülü ile standartlaştırılır.

5.2. Temel Bileşenler Analizi

Temel bileşenler analizi, orijinal değişkenlerin özelliklerini tutarak veri setini daha basit bir şekilde açıklayabilmek için orijinal değişkenler arasındaki maksimum varyansı veren lineer bileşimi arayan çok değişkenli istatistiksel bir tekniktir (Sampath ve Henrik 2006). $\mathbf{X}$, $n \times p$ boyutlu veri matrisi ve $\mathbf{V}_1$, $\boldsymbol{\Sigma}$ kovaryans matrisinin özvektörlerinden oluşan $p \times p$ boyutlu bir matris olmak üzere temel bileşen skorları,

$$\mathbf{T}_X = \mathbf{XV}_1 \quad (5.9)$$

şeklinde elde edilir. $\boldsymbol{\Sigma}$ matrisinin özvektörleri yerine $\mathbf{R}$ korelasyon matrisinin özvektörlerinden oluşan $\mathbf{V}_2$ matrisinin kullanılması durumunda temel bileşen skorları,

$$\mathbf{T}_Z = \mathbf{ZV}_2 \quad (5.10)$$

formülü ile elde edilir. Burada $\mathbf{Z}$, $n \times p$ boyutlu standartlaştırılmış veri matrisidir. $\boldsymbol{\Sigma}$ kovaryans matrisinin özdeğerleri $\lambda_1 > \lambda_2 > \dots > \lambda_p$ olmak üzere, k . temel bileşenin varyansı açıklama oranı,

$$AO_k = \frac{\lambda_k}{\sum_{i=1}^p \lambda_i} \quad (5.11)$$

eşitliği yardımıyla elde edilir.

5.3. Simülasyon Çalışması

Yapılacak simülasyon çalışmasında değişken sayısı 10 olan 100 birimlik veri setleri, veri içerisinde 2 veya 3 küme olacak şekilde üretilecektir. Veri setleri üretilirken göz önünde tutulacak özellikler Tablo 5.1’de verilmiştir.

Tablo 5.1. Veri setlerinin üretiminde göz önünde tutulacak durumlar

Durum	Küme Sayısı	Ortalama Vektörü	Kovaryans Matrisi
Durum I	$2(n_1 = 30, n_2 = 70)$	Farklı	Eşit
Durum II	$2(n_1 = 30, n_2 = 70)$	Farklı	Farklı
Durum III	$3(n_1 = 25, n_2 = 35, n_3 = 40)$	Farklı	Eşit
Durum IV	$3(n_1 = 25, n_2 = 35, n_3 = 40)$	Farklı	Farklı

Veri setleri üretildikten sonra standartlaştırılacak, hem korelasyon matrisi hem de kovaryans matrisi kullanılarak iki defa temel bileşenler analizi uygulanacaktır. Temel bileşenler analizi sonrası k ortalamalar tekniği ile kümelemeler gerçekleştirilecektir. Gerçekleştirilecek kümeleme uygulamalarında göz önünde tutulacak veri matrisleri Tablo 5.2’de gösterilmiştir.

Tablo 5.2. KA uygulamasının gerçekleştirileceği veri matrisleri

KA1	KA3	KA4	KA5	...	KA12
X	T_{X1}	T_{X2}	T_{X3}	...	T_{X10}
KA2	KA13	KA14	KA15	...	KA22
Z	T_{Z1}	T_{Z2}	T_{Z3}	...	T_{Z10}

T_{X3} : Kovaryans matrisi kullanılarak elde edilen bileşenlerden ilk üçünden oluşan veri matrisi

T_{Z3} : Korelasyon matrisi kullanılarak elde edilen bileşenlerden ilk üçünden oluşan veri matrisi

Çalışmada, elde edilen kümeleme sonuçlarının gerçek küme üyelikleri ile karşılaştırması yapılacaktır. Kümeleme sonuçlarının gerçek küme üyelikleri ile karşılaştırılmasında doğru sınıflandırma yüzdesi kullanılacaktır. Ayrıca orijinal değişkenler ile kovaryans matrisinden elde edilen bileşenlerin kümeleme sonuçları arasındaki uyum ve standartlaştırılmış değişkenler ile korelasyon matrisinden elde

edilen bileşenlerin kümeleme sonuçları arasındaki uyum eşleştirme katsayısı ile incelenecektir. Bu simülasyon çalışması için gerekli algoritma SPSS Syntax’de yazılmıştır. Tablo 5.1’de verilen her durum için 100 tekrar yapılacaktır. Toplamda 400 farklı veri seti kullanılarak 8800 adet kümeleme analizi sonucu incelenecektir.

5.4. Uygulama

Kümeleme sonucu oluşan küme üyelikleri ile gerçek küme üyeliklerinin karşılaştırılması sonucu doğru sınıflandırılan birim sayısı r ve $\mathbf{X}$ veri matrisindeki birim sayısı n olmak üzere doğru sınıflandırma yüzdesi

$$Ds(\%) = \frac{r}{n} \times 100 \quad (5.12)$$

eşitliği ile elde edilir. Simülasyon çalışmasında $n=100$ alındığı için doğru sınıflandırma yüzdesi uygulamamızda doğru sınıflandırılan birim sayısına eşittir. t . tekrarda ($t=1, \dots, 100$), üretilen veri matrisinin kümelenmesinde elde edilen doğru sınıflandırılan birim sayısı r_t olmak üzere 100 tekrar sonucu elde edilen ortalama doğru sınıflandırma yüzdesi

$$\overline{Ds}(\%) = \frac{\sum_{t=1}^{100} r_t}{100} \quad (5.13)$$

eşitliği ile elde edilir. Küme sayısının 2 ve kümelere ait kovaryans matrislerinin eşit olması $\Sigma_1 = \Sigma_2$ durumunda, kümeleme sonucu oluşan küme üyelikleri ile gerçek küme üyeliklerinin karşılaştırılmasıyla elde edilen doğru sınıflandırma yüzdesine ait ortalama, ortalama için %95 güven aralığı ve kümelemede kullanılan bileşenlerin birikimli varyans açıklama oranları Tablo 5.3’de verilmiştir.

Tablo 5.3. Durum 1 için doğru sınıflandırma yüzdesine ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin birikimli varyans açıklama oranları

Matris	B. Varyans Açıklama Oranı	Alt Sınır	$\overline{Ds}(\%)$	Üst Sınır
X		67.33	69.77	72.21
Z		64.41	66.61	68.81
T_{X1}	22.76	69.92	72.11	74.30
T_{X2}	39.10	68.26	70.47	72.68
T_{X3}	51.91	68.59	70.85	73.11
T_{X4}	62.84	67.56	69.99	72.42
T_{X5}	72.14	66.73	69.11	71.49
T_{X6}	80.06	66.75	69.15	71.55
T_{X7}	86.72	67.24	69.67	72.10
T_{X8}	92.18	67.29	69.75	72.21
T_{X9}	96.65	66.99	69.39	71.79
T_{X10}	100.00	67.33	69.77	72.21
T_{Z1}	21.27	66.85	69.03	71.21
T_{Z2}	37.35	64.37	66.76	69.15
T_{Z3}	50.47	64.84	67.16	69.48
T_{Z4}	61.62	65.30	67.64	69.98
T_{Z5}	71.21	64.90	67.19	69.48
T_{Z6}	79.38	65.83	68.04	70.25
T_{Z7}	86.29	65.48	67.78	70.08
T_{Z8}	91.97	64.75	66.99	69.23
T_{Z9}	96.60	63.55	66.05	68.55
T_{Z10}	100.00	64.41	66.61	68.81

Tablo 5.3 incelendiğinde, küme sayısının 2 ve kümelere ait kovaryans matrislerinin eşit olması $\Sigma_1 = \Sigma_2$ koşulunda kümelemede ham veri yerine standartlaştırılmış verinin kullanılması durumunda doğru sınıflandırma yüzdesi düşmektedir. Kümelemede orijinal değişkenler yerine kümeleme analizi öncesi uygulanan temel bileşenler analizinden elde edilecek temel bileşenlerin kullanılması durumunda en yüksek doğru sınıflandırma yüzdesi sadece ilk temel bileşenin kullanılmasıyla elde edilmiştir. Doğru sınıflandırma yüzdesi bakımından elde edilen bu sonuç Jolliffe ve ark (1980)'nin sonuçları ile benzerdir.

Küme sayısının 2 ve kümelere ait kovaryans matrislerinin farklı olması ($\Sigma_1 \neq \Sigma_2$) durumunda, kümeleme sonucu oluşan küme üyelikleri ile gerçek küme üyeliklerinin karşılaştırılmasıyla elde edilen doğru sınıflandırma yüzdesine ait

ortalama, ortalama için %95 güven aralığı ve kümelemede kullanılan bileşenlerin birikimli varyans açıklama oranları Tablo 5.4’de verilmiştir.

Tablo 5.4.Durum 2 için doğru sınıflandırma yüzdesine ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin ortalama birikimli varyans açıklama oranları

Matris	B. Varyans Açıklama Oranı	Alt Sınır	$\overline{Ds}(\%)$	Üst Sınır
X		59.88	61.41	62.94
Z		58.80	60.27	61.74
T_{X1}	21.46	60.43	62.04	63.65
T_{X2}	37.48	59.00	60.48	61.96
T_{X3}	50.70	59.32	60.83	62.34
T_{X4}	61.87	59.42	60.98	62.54
T_{X5}	71.49	59.91	61.51	63.11
T_{X6}	79.63	60.08	61.65	63.22
T_{X7}	86.44	60.30	61.92	63.54
T_{X8}	92.03	59.27	60.81	62.35
T_{X9}	96.55	59.75	61.28	62.81
T_{X10}	100.00	59.88	61.41	62.94
T_{Z1}	20.51	58.76	60.22	61.68
T_{Z2}	36.28	58.51	59.90	61.29
T_{Z3}	49.54	59.08	60.50	61.92
T_{Z4}	60.83	59.11	60.60	62.09
T_{Z5}	70.65	59.15	60.56	61.97
T_{Z6}	78.99	58.35	59.77	61.19
T_{Z7}	86.03	58.60	60.08	61.56
T_{Z8}	91.80	58.31	59.78	61.25
T_{Z9}	96.47	59.17	60.58	61.99
T_{Z10}	100.00	58.80	60.27	61.74

Tablo 5.4 incelendiğinde, kümelere ait kovaryans matrislerinin farklı olması durumunda ortalama doğru sınıflandırma yüzdesinin eşit olma durumuna göre düştüğü görülmektedir. Burada da ham veri yerine standartlaştırılmış verilerin kullanılmasının kümelemeyi olumsuz etkilediği görülmektedir. Ancak doğru sınıflandırma yüzdesindeki düşüş kovaryans matrislerinin eşit olması durumundaki düşüşe göre daha azdır. Kovaryans matrisi kullanılarak (5.9) nolu eşitlik ile hesaplanan temel bileşen skorlarıyla kümeleme yapıldığında en yüksek doğru sınıflandırma yüzdesi orijinal değişkenler yerine sadece tek bileşenin kullanılması durumunda elde edilmiştir. Korelasyon matrisi kullanılarak hesaplanan temel bileşen

skorlarıyla kümeleme yapıldığında en yüksek doğru sınıflandırma yüzdesi orijinal değişkenler yerine ilk dört temel bileşenin kullanılması durumunda elde edilmiştir.

Küme sayısının 3 ve kümelere ait kovaryans matrislerinin eşit olması ($\Sigma_1 = \Sigma_2 = \Sigma_3$) durumunda, kümeleme sonucu oluşan küme üyelikleri ile gerçek küme üyeliklerinin karşılaştırılmasıyla elde edilen doğru sınıflandırma yüzdesine ait ortalama, ortalama için %95 güven aralığı ve kümelemede kullanılan bileşenlerin birikimli varyans açıklama oranları Tablo 5.5’de verilmiştir.

Tablo 5.5.Durum 3 için doğru sınıflandırma yüzdesine ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin ortalama birikimli varyans açıklama oranları

Matris	B. Varyans Açıklama Oranı	Alt Sınır	$\overline{Ds}(\%)$	Üst Sınır
X		70.23	72.35	74.47
Z		67.40	69.53	71.66
T_{X1}	36.92	63.55	65.17	66.79
T_{X2}	50.63	71.34	73.49	75.64
T_{X3}	61.03	70.90	72.95	75.00
T_{X4}	69.84	72.02	74.17	76.32
T_{X5}	77.33	71.29	73.49	75.69
T_{X6}	83.77	70.21	72.43	74.65
T_{X7}	89.18	70.72	72.83	74.94
T_{X8}	93.63	70.72	72.72	74.72
T_{X9}	97.21	71.08	73.19	75.30
T_{X10}	100.00	70.23	72.35	74.47
T_{Z1}	33.24	61.16	62.70	64.24
T_{Z2}	47.65	67.77	69.92	72.07
T_{Z3}	58.95	68.43	70.58	72.73
T_{Z4}	68.54	68.73	70.86	72.99
T_{Z5}	76.58	67.32	69.48	71.64
T_{Z6}	83.43	67.51	69.61	71.71
T_{Z7}	89.04	67.61	69.68	71.75
T_{Z8}	93.66	68.37	70.41	72.45
T_{Z9}	97.32	67.46	69.58	71.70
T_{Z10}	100.00	67.40	69.53	71.66

Küme sayısının 3 ve kümelere ait kovaryans matrislerinin eşit olması ($\Sigma_1 = \Sigma_2 = \Sigma_3$) durumunda orijinal değişkenler yerine hem kovaryans matrisi hem de korelasyon matrisi ile hesaplanan temel bileşenlerden ilk dört tanesinin kullanılması ile elde edilen kümelemelerde en yüksek doğru sınıflandırma yüzdesi elde edilmiştir. Tablo 5.5 incelendiğinde burada da ham veri yerine

standartlaştırılmış verinin kullanılmasının kümeleme kalitesini bozduğu görülmektedir.

Küme sayısının 3 ve kümelere ait kovaryans matrislerinin farklı olması (*en az biri* $\Sigma_i \neq \Sigma_j$) durumunda, kümeleme sonucu oluşan küme üyelikleri ile gerçek küme üyeliklerinin karşılaştırılmasıyla elde edilen doğru sınıflandırma yüzdesine ait ortalama, ortalama için %95 güven aralığı ve kümelemede kullanılan bileşenlerin birikimli varyans açıklama oranları Tablo 5.6’da verilmiştir.

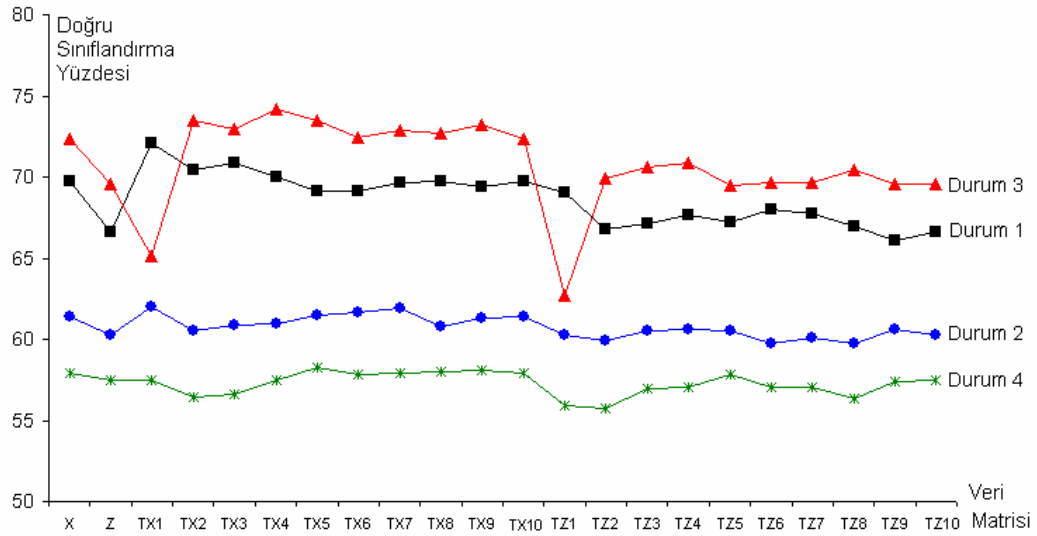
Tablo 5.6. Durum 4 için doğru sınıflandırma yüzdesine ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin ortalama birikimli varyans açıklama oranları

Matris	B. Varyans Açıklama Oranı	Alt Sınır	$\overline{Ds}(\%)$	Üst Sınır
X		56.50	57.87	59.24
Z		55.97	57.45	58.93
T_{X1}	30.23	56.28	57.51	58.74
T_{X2}	45.10	55.03	56.43	57.83
T_{X3}	57.12	55.22	56.61	58.00
T_{X4}	67.10	56.10	57.51	58.92
T_{X5}	75.45	56.68	58.23	59.78
T_{X6}	82.60	56.40	57.86	59.32
T_{X7}	88.46	56.42	57.94	59.46
T_{X8}	93.26	56.37	57.99	59.61
T_{X9}	97.12	56.55	58.07	59.59
T_{X10}	100.00	56.50	57.87	59.24
T_{Z1}	28.06	54.65	55.87	57.09
T_{Z2}	43.15	54.38	55.71	57.04
T_{Z3}	55.51	55.42	56.92	58.42
T_{Z4}	65.87	55.69	57.04	58.39
T_{Z5}	74.57	56.20	57.83	59.46
T_{Z6}	82.04	55.61	57.03	58.45
T_{Z7}	88.08	55.56	57.01	58.46
T_{Z8}	93.07	54.97	56.39	57.81
T_{Z9}	97.05	55.96	57.40	58.84
T_{Z10}	100.00	55.97	57.45	58.93

Tablo 5.6 incelendiğinde ortalama doğru sınıflandırma yüzdesi bakımından orijinal değişkenler yerine ilk birkaç temel bileşenin kullanılması ile gerçekleştirilen kümelemenin daha başarılı olduğu görülmektedir. Küme sayısının 3 ve kümelere ait kovaryans matrislerinin farklı olması (*en az biri* $\Sigma_i \neq \Sigma_j$) durumunda da ham veri

yerine standartlaştırılmış verinin kullanılması doğru sınıflandırma yüzdesini düşürmektedir.

Dört farklı durum için elde edilen ortalama doğru sınıflandırma yüzdesi değerlerinin birlikte değerlendirildiği çizgi grafiği Şekil 5.1’de verilmiştir.



Şekil 5.1. Farklı dört durum için ortalama doğru sınıflandırma yüzdesi

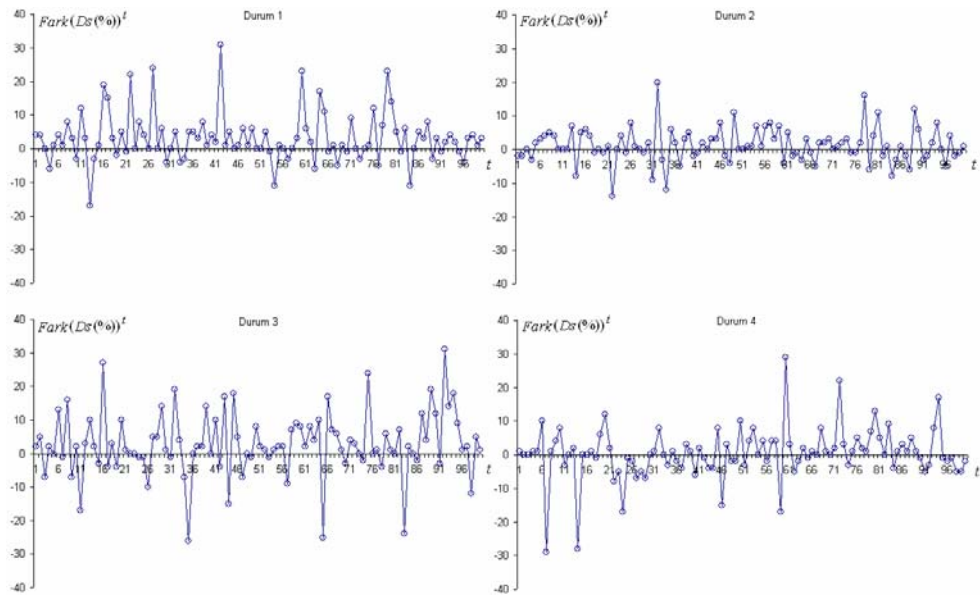
Şekil 5.1 incelendiğinde, en yüksek ortalama doğru sınıflandırma yüzdesinin kovaryans matrislerinin eşit ve küme sayısının 2 olduğu Durum 3’de elde edildiği görülmektedir. Kümelerin varyans kovaryans matrislerinin farklı olması durumunda ortalama doğru sınıflandırma yüzdeslerinin düştüğü görülmektedir. Kovaryans matrisinden elde edilen temel bileşenler ile gerçekleştirilen kümelemeler içerisinde en yüksek ortalama doğru sınıflandırma yüzdesi, Durum 1 ile Durum 2’de sadece ilk temel bileşenin kullanılmasıyla, Durum 3’de ilk dört temel bileşen kullanılmasıyla ve Durum 4’de ilk beş temel bileşenin kullanılmasıyla elde edilmiştir. Korelasyon matrisinden elde edilen bileşenler için en yüksek ortalama doğru sınıflandırma yüzdesi Durum 1’de ilk temel bileşenin kullanılmasıyla, Durum 2 ile Durum 3’de ilk dört temel bileşenin kullanılmasıyla ve Durum 4’de ilk beş temel bileşenin kullanılmasıyla elde edilmiştir. Doğru sınıflandırma yüzdesi bakımından kümelemede orijinal değişkenler yerine temel bileşenler analizinden elde edilen birkaç temel bileşenin kullanılmasıyla daha başarılı kümelemeler elde edilmiştir.

Kümelemede ham veri yerine standartlaştırılmış verinin kullanılması durumunda doğru sınıflandırma yüzdesinin düştüğü görülmüştür.

Simülasyonda t . tekrarda ham veri ile gerçekleştirilen kümelemeden elde edilen doğru sınıflandırma yüzdesi ile standartlaştırılmış veri ile gerçekleştirilen kümelemeden elde edilen doğru sınıflandırma yüzdesi arasındaki fark

$$Fark(Ds(\%))^t = Ds(\%)^t_x - Ds(\%)^t_z, \quad t = 1, \dots, 100 \quad (5.14)$$

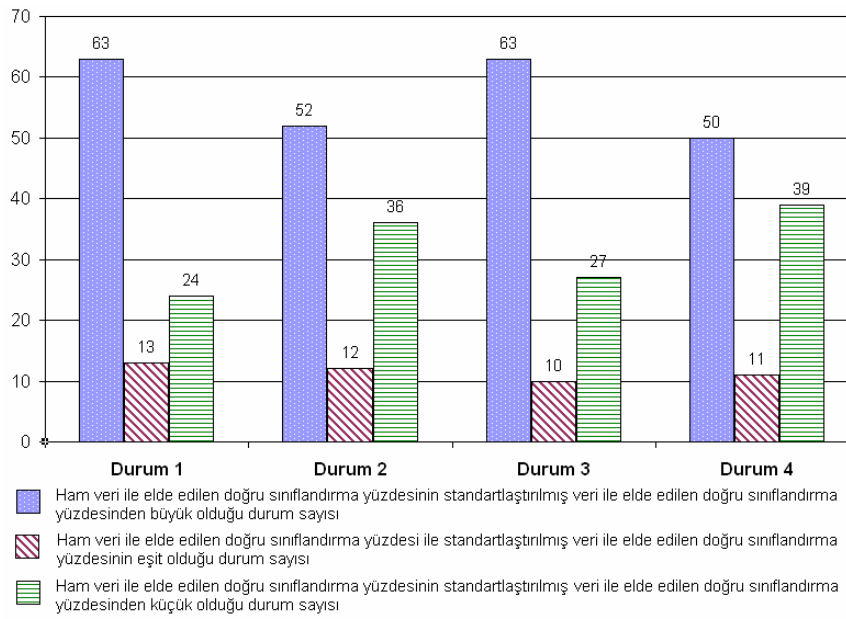
eşitliği ile elde edilir. Eşitlikte yer alan $Ds(\%)^t_x$, t . tekrarda ham veri matrisi X kullanılarak gerçekleştirilen kümelemeden elde edilen doğru sınıflandırma yüzdesini gösterir. $Fark(Ds(\%))^t$ değerlerine ait değişim grafiği Şekil 5.2’de verilmiştir.



Şekil 5.2. t . tekrarda ham veri ile elde edilen doğru sınıflandırma yüzdesi ile standartlaştırılmış veri ile doğru sınıflandırma yüzdesi arasındaki farkların değişimi

$Fark(Ds(\%))^t$ değerinin pozitif olması, t . tekrarda ham veri ile gerçekleştirilen kümelemeden elde edilen doğru sınıflandırma yüzdesinin, standartlaştırılmış veri kullanılarak gerçekleştirilen kümelemedeki doğru sınıflandırma yüzdesinden daha yüksek olduğunu gösterir. Şekil 5.2 incelendiğinde

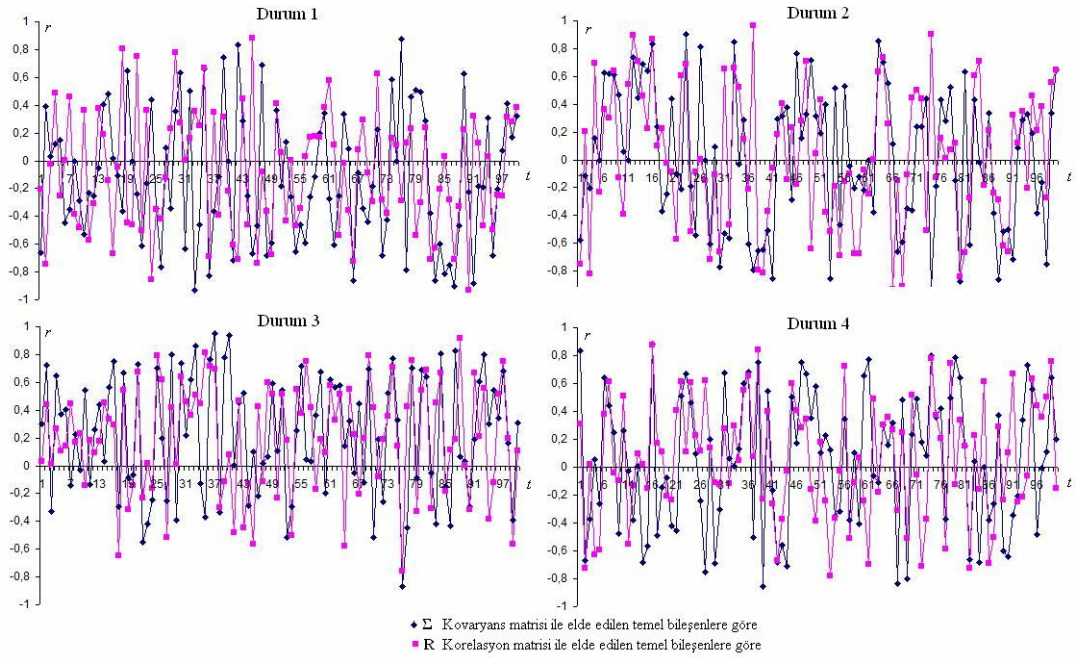
dört durumun dördünde de $Fark(Ds(\%))^t$ değerlerinin pozitif olduğu durum sayısı daha fazladır. Bununla birlikte, kümelere ait kovaryans matrislerinin farklı olması durumunda ham veri ile standartlaştırılmış verilere ait doğru sınıflandırma yüzdeleri arasındaki farklar azalmaktadır. Bu üstünlük değerlendirilmesine ait grafik Şekil 5.3’de verilmiştir.



Şekil 5.3. Ham veri ile kümelemenin standartlaştırılmış veri ile kümelemeyle doğru sınıflandırma yüzdesi bakımından karşılaştırılması

Şekil 5.3’den de görüldüğü gibi ham veri yerine standartlaştırılmış veri ile kümeleme yapıldığında doğru sınıflandırma yüzdesinin düşmesi beklenmektedir.

Birikimli varyans açıklama oranının değişimine göre doğru sınıflandırma yüzdesindeki değişimin nasıl olduğunu görebilmek için, her bir veri seti için kümelemede kullanılan bileşen sayısına ait birikimli varyans açıklama oranları ile kullanılan bileşen sayısına göre elde edilen doğru sınıflandırma yüzdeleri arasındaki korelasyonlar hesaplanmış ve bu korelasyon değerlerinin üretilen her bir veri seti için değişimi Şekil 5.4’de verilmiştir.



Şekil 5.4. Kümelemede kullanılan bileşen sayısına ait birikimli varyans açıklama oranları ile kümelemede kullanılan bileşen sayısına göre elde edilen doğru sınıflandırma yüzdeleri arasındaki korelasyonlar

Şekil 5.4 incelendiğinde, kümelemeden elde edilecek doğru sınıflandırma yüzdesindeki değişimin kümelemede kullanılan bileşen sayısına ait birikimli varyans açıklama oranı ile açıklanamayacağı görülmektedir. Yani kümeleme analizinde orijinal değişkenler yerine temel bileşenler kullanılması durumunda, kullanılan temel bileşen sayısı arttıkça doğru sınıflandırma yüzdesi artacaktır veya azalacaktır şeklinde bir açıklama yapmak mümkün değildir.

Uygulamada kümeleme analizi öncesi küme yapısı hakkında herhangi bir ön bilgi bulunmaz. Bu nedenle kümelemede orijinal değişkenler yerine temel bileşenlerin kullanılmasının kümelemeye olan etkisi incelenirken orijinal değişkenlerle gerçekleştirilen kümelemeden elde edilen küme üyelikleri ile temel bileşenlerle gerçekleştirilen kümelemeden elde edilen küme üyeliklerinin de karşılaştırılması uygun olacaktır.

Eşleştirme katsayısı, orijinal değişkenler ve temel bileşenlerle elde edilen her iki kümeleme sonucu aynı kümelerde yer alan birimlerin sayısı m olmak üzere

$$E(\%) = \frac{m}{n} \times 100 \quad (5.15)$$

eşitliği ile elde edilir. Simülasyon çalışmasında $n = 100$ alındığı için eşleştirme katsayısı uygulamamızda orijinal değişkenler ve temel bileşenlerle elde edilen her iki kümelemede aynı kümede yer alan birim sayısı olan m değerine eşit olacaktır.

Kovaryans matrisi Σ kullanılarak elde edilen temel bileşenlerle kümeleme için eşleştirme katsayısı hesaplanırken, kümelemede temel bileşenlerin kullanılmasıyla elde edilen küme üyelikleri ham veri ile gerçekleştirilen kümeleme sonucu oluşan küme üyelikleri ile karşılaştırılır. Korelasyon matrisi R kullanılarak elde edilen bileşenlerle kümeleme için eşleştirme katsayısı hesaplanırken ise, kümelemede temel bileşenlerin kullanılmasıyla elde edilen küme üyelikleri standartlaştırılmış veri ile gerçekleştirilen kümeleme sonucu oluşan küme üyelikleri ile karşılaştırılır.

t . tekrarda ($t = 1, \dots, 100$), üretilen veri matrisi için orijinal değişkenler ve temel bileşenlerle elde edilen her iki kümelemede aynı kümede yer alan birimlerin sayısı m_t olmak üzere ortalama eşleştirme katsayısı

$$\bar{E}(\%) = \frac{\sum_{t=1}^{100} m_t}{100} \quad (5.16)$$

eşitliği ile elde edilir. Küme sayısının 2 ve kümelere ait kovaryans matrislerinin eşit olması ($\Sigma_1 = \Sigma_2$) durumunda, eşleştirme katsayısına ait ortalama, ortalama için %95 güven aralığı ve kümelemede kullanılan bileşenlerin birikimli varyans açıklama oranları Tablo 5.7’de verilmiştir.

Tablo 5.7. Durum 1 için eşleştirme katsayısına ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin ortalama birikimli varyans açıklama oranları

Matris	B. Varyans Açıklama Oranı	Alt Sınır	$\bar{E}(\%)$	Üst Sınır
T_{X1}	22.76	87.52	89.56	91.60
T_{X2}	39.10	85.28	87.77	90.26
T_{X3}	51.91	86.22	88.69	91.16
T_{X4}	62.84	86.81	89.35	91.89
T_{X5}	72.14	86.43	89.07	91.71
T_{X6}	80.06	88.56	90.99	93.42
T_{X7}	86.72	89.50	91.91	94.32
T_{X8}	92.18	91.11	93.27	95.43
T_{X9}	96.65	95.85	97.09	98.33
T_{X10}	100.00	100.00	100.00	100.00
T_{Z1}	21.27	83.40	86.07	88.74
T_{Z2}	37.35	84.10	86.63	89.16
T_{Z3}	50.47	84.55	87.24	89.93
T_{Z4}	61.62	83.87	86.68	89.49
T_{Z5}	71.21	85.25	88.16	91.07
T_{Z6}	79.38	87.10	89.60	92.10
T_{Z7}	86.29	89.35	91.71	94.07
T_{Z8}	91.97	90.66	92.90	95.14
T_{Z9}	96.60	94.66	96.14	97.62
T_{Z10}	100.00	100.00	100.00	100.00

Tablo 5.7 incelendiğinde kümelemede orijinal değişkenler yerine kullanılan temel bileşen sayısı arttıkça eşleştirme katsayısının genelde arttığı görülmektedir. Kovaryans matrisi Σ kullanılarak elde edilen bileşenlerle kümelemede en düşük eşleştirme katsayısı yaklaşık %88 olarak ilk iki bileşenin kullanıldığı durumda elde edilmiştir. Korelasyon matrisi R kullanılarak elde edilen bileşenlerle kümelemede en düşük eşleştirme katsayısı yaklaşık %86 olarak sadece ilk temel bileşenin kullanıldığı durumda elde edilmiştir.

Küme sayısının 2 ve kümelere ait kovaryans matrislerinin farklı olması ($\Sigma_1 \neq \Sigma_2$) durumunda, eşleştirme katsayısına ait ortalama, ortalama için %95 güven aralığı ve kümelemede kullanılan bileşenlerin birikimli varyans açıklama oranları Tablo 5.8’de verilmiştir.

Tablo 5.8. Durum 2 için eşleştirme katsayısına ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin ortalama birikimli varyans açıklama oranları

Matris	B. Varyans Açıklama Oranı	Alt Sınır	$\bar{E}(\%)$	Üst Sınır
T_{X1}	21.46	85.83	88.02	90.21
T_{X2}	37.48	86.23	88.42	90.61
T_{X3}	50.70	84.97	87.54	90.11
T_{X4}	61.87	86.19	88.63	91.07
T_{X5}	71.49	88.16	90.41	92.66
T_{X6}	79.63	88.67	90.84	93.01
T_{X7}	86.44	88.84	91.14	93.44
T_{X8}	92.03	92.75	94.42	96.09
T_{X9}	96.55	93.53	95.29	97.05
T_{X10}	100.00	100.00	100.00	100.00
T_{Z1}	20.51	83.63	85.97	88.31
T_{Z2}	36.28	82.87	85.33	87.79
T_{Z3}	49.54	83.34	85.81	88.28
T_{Z4}	60.83	84.97	87.39	89.81
T_{Z5}	70.65	86.96	89.33	91.70
T_{Z6}	78.99	88.26	90.64	93.02
T_{Z7}	86.03	87.60	90.14	92.68
T_{Z8}	91.80	89.80	92.08	94.36
T_{Z9}	96.47	92.96	94.76	96.56
T_{Z10}	100.00	100.00	100.00	100.00

Tablo 5.8 incelendiğinde kümelemede orijinal değişkenler yerine kullanılan temel bileşen sayısı arttıkça eşleştirme katsayısının genelde arttığı görülmektedir. Kovaryans matrisi Σ kullanılarak elde edilen bileşenlerle kümelemede en düşük eşleştirme katsayısı yaklaşık %87.5 olarak ilk üç bileşenin kullanıldığı durumda elde edilmiştir. Korelasyon matrisi R kullanılarak elde edilen bileşenlerle kümelemede en düşük eşleştirme katsayısı yaklaşık %85.3 olarak sadece ilk temel bileşenin kullanıldığı durumda elde edilmiştir.

Küme sayısının 3 ve kümelere ait kovaryans matrislerinin eşit olması ($\Sigma_1 = \Sigma_2 = \Sigma_3$) durumunda eşleştirme katsayısına ait ortalama, ortalama için %95 güven aralığı ve kümelemede kullanılan bileşenlerin birikimli varyans açıklama oranları Tablo 5.9'da verilmiştir.

Tablo 5.9. Durum 3 için eşleştirme katsayısına ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin ortalama birikimli varyans açıklama oranları

Matris	B. Varyans Açıklama Oranı	Alt Sınır	$\bar{E}(\%)$	Üst Sınır
T_{X1}	36.92	72.15	74.39	76.63
T_{X2}	50.63	79.30	81.92	84.54
T_{X3}	61.03	79.58	82.28	84.98
T_{X4}	69.84	82.26	84.77	87.28
T_{X5}	77.33	82.08	84.65	87.22
T_{X6}	83.77	83.24	85.88	88.52
T_{X7}	89.18	85.85	88.36	90.87
T_{X8}	93.63	87.15	89.67	92.19
T_{X9}	97.21	88.62	91.02	93.42
T_{X10}	100.00	100.00	100.00	100.00
T_{Z1}	33.24	69.09	71.20	73.31
T_{Z2}	47.65	78.43	81.14	83.85
T_{Z3}	58.95	77.67	80.47	83.27
T_{Z4}	68.54	76.88	79.56	82.24
T_{Z5}	76.58	79.57	82.48	85.39
T_{Z6}	83.44	81.84	84.50	87.16
T_{Z7}	89.05	80.95	83.73	86.51
T_{Z8}	93.67	84.29	86.93	89.57
T_{Z9}	97.33	86.54	89.20	91.86
T_{Z10}	100.00	100.00	100.00	100.00

Tablo 5.9 incelendiğinde kümelemede orijinal değişkenler yerine kullanılan temel bileşen sayısı arttıkça eşleştirme katsayısının genelde arttığı görülmektedir. Ancak küme sayısının artması ortalama eşleştirme katsayısı değerini düşürmüştür. Kovaryans matrisi Σ kullanılarak elde edilen bileşenlerle kümelemede en düşük eşleştirme katsayısı yaklaşık %74.4 olarak ilk bileşenin kullanıldığı durumda elde edilmiştir. Korelasyon matrisi R kullanılarak elde edilen bileşenlerle kümelemede en düşük eşleştirme katsayısı yaklaşık %71 olarak ilk temel bileşenin kullanıldığı durumda elde edilmiştir.

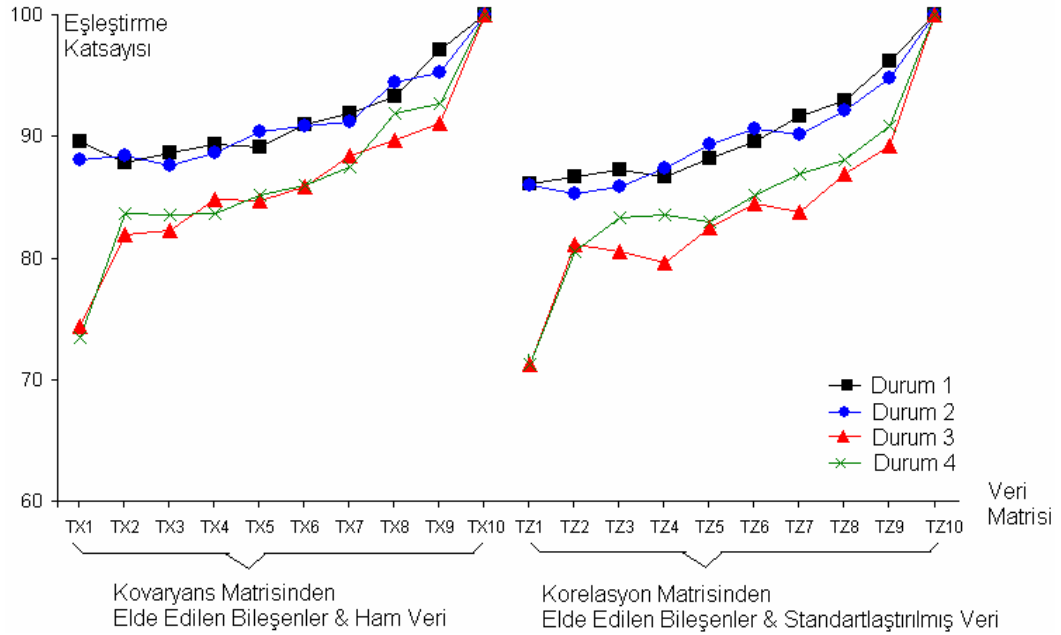
Küme sayısının 3 ve kümelere ait kovaryans matrislerinin farklı olması (*en az biri* $\Sigma_i \neq \Sigma_j$) durumunda, eşleştirme katsayısına ait ortalama, ortalama için %95 güven aralığı ve kümelemede kullanılan bileşenlerin birikimli varyans açıklama oranları Tablo 5.10'da verilmiştir.

Tablo 5.10. Durum 4 için eşleştirme katsayısına ait ortalama, ortalama için %95 güven aralıkları ve kümelemede kullanılan bileşenlerin ortalama birikimli varyans açıklama oranları

Matris	B. Varyans Açıklama Oranı	Alt Sınır	$\bar{E}(\%)$	Üst Sınır
T_{X1}	30.23	71.56	73.45	75.34
T_{X2}	45.10	81.41	83.65	85.89
T_{X3}	57.12	81.38	83.49	85.60
T_{X4}	67.10	81.29	83.66	86.03
T_{X5}	75.45	82.81	85.19	87.57
T_{X6}	82.60	83.72	85.94	88.16
T_{X7}	88.46	84.97	87.48	89.99
T_{X8}	93.26	89.93	91.85	93.77
T_{X9}	97.12	90.77	92.69	94.61
T_{X10}	100	100	100	100
T_{Z1}	28.06	69.14	71.25	73.36
T_{Z2}	43.15	77.97	80.51	83.05
T_{Z3}	55.51	80.74	83.31	85.88
T_{Z4}	65.87	80.91	83.50	86.09
T_{Z5}	74.57	80.46	83.01	85.56
T_{Z6}	82.04	82.60	85.18	87.76
T_{Z7}	88.08	84.44	86.91	89.38
T_{Z8}	93.07	85.69	88.02	90.35
T_{Z9}	97.05	88.55	90.85	93.15
T_{Z10}	100	100	100	100

Tablo 5.10 incelendiğinde, kovaryans matrisi Σ kullanılarak elde edilen bileşenlerle kümelemede en düşük eşleştirme katsayısı yaklaşık %73.5 olarak ilk bileşenin kullanıldığı durumda elde edildiği görülmektedir. Korelasyon matrisi R kullanılarak elde edilen bileşenlerle kümelemede ise en düşük eşleştirme katsayısı yaklaşık %71 olarak ilk temel bileşenin kullanıldığı durumda elde edilmiştir.

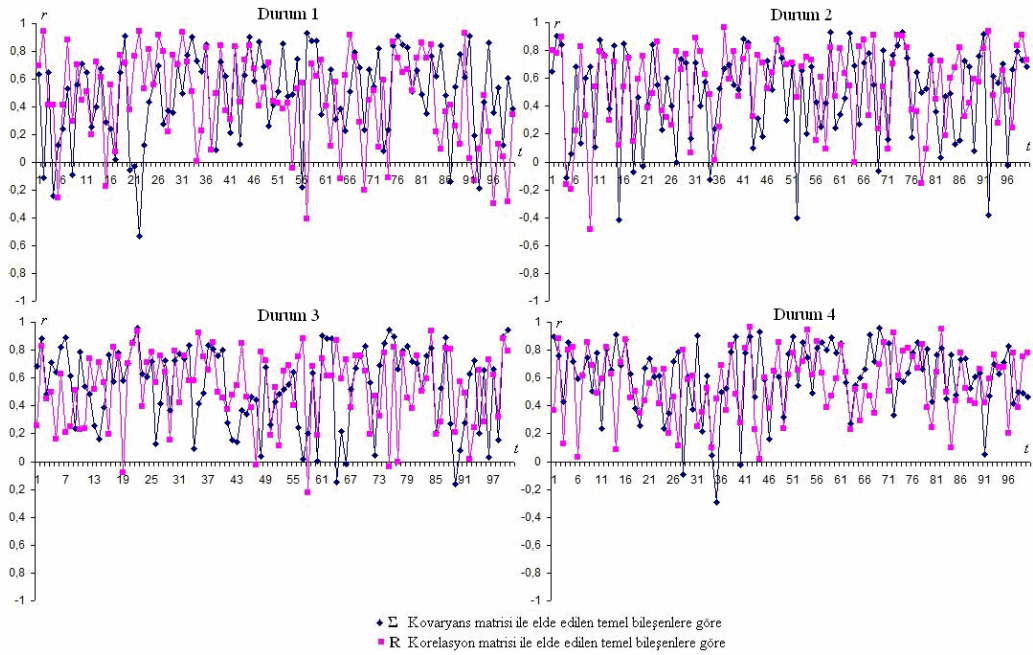
Dört farklı durum için elde edilen ortalama eşleştirme katsayısı değerlerinin birlikte değerlendirildiği çizgi grafiği Şekil 5.5’de verilmiştir.



Şekil 5.5. Farklı dört durum için ortalama eşleştirme katsayısı değerleri

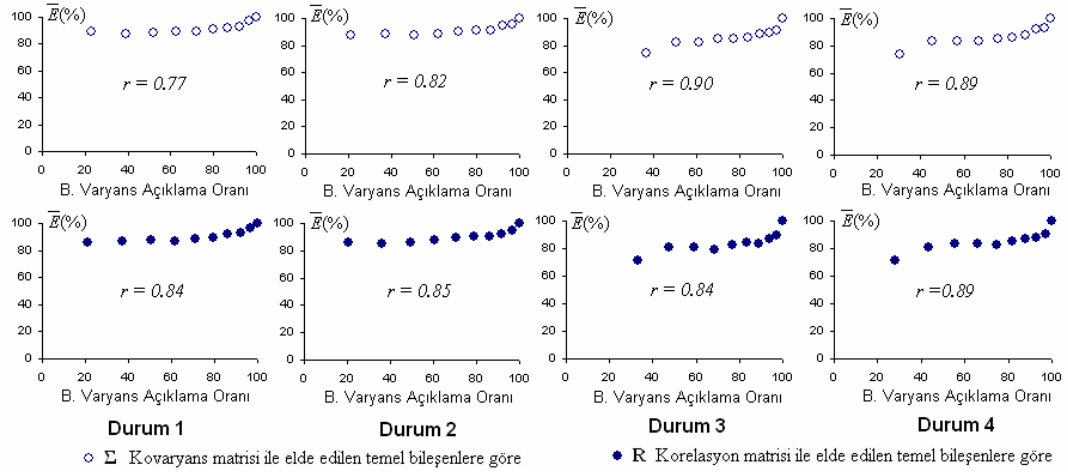
Şekil 5.5 incelendiğinde küme sayısının 3 olduğu Durum 3 ve Durum 4’de ortalama eşleştirme katsayısının, küme sayısının 2 olduğu Durum 1 ve Durum 2’ye göre daha düşük olduğu görülmektedir. Ayrıca orijinal değişkenler yerine kümelemede kullanılacak temel bileşen sayısının artması durumunda eşleştirme katsayısı değerinin artışı görülmektedir.

Orijinal değişkenler yerine temel bileşenlerle gerçekleştirilen kümelemelerden elde edilen eşleştirme katsayılarının, kümelemelerde kullanılan bileşenlerin birikimli varyans açıklama oranlarına göre değişimini gösteren saçılım grafikleri Şekil 5.6’da verilmiştir.



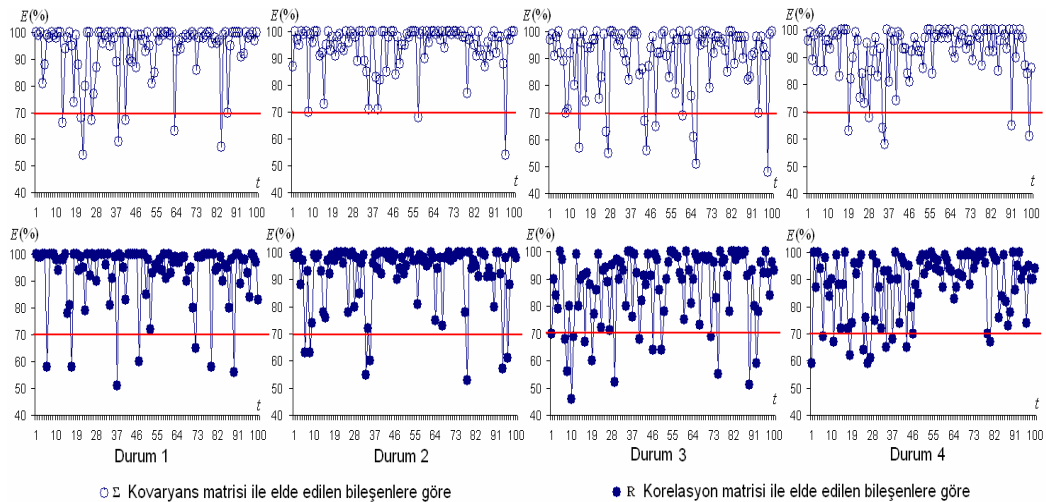
Şekil 5.6. Kümelerden elde edilen eşleştirme katsayıları ile kümelerden kullanılan bileşenlerin birikimli varyans açıklama oranları arasındaki saçılım grafikleri

Şekil 5.6 incelendiğinde küme sayısının 2 olduğu Durum 1 ve Durum 2’de negatif korelasyonla, küme sayısının 3 olduğu Durum 3 ve Durum 4’e göre daha fazla karşılaşılmıştır. Bunun yanında her dört durumda da korelasyonun genel olarak pozitif yönlü olduğu görülmektedir. Genel olarak beklendiği gibi, orijinal değişkenler yerine temel bileşenlerle gerçekleştirilen kümelemede kullanılacak bileşen sayısının arttırılması durumunda eşleştirme katsayısı artmıştır. Bu sonucu daha açık olarak görebilmek için simülasyon çalışmasından elde edilen ortalama eşleştirme katsayıları ile ortalama birikimli varyans açıklama oranları arasındaki saçılım grafikleri Şekil 5.7’de verilmiştir.



Şekil 5.7. Ortalama eşleştirme katsayıları ile ortalama birikimli varyans açıklama oranları arasındaki saçılım grafikleri

Şekil 5.7 incelendiğinde ortalama eşleştirme katsayıları ile ortalama birikimli varyans açıklama oranları arasında yüksek pozitif korelasyonların olduğu görülmektedir. Buna göre kümelemede orijinal değişkenler yerine kullanılacak temel bileşen sayısı arttıkça, eşleştirme katsayısı artmaktadır. Kümelemede orijinal değişkenler yerine varyans açıklama oranı en yüksek olan ilk sekiz temel bileşenin kullanımı durumunda her bir veri seti için elde edilen eşleştirme katsayılarının değişimi Şekil 5.8’de verilmiştir.



Şekil 5.8. Kümelemede orijinal değişkenler yerine ilk sekiz temel bileşenin kullanılmasıyla elde edilen eşleştirme katsayılarının değişimi

Şekil 5.8 incelendiğinde, orijinal değişkenler yerine ilk sekiz bileşen kullanılarak gerçekleştirilen kümelemelerden elde edilen eşleştirme katsayısı değerinin %70'e eşit veya altında olduğu durum sayısı 72'dir. Buna göre ortalama birikimli varyans açıklama oranı yaklaşık %92 olan ilk sekiz bileşenin kullanıldığı durumda dahi eşleştirme katsayısı değerinin %70'e eşit veya altında olması olasılığı %9'dur. Bu rakam oldukça yüksektir. Kümelemede orijinal değişkenler yerine yaklaşık %92'lik birikimli varyans açıklama oranına sahip ilk sekiz temel bileşenin kullanıldığı durumda dahi her 10 veri matrisinden yaklaşık birinde 100 birimden en az 30 tanesinin farklı kümelerle atanması beklenmektedir.

5.5. Sonuçlar

Bu bölümde kümelemede orijinal değişkenler yerine boyut indirgeyerek temel bileşenlerin kullanılmasının kümelemeye olan etkisi, doğru sınıflandırma yüzdesi ve eşleştirme katsayısı bakımından incelenmiştir. Konu ile ilgili önceki çalışmaların bazılarında kümelemede orijinal değişkenler yerine ilk birkaç temel bileşenin kullanılmasının avantajlı olduğu belirtilirken bazılarında dezavantajlı olduğu belirtilmiştir. Simülasyon çalışması sonunda, doğru sınıflandırma yüzdesi bakımından boyut indirgemenin avantajlı, eşleştirme katsayısı bakımından değerlendirildiğinde ise avantajlı olmadığı görülmüştür. Elde edilen diğer sonuçlar,

- Kümelemede ham veri yerine standartlaştırılmış verinin kullanılması kümeleme performansını düşürmektedir. Ancak, kümelerin kovaryans matrislerinin farklı olması durumunda standartlaştırmanın kümelemeye olumsuz etkisi azalmaktadır.
- Küme sayısı artıkça kümelemede orijinal değişkenler yerine ilk birkaç temel bileşenin kullanılmasıyla elde edilen kümelemelerdeki eşleştirme katsayısı düşmektedir.
- Kümelemede orijinal değişkenler yerine kullanılacak temel bileşen sayısı artıkça eşleştirme katsayısı artmaktadır.

şeklinde özetlenebilir.

6. KÜME SAYISININ TAHMİNİNDE YENİ BİR YÖNTEM

Kümeleme analizindeki önemli problemlerden biride küme sayısının belirlenmesidir. Literatür incelendiğinde küme sayının belirlenmesi ile ilgili birçok farklı yaklaşımın olduğu görülmektedir. Değişken ve gözlem sayısının artması durumunda uygun küme sayısının belirlenmesi oldukça güçleşmektedir. Kümeleme analizi öncesi küme yapısı hakkında herhangi bir ön bilginin olmaması, veri yapısındaki farklılıklar uygun küme sayısının tahmini için genel ve kabul görmüş bir formülün verilmesini zorlaştırmaktadır (Ünler 2006).

6.1. Önemli Bazı Küme Sayısı Tahmin Yöntemleri

Literatürde küme sayısı tahmini için önerilen bir çok yöntem bulunmaktadır. Bu bölümde, Marriott (1971), Calinski ve Harabasz (1974), Mojena (1977), Kaufman ve Rousseeuw (1990) ile Tibshirani ve ark. (2001) tarafından önerilen yöntemler incelenecektir.

6.1.1. Marriott İndeksi

Marriott (1971) uygun küme sayısının seçimi için, verinin kümelendirilmesi sonrası elde edilen birleştirilmiş küme içi kareler ve çarpımlar matrisinin determinantını esas alan d Marriott indeksini önermiştir.

$$d_k = k^2 |\mathbf{E}|, \quad k = 2, 3, \dots \quad (6.1)$$

Eşitlikte yer alan $|\mathbf{E}|$ birleştirilmiş küme içi kareler ve çarpımlar matrisinin determinantını göstermektedir. Marriott (1971) eşitlik (6.1)'de verilen d_k indeks değerini minimum yapan k değerini, uygun küme sayısı olarak önermektedir.

6.1.2. Varyans Açıklama Oran Kriteri

Kümelemenin geçerliliği için Calinsky ve Harabasz (1974) tarafından önerilen varyans açıklama oran kriteri uygun küme sayısının belirlenmesi içinde kullanılmaktadır.

$$CH(k) = \frac{iz(\mathbf{B})(n-k)}{iz(\mathbf{W})(k-1)}, \quad k = 2, 3, \dots \quad (6.2)$$

Eşitlikte yer alan $\mathbf{B}$ kümeler arası çarpımlar ve kareler matrisini, $\mathbf{W}$ kümeler içi çarpımlar ve kareler matrisini ve k küme sayısını göstermektedir. Bu yaklaşıma göre $CH(k)$ indeksini en büyük yapan k değeri, uygun küme sayısı olarak seçilmektedir.

6.1.3. Üst Kuyruk Yöntemi

Mojena (1977) yaptığı çalışmada aşamalı kümeleme tekniklerinde birleşme uzaklıkları $\alpha_0, \alpha_1, \dots, \alpha_{n-1}$, birleşme uzaklıklarının ortalama ve standart sapması sırasıyla $\bar{\alpha}$ ve s_α olmak üzere,

$$\alpha_j > \bar{\alpha} + cs_\alpha, \quad j = 0, 1, \dots, n-1 \quad (6.3)$$

eşitsizliğini sağlayan durum sayısını uygun küme sayısını olarak önermiştir. Mojena'nın yöntemi, üst kuyruk yöntemi (upper tail method) olarak isimlendirilmektedir. Eşitsizlik (6.3)'de yer alan c sabiti için Mojena (1977), 2.75 ve 3.50 aralığını önermiştir. Fakat Milligan ve Cooper (1985) gerçekleştirdikleri simülasyon çalışması sonucuna dayanarak bu değeri 1.25 olarak önermişlerdir.

6.1.4. Silhouette İstatistiği

Kaufman ve Rousseeuw (1990) veri seti içerisindeki küme sayısını tahmin etmek için birimin kendi kümesi içerisindeki diğer birimlere olan uzaklığı ve diğer kümelerdeki birimlere olan uzaklığına dayanan Silhouette istatistiğini önermişlerdir. $a(i)$; i . birimin kendi kümesindeki diğer birimlere olan ortalama uzaklığını ve $b(i)$; i . birimin diğer kümelerdeki birimlere olan ortalama uzaklığını göstermek üzere i . birim için Silhouette indeksi

$$silh(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (6.4)$$

eşitliği ile hesaplanır. Eğer $silh(i)$ değeri 1'e yaklaşırsa i . birimin atandığı kümeye daha uygun olduğu, $silh(i)$ değeri 0'a yaklaşırsa veya negatif olursa i . birimin atandığı kümeye uygun olmadığı sonucuna varılır. Küme sayısının tahmininde ortalama Silhouette indeksi göz önünde bulundurulur.

$$\overline{silh}(k) = \frac{1}{n} \sum_{i=1}^n silh(i), \quad k = 1, 2, \dots \quad (6.5)$$

Eşitlikte yer alan n toplam birim sayısını göstermektedir. Kaufman ve Rousseeuw (1990), ortalama Silhouette indeksini maksimum yapan k değerini uygun küme sayısı olarak önermektedir.

6.1.5. Gap İstatistiği

Tibshirani ve ark (2001), bir veri kümesinde küme sayısını tahmin etmek için Gap istatistik yöntemini önermişlerdir. Veriyi kümelemede kullanılan tüm yöntemlerde uygulanabilir olan Gap istatistik yönteminde verinin yapısına uygun bir referans sıfır dağılımı oluşturulur. Bu referans dağılımından orijinal veri ile aynı hacimli rasgele örneklem üretilir. Bu rasgele örnekleme orijinal veriye uygulanan

kümeleme yöntemi uygulanarak orijinal verinin küme içi saçılımı ile rasgele örneklemden elde edilen kümeleme yapısındaki küme içi saçılımlar karşılaştırılır. Gap istatistik değerini maksimum yapan k sayısı uygun küme sayısı olarak alınır. Varsayalım ki n birimden oluşan veri, küme sayısı k olmak üzere $G_1, G_2, \dots, G_k$ şeklinde kümelere ayrılsın. Gap istatistiğinde gözlemlerin bulundukları küme merkezlerine olan uzaklıklarının toplamı W_k ,

$$W_k = \sum_{k=1}^K \frac{1}{2n_k} D_k \quad (6.6)$$

şeklinde ifade edilir. Eşitlikte yer alan n_k , k . kümede yer alan birim sayısını ve D_k , k . kümedeki gözlem çiftleri arasındaki uzaklıkların toplamını göstermektedir. Aynı kümede yer alan gözlem çiftleri arasındaki uzaklıklar toplamı D_k ,

$$D_k = \sum_{\mathbf{x}_i, \mathbf{x}_j \in G_k} d(\mathbf{x}_i, \mathbf{x}_j) \quad (6.7)$$

eşitliği ile hesaplanır. Gap istatistik yönteminin adımları aşağıdaki gibi özetlenebilir.

1. n gözlem içeren veriye bir kümeleme yöntemi, küme sayısı k ($k = 1, 2, \dots, K$) alınarak kümeler elde edilir.
2. Birinci adımda alınan k küme sayısına göre $\log(W_k)$ değeri hesaplanır.
3. n hacimlik bir rasgele örneklem uygun bir yöntemle üretilir. Bu rasgele örneklem $\mathbf{Y}$ ile gösterilir.
4. Birinci adımda hangi kümeleme yöntemi uygulanmışsa $\mathbf{Y}$ rasgele örnekleme için uygulanılır ve kümeler elde edilir.
5. Bu örneklem içinde eşitlik (6.6) ile ifade edilen W_k değeri hesaplanır ve doğal logaritması alınır.
6. $b = 1, 2, \dots, B$ olmak üzere 3-5 adımları B defa tekrar edilir. $\log(W_{k,b}^{\mathbf{Y}})$ gösterimi, $\mathbf{Y}$ rasgele örnekleme ait b . tekrarda elde edilen logaritmik

W_k değerini göstermek üzere, bu şekilde B sayıda $\log(W_{k,b}^Y)$ değeri hesaplanır.

7. Altıncı adımda hesaplanan B sayıda $\log(W_{k,b}^s)$ değerinin ortalaması ve standart sapması sırasıyla,

$$\bar{W}_k^Y = \frac{1}{B} \sum_{b=1}^B \log(W_{k,b}^Y) \quad (6.8)$$

$$sd_k = \left(\frac{1}{B} \sum_{b=1}^B [\log(W_{k,b}^Y) - \bar{W}_k^Y]^2 \right)^{\frac{1}{2}} \quad (6.9)$$

eşitlikleri ile hesaplanır.

8. Tahmin edilmiş Gap istatistiği,

$$gap(k) = \bar{W}_k^Y - \log(W_k) \quad (6.10)$$

eşitliği ile elde edilir.

9. Uygun küme sayısı,

$$gap(k) \geq gap(k+1) - s_{k+1} \quad (6.11)$$

koşulunu sağlayan en küçük k değeridir. Eşitlikte yer alan s_{k+1} ,

$$s_{k+1} = sd_{k+1} \left(1 + \frac{1}{B} \right)^{\frac{1}{2}} \quad (6.12)$$

eşitliği ile hesaplanır.

6.2. Küme Sayısı Tahmini İçin Önerilen Yöntem

Kümeleme analizinde uygun küme sayısının belirlenmesi için önerdiğimiz yöntem $1 \times p$ boyutlu gözlem vektörünün 1×1 boyutunda tek değerle temsil edilmesi esasına dayanmaktadır. Farz edelim ki $n \times p$ boyutlu ham veri matrisi $\mathbf{X}$,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} = [\mathbf{X}_1 \quad \mathbf{X}_2 \quad \cdots \quad \mathbf{X}_j \quad \cdots \quad \mathbf{X}_p] \quad (6.13)$$

şeklinde verilsin. Veri matrisindeki x_{ij} gösteriminde i birimleri, j değişkeni göstermektedir. Örneklem ortalama vektörü $\bar{\mathbf{x}}$,

$$\bar{\mathbf{x}} = \frac{1}{n} \mathbf{X}' \mathbf{j} \quad (6.14)$$

eşitliği ile tanımlanır. Eşitlikte yer alan $\mathbf{j}$, $n \times 1$ boyutlu bir vektördür. Örneklem kovaryans matrisi $\mathbf{S}$,

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \quad (6.15)$$

formundadır. Örneklem kovaryans matrisinin esas köşegen elemanları değişkenlere ait varyansı gösterirken, diğer elemanlar değişkenler arasındaki kovaryansı göstermektedir ve $s_{ij} = s_{ji}$ olduğundan örneklem kovaryans matrisi simetrik bir matristir. Örneklem kovaryans matrisi $\mathbf{S}$,

$$\mathbf{S} = \text{cov}(\mathbf{X}) = \frac{1}{n-1} \mathbf{X}'(\mathbf{I} - \frac{1}{n} \mathbf{J})\mathbf{X} \quad (6.16)$$

eşitliği ile hesaplanır. Eşitlikte yer alan $\mathbf{I}$, $n \times n$ boyutlu birim matrisi, $\mathbf{J}$ ise $n \times n$ boyutlu bir matrisini göstermektedir. Çokdeğişkeli istatistiksel analizlerde, örneklem kovaryans matrisinin determinantı $|\mathbf{S}|$ **genelleştirilmiş varyans** olarak isimlendirilirken, $\mathbf{S}$ matrisinin esas köşegen üzerindeki elemanlarının toplamı **toplam örneklem varyansı** olarak isimlendirilir. Çok değişkenli veri matrislerinde genellikle değişkenler ilişkilidir yani $s_{ij} \neq 0$. Bundan dolayı her bir değişkenin toplam değişime olan yalın katkısı belirlenemez. Dolayısıyla $p \times 1$ boyutlu gözlem vektörünü tek bir temsil değerine dönüştürmede değişken değerlerini doğrudan toplamak uygun olmaz. Gözlem vektörünü tek bir temsil değerine dönüştürürken orijinal değişkenler yerine birbiri ile ilişkisiz olan ve orijinal değişkenlerin yerine kullanılabilecek yeni bileşenler kullanılmalıdır. Bu amaçla boyut indirgemek ve bağımlılık yapısını yok etmek amaçlı kullanılan temel bileşenler analizinden yararlanılabilir. Ancak burada temel bileşenler analizi sadece bağımlılık yapısını yok etmek amacı ile kullanılacaktır.

Varsayalım ki $\mathbf{S}$ örneklem kovaryans matrisinin özdeğer-özvektör çiftleri $\lambda_1 > \lambda_2 > \dots > \lambda_p > 0$ koşulunda $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \dots, (\lambda_p, \mathbf{e}_p)$ şeklinde verilsin. Orijinal değişkenler yerine kullanılacak olan yeni bileşenler,

$$\begin{aligned} \mathbf{Y}_1 &= \mathbf{X}\mathbf{e}_1 = \mathbf{X}_1 e_{11} + \mathbf{X}_2 e_{12} + \dots + \mathbf{X}_p e_{1p} \\ \mathbf{Y}_2 &= \mathbf{X}\mathbf{e}_2 = \mathbf{X}_1 e_{21} + \mathbf{X}_2 e_{22} + \dots + \mathbf{X}_p e_{2p} \\ &\vdots \\ \mathbf{Y}_p &= \mathbf{X}\mathbf{e}_p = \mathbf{X}_1 e_{p1} + \mathbf{X}_2 e_{p2} + \dots + \mathbf{X}_p e_{pp} \end{aligned} \quad (6.17)$$

$$\text{veya } \mathbf{E} = \begin{bmatrix} e_{11} & e_{21} & \dots & e_{p1} \\ e_{12} & e_{22} & \dots & e_{p2} \\ \vdots & \vdots & & \vdots \\ e_{1p} & e_{2p} & \dots & e_{pp} \end{bmatrix} \text{ ve } \mathbf{Y} = [\mathbf{Y}_1 \quad \mathbf{Y}_2 \quad \dots \quad \mathbf{Y}_p] \text{ olmak üzere}$$

$$\mathbf{Y} = \mathbf{X}\mathbf{E} \quad (6.18)$$

eşitlikleri ile elde edilir. Eğer özdeğerler $\lambda_1, \lambda_2, \dots, \lambda_p$ birbirinden farklı ise özvektörler bağımsızdır. Eğer örneklem kovaryans matrisinin özdeğerlerinden aynı olanlar varsa bu özdeğerlere karşı ortak bir özvektör kullanılır. Böylece $l \neq k$ olmak üzere $\mathbf{e}_l$ ve $\mathbf{e}_k$ gibi herhangi iki özvektör için $\mathbf{e}_l' \mathbf{e}_k = 0$ geçerli olacaktır. Bu durumda iki bileşen arasındaki kovaryans $\mathbf{S}\mathbf{e}_k = \lambda_k \mathbf{e}_k$ olduğu için,

$$\text{cov}(\mathbf{Y}_l, \mathbf{Y}_k) = \mathbf{e}_l' \mathbf{S} \mathbf{e}_k = \mathbf{e}_l' \lambda_k \mathbf{e}_k = \lambda_k \mathbf{e}_l' \mathbf{e}_k = 0 \quad (6.19)$$

olur ve böylelikle yeni bileşenlerin birbirinden bağımsız oldukları gösterilmiş olur (Johnson ve Wichern 2007). Yeni bileşenlerin varyansları $\mathbf{S}$ örneklem kovaryans matrisinin özdeğerlerine karşılık gelir. Bu durumda yeni bileşenlerin örneklem kovaryans matrisi,

$$\text{cov}(\mathbf{Y}) = \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix} \quad (6.20)$$

şeklindedir. Yeni bileşenlerin toplam örneklem varyansı ve genelleştirilmiş varyansı , orijinal değişkenlerin toplam örneklem varyansı ve genelleştirilmiş varyansına eşittir.

$$\text{Toplam Örneklem Varyansı} = \text{tr}(\mathbf{S}) = \text{tr}(\mathbf{\Lambda}) = \sum_{r=1}^p \lambda_r \quad (6.21)$$

$$\text{Genelleştirilmiş Varyans} = |\mathbf{S}| = |\mathbf{\Lambda}| = \prod_{r=1}^p \lambda_r \quad (6.22)$$

Herhangi bir bileşenin toplam değişime olan katkısını gösteren varyans açıklama oranı,

$$w_r = \frac{\lambda_r}{tr(\Lambda)}, \quad r = 1, 2, \dots, p \quad (6.23)$$

eşitliği ile elde edilir. Bileşenlere ait varyans açıklama oranlarının ağırlık olarak kullanılmasıyla temsil vektörü $\mathbf{SV}$,

$$\mathbf{SV} = \mathbf{Y}\mathbf{w}' \quad (6.24)$$

şeklinde elde edilir. Eşitlikte yer alan $\mathbf{w}$, bileşenlerin varyans açıklama oranlarından oluşan vektörü gösterir ve $\mathbf{w} = [w_1 \ w_2 \ \dots \ w_p]$ şeklindedir. Temsil vektörü $\mathbf{SV}$ $i = 1, 2, \dots, n$ olmak üzere sv_i değerlerinden oluşur. Temsil vektörünün elemanı sv_i , $p \times 1$ boyutlu i . gözlem vektörünü temsil eden temsil değeridir. Temsil değerlerinin varyansı,

$$\begin{aligned} \sigma_{sv}^2 &= \text{cov}(\mathbf{SV}) = \text{cov}(\mathbf{Y}\mathbf{w}') = \mathbf{w} \text{cov}(\mathbf{Y})\mathbf{w}' = \mathbf{w}\Lambda\mathbf{w}' \\ \sigma_{sv}^2 &= tr(\Lambda)^{-2} \sum_{i=1}^p \lambda_i^3 \end{aligned} \quad (6.25)$$

şeklinde elde edilir.

Ham veri matrisi $\mathbf{X}$ 'in $n \times 1$ boyutlu temsil vektörüne dönüştürülmesinden sonra, temsil değerleri küçükten büyüğe doğru sıralanır. Küçükten büyüğe doğru sıralanan temsil değerleri $sv_{(i)}$ gösterimi ile gösterildiğinde,

$$sv_{(i+1)} > sv_{(i)} \quad (6.26)$$

olacaktır. Temsil değerlerine ait frekanslar f_i ile gösterildiğinde

$$f_i = 1, \quad i = 1, \dots, n \quad (6.27)$$

şeklinde ifade edilir. Küme sayısı tahmini için önerdiğimiz yöntemde, ardışık iki temsil değeri arasındaki farklar,

$$d_i = sv_{(i+1)} - sv_{(i)}, \quad i = 1, \dots, n-1 \quad (6.28)$$

eşitliği ile elde edilir. Küme sayısı tahmini için önerdiğimiz yöntemde, elde edilen farklardan hangilerinin önemli olduğuna karar verebilmek için ilgilenilen iki küme merkezi arasındaki beklenen minimum fark EMD ,

$$EMD = \gamma \sigma_{sv} n^{-0.25} \quad (6.29)$$

eşitliği ile tanımlanmıştır. Eşitlikte yer alan γ , ($\gamma > 0$) sabit bir değerdir ve bu sabit değer iç içe kümeleme olmaması varsayımı altında $\gamma = 3.0$ alınabilir. İç içe kümeleme kavramı ile veri setinin kümelendirilmesi sonrası elde edilen kümelerin kendi içerisinde tekrar parçalanması ifade edilmektedir. “Hasta” ve “Hasta değil” diye iki kümeye ayrılan kitlede “Hasta” kümesinin kendi içerisinde “Ağır hasta” ve “Hafif hasta” olarak ikiye parçalanması ile küme sayısının üç olması, iç içe kümeleme durumuna örnek olarak verilebilir. Uygulamalarımızda kümelerin kendi içerisinde bir kere parçalandığı veri setleri için $\gamma = 2.5$, kümelerin kendi içerisinde birden fazla kez parçalandığı veri setlerinde ise $\gamma = 1.2$ alındığında gerçek küme sayıları doğru olarak tahmin edilmiştir.

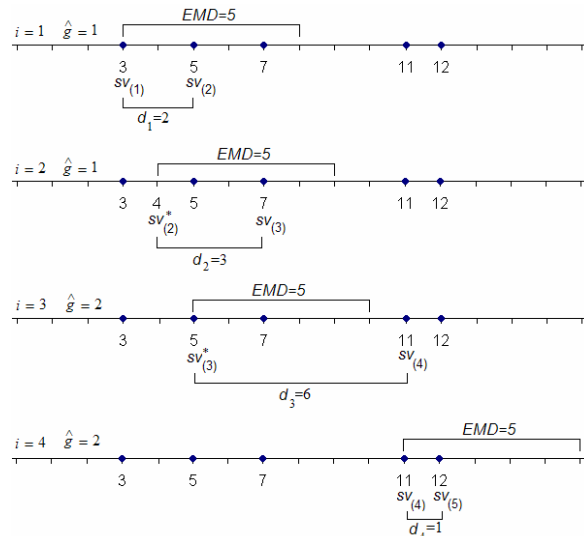
Küme sayısı tahmini için önerdiğimiz algoritmanın başlangıcında küme sayısı tahminimiz 1’dir. Önerdiğimiz algoritmada, ilk ardışık farktan başlanarak d_i değerlerinin önemli olup olmadığı incelenmelidir. Algoritmada ardışık iki temsil değeri arasındaki fark önemli bulunduğunda yani $d_i > EMD$ ise küme sayısı

tahminimiz 1 arttırılır. Eğer ardışık iki temsil değeri arasındaki fark önemsiz bulunursa yani $d_i \leq EMD$ ise $sv_{(i+1)}$ ve f_{i+1} değerleri sırasıyla,

$$sv_{(i+1)} = \frac{f_i sv_{(i)} + f_{i+1} sv_{(i+1)}}{f_i + f_{i+1}} \quad (6.30)$$

$$f_{i+1} = f_i + f_{i+1} \quad (6.31)$$

eşitlikleri ile güncellenir ve bir sonraki ardışık fark güncellenen değerlere göre belirlenir. Eşitlik (6.30) ve (6.31) bilgisayar algoritma dilindedir. Buna göre (6.30) nolu eşitliğin solundaki $sv_{(i+1)}$ değeri yeni değer iken (6.30) nolu eşitliğin sağındaki $sv_{(i+1)}$ değeri eski değerdir. Eşitlik (6.31) içinde tanımlama benzer şekildedir. Ardışık farkların incelenmesi tamamlandığında oluşan küme sayısı, $\mathbf{X}$ veri seti için önereceğimiz uygun küme sayısıdır. Küme sayısı tahmini için önerdiğimiz bu yöntem, **ardışık minimum fark yöntemi** olarak isimlendirilebilir. Önerilen yöntemi bir örnek ile inceleyelim. Varsayalım ki $EMD = 5$ ve n birimden oluşan veri seti için küçükten büyüğe sıralanmış ilk beş temsili değer $sv_{(1)} = 3$, $sv_{(2)} = 5$, $sv_{(3)} = 7$, $sv_{(4)} = 11$ ve $sv_{(5)} = 12$ olsun. Önerilen yönteme göre ilk beş temsili değer incelenmesi Şekil 6.1’de verilmiştir.



Şekil 6.1. Ardışık minimum fark yöntemi ile ilk beş temsil değerinin incelenmesi

Şekil 6.1’de önerilen yöntemle göre ilk adımda ($i=1$), ilk iki temsil değeri arasındaki ardışık fark $d_1 = sv_{(2)} - sv_{(1)} = 2$, $EMD = 5$ ’den küçük bulunmuştur. İkinci adımda ($i=2$), öncelikle yöntem gereği ikinci temsil değeri güncellenerek $sv_{(2)}^* = 4$ bulunmuştur. Güncellenen ikinci temsil değeri ile üçüncü temsil değeri arasındaki fark $d_2 = 3$, $EMD = 5$ ’den küçük bulunmuş ve küme sayısı tahminimiz 1 ($\hat{g}=1$) değişmemiştir. Üçüncü adımda da ($i=3$), öncelikle yöntem gereği temsil değeri güncellenerek $sv_{(3)}^* = 5$ bulunmuştur. Güncellenen üçüncü temsil değeri ile dördüncü temsil değeri arasındaki fark $d_3 = 6$, $EMD = 5$ ’den büyük bulunmuş ve küme sayısı tahminimiz 1 artırılarak 2 ($\hat{g}=1+1=2$) olmuştur. Üçüncü adımda $d_3 > EMD$ olduğundan dördüncü adımda ($i=4$), $sv_{(4)}$ güncellenmemiş ve dördüncü temsil değeri ile beşinci temsil değeri arasındaki ardışık fark $d_4 = sv_{(5)} - sv_{(4)} = 1$, $EMD = 5$ ’den küçük bulunmuştur. Şekil 6.1’e göre ilk beş temsil değerinin incelenmesi sonrası küme sayısı tahminimiz 2 ($\hat{g}=2$) olmuştur. Matlab 7 paket programında ardışık farklar yöntemi için yazılmış kod Şekil 6.2’de verilmiştir.

```
clear all
clc
load data.txt;%İlgili veri dosyasının yüklenmesi
x=data;
kov=cov(x); %X'in kovaryans matrisinin elde edilmesi
[n p]=size(x);
[ozvektor,lambda,w] = pcacov(kov);%X'in kovaryans matrisi ile
                                Temel bileşenler analizi
y=x*ozvektor;%Temel bileşen skorları
w=w./100;%Varyans açıklama oranı
sv=y*w;%Temsil değerleri
sv=sort(sv);%Temsil değerlerinin sıralanması
c=3;%Özel değerin belirlenmesi
EMD=(c*std(sv)*n^(-0.25));%Beklenen minimum fark
f=ones(n,1); % Sıralanmış temsil değerlerine ait frekanslar
g=1;%Başlangıçta küme sayısı
for i=1:n-1
    d(i)=sv(i+1)-sv(i);
    if d(i)>EMD
        g=g+1;
    else
        sv(i+1)=(f(i)*sv(i)+f(i+1)*sv(i+1))/(f(i)+f(i+1));
        f(i+1)=f(i)+f(i+1);
    end
end
kumesayisi=g
```

Şekil 6.2. Ardışık farklar yöntemi ile küme sayısı tahmini için yazılmış Matlab kodu

6.3. Önerilen Yöntem ile Küme Sayısı Tahminleri

Küme sayısının tahmini için önerdiğimiz ardışık farklar yönteminin etkinliğini, kümeleme analizi literatüründe yaygın olarak kullanılan ve küme sayıları bilinen gerçek veri setleri üzerinde göstereceğiz. Ardışık farklar yöntemi ile küme sayısı tahmini; Diyabet (Diabetes), E.coli, Gayzer (Geyser), Cam (Glass) İris Çiçeği (Iris), Araç Görüntüsü (Vehicle Silhouettes), Şarap (Wine) ve Maya (Yeast) veri setleri için yapılacaktır. İncelenen veri setlerinin örneklem büyüklükleri (n), değişken sayıları (p) ve küme sayıları (k) Tablo 6.1’de verilmiştir.

Tablo 6.1. İncelenen veri setlerinin örneklem büyüklükleri, değişken sayıları ve küme sayıları

Veri Setleri	Örneklem Büyüklüğü (n)	Değişken Sayısı (p)	Küme Sayısı (k)
Diyabet*	145	4	3
E.coli**	336	7	8
Gayzer	272	2	3
Cam**	214	9	6
İris	150	4	3
Araç	846	18	4
Şarap	178	13	3
Maya*	1484	8	10

*: İlgili veri setleri için verilen küme sayıları, kümelerin kendi içerisinde bir kere parçalanması ile elde edilmiş küme sayılarıdır. İç içe kümelemenin olmadığı durumda Diyabet verisi 2, Maya verisi 8 kümeden oluşmaktadır.

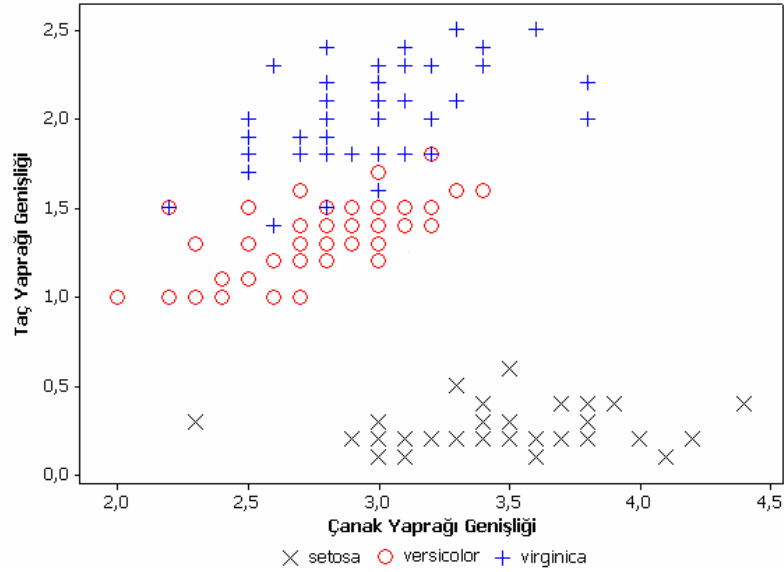
** : İlgili veri setleri için verilen küme sayıları, kümelerin kendi içerisinde birden fazla kez parçalanması ile elde edilmiş küme sayılarıdır. İç içe kümelemenin olmadığı durumda E.coli verisi 4, Cam verisi 2 kümeden oluşmaktadır.

İncelenen veri setleri için iç içe kümelemenin olmadığı varsayımı altında gerçek küme sayıları (k), kovaryans matrislerinden elde edilen özdeğerler, temsil değerlerine ait standart sapmalar (σ_{sv}), özel değer $\gamma = 3.0$ alınarak iki küme merkezi arasında beklenen minimum farklar (EMD) ve iç içe kümelemenin olmadığı varsayımı altında ardışık minimum fark yöntemi ile gerçekleştirilen küme sayısı tahminleri ($\hat{k}$) Tablo 6.2’de verilmiştir.

Tablo 6.2. Veri setleri için iç içe kümelemenin olmadığı varsayımı altında gerçek küme sayıları(k),özdeğerler, temsil değerlerinin standart sapmaları(σ_{sv}), iç içe kümelemenin olmadığı varsayımı altında özel değer $\gamma = 3.0$ alınarak elde edilen beklenen minimum farklar (EMD) ve ardışık minimum fark yöntemi ile gerçekleştirilen küme sayısı tahminleri ($\hat{k}$)

Veri Setleri	k	Özdeğerler				σ_{sv}	EMD	$\hat{k}$
Diyabet	2	112890	14187	3085.5	250.32	291.1371	251.6963	2
E.coli	4	0.089703 0.0085336	0.042439 0.0049353	0.014633 0.00065892	0.012884	0.1632	0.1143	4
Gayzer	3	185.88	0.24422			13.616	10.0584	3
Cam	2	3.0020000 0.2084600 0.0000029	1.6592000 0.1011600	0.6795800 0.0089986	0.6431700 0.0014774	0.9003	0.7062	2
İris	3	4.2282	0.24267	0.07821	0.023835	1.9015	1.63	2
Araç	4	34900 72.384 23.577 4.5211 0.23979	631.97 47.506 11.938 2.3868 0.071225	416.88 41.453 9.3676 1.5676	130.13 26.758 5.8297 0.69127	179.477	99.836	4
Şarap	3	99202 1.2288 0.1121 0.0082037	172.54 0.84107 0.071703	9.4381 0.27897 0.037575	4.9912 0.15138 0.021066	314.362	258.1941	3
Maya	8	0.029356 0.0060038	0.017754 0.0057166	0.011187 0.0030834	0.0078507 0.0023097	0.0692	0.0335	8

Tablo 6.2 incelendiğinde İris verisi dışındaki tüm veri setleri için iç içe kümeleme olmadığı varsayımı altında $\gamma = 3.0$ alınarak gerçekleştirilen küme sayısı tahminlerinin doğru olduğu görülmektedir. İris verisinin kullanıldığı çok sayıda çalışmada küme sayısının 2 olarak alındığı görülebilir(Kothari ve Pitts 1999, Sun ve ark 2001, Sugar ve James 2003, Mufti ve ark 2005, Yan ve Ye 2007). İris verisinin küme sayısı ile ilgili olan bu farklılığın nedenini görmek için İris verisinin taç yaprağı genişliği ve çanak yaprağı genişliği değişkenleri bakımından elde edilen saçılım grafiği Şekil 6.3’de verilmiştir.



Şekil 6.3. İris verisinin taç yaprağı genişliği ve çanak yaprağı genişliğine göre saçılımı

İris verisinin taç yaprağı genişliği ve çanak yaprağı genişliğine göre saçılımı incelendiğinde “Setosa” türünün diğer iki türden iyi bir şekilde ayrıldığı fakat “Versicolor” ve “Virginica” türlerinin iç içe olduğu görülmektedir.

İris, Diyabet ve Maya veri setlerinde kümelerin kendi içerisinde bir kere parçalanması ile oluşan küme sayıları, kovaryans matrislerinden elde edilen özdeğerler, temsil değerlerine ait standart sapmalar (σ_{sv}), özel değer $\gamma = 2.5$ alınarak elde edilen iki küme merkezi arasında beklenen minimum farklar (EMD) ve ardışık minimum fark yöntemi ile elde edilen küme sayısı tahminleri Tablo 6.3’de verilmiştir.

Tablo 6.3. Veri setleri için küme sayıları (k), özdeğerler, temsil değerlerinin standart sapmaları (σ_{sv}), $\gamma = 2.5$ alınarak elde edilen beklenen minimum farklar (EMD) ve ardışık minimum fark yöntemi ile gerçekleştirilen küme sayısı tahminleri ($\hat{k}$)

Veri Setleri	k	Özdeğerler				σ_{sv}	EMD	$\hat{k}$
Diyabet	3	112890	14187	3085.5	250.32	291.1371	209.7469	3
Iris	3	4.2282	0.24267	0.07821	0.023835	1.9015	1.3583	3
Maya	10	0.029356 0.0060038	0.017754 0.0057166	0.011187 0.0030834	0.0078507 0.0023097	0.0692	0.0279	10

Tablo 6.3 incelendiğinde iki küme merkezi arasında beklenen minimum farkın belirlenmesinde $\gamma = 2.5$ alındığında, İris ve kümelerin kendi içerisinde bir kere parçalanması ile küme sayıları sırasıyla 3 ve 10 olan Diyabet ve Maya veri setleri için gerçekleştirilen küme sayısı tahminleri doğru olarak elde edilmiştir.

E.coli ve Cam veri setlerinde kümelerin kendi içerisinde birden fazla kez parçalanması ile oluşan küme sayıları, kovaryans matrislerinden elde edilen özdeğerler, temsil değerlerine ait standart sapmalar (σ_{sv}), özel değer $\gamma = 1.2$ alınarak elde edilen iki küme merkezi arasında beklenen minimum farklar (EMD) ve ardışık minimum fark yöntemi ile elde edilen küme sayısı tahminleri Tablo 6.4’de verilmiştir.

Tablo 6.4. Veri setleri için küme sayıları (k), özdeğerler, temsil değerlerinin standart sapmaları (σ_{sv}), $\gamma = 1.2$ alınarak elde edilen beklenen minimum farklar (EMD) ve ardışık minimum fark yöntemi ile gerçekleştirilen küme sayısı tahminleri ($\hat{k}$)

Veri Setleri	k	Özdeğerler				σ_{sv}	EMD	$\hat{k}$
E.coli	8	0.089703	0.042439	0.014633	0.012884	0.1632	0.0457	8
		0.0085336	0.0049353	0.00065892				
Cam	6	3.0020000	1.6592000	0.6795800	0.6431700	0.9003	0.2825	6
		0.2084600	0.1011600	0.0089986	0.0014774			
		0.0000029						

Tablo 6.4 incelendiğinde iki küme merkezi arasında beklenen minimum farkın belirlenmesinde $\gamma = 1.2$ alındığında, kümelerin kendi içerisinde birden fazla kez parçalanması ile oluşan küme sayıları sırasıyla 8 ve 6 olan E.coli ve Cam veri setleri için gerçekleştirilen küme sayısı tahminleri doğru olarak elde edilmiştir.

6.4. Önerilen Yöntemin Diğer Yöntemlerle Karşılaştırılması

Ardışık minimum fark yöntemi ile küme sayısı tahminleri yaptığımız veri setleri için Marriott indeksi, varyans açıklama oran kriteri, üst kuyruk yöntemi, Silhouette indeksi ve Gap istatistiği ile küme sayısı tahminleri yapılacaktır. Üst kuyruk yöntemi dışındaki diğer yöntemler deneme yanılmaya dayalı yöntemlerdir.

Bu yöntemlerle küme sayısı tahminleri yapmak için $k = 1, 2, \dots, 20$ alınarak kümeleme yapılacaktır. Ardışık minimum farklar yöntemi ile karşılaştıracağımız yöntemler ile edilecek küme sayısı tahminleri, kümelemede kullanılacak olan kümeleme tekniği ve uzaklık ölçüsüne bağlıdır. Yani aynı veri seti için farklı kümeleme algoritması veya farklı uzaklık ölçüsü kullanılması durumunda farklı küme sayısı tahminleri vereceklerdir. Yöntemlerin bu dezavantajını gösterebilmek için ilgili yöntemlerle aynı uzaklık ölçüsü ile farklı kümeleme algoritmalarında ve aynı kümeleme algoritmasında farklı uzaklık ölçüsünün kullanıldığı durumlarda küme sayısı tahminleri gerçekleştirilecektir.

Mojena (1977) tarafından aşamalı kümeleme algoritmalarında uygun küme sayısını belirlemesi için önerilen üst kuyruk yöntemiyle, incelenen veri setlerinde Öklid uzaklığının kullanıldığı “Tek bağlantı” ve “Ward” kümeleme algoritmalarında farklı c sabiti değerlerine göre elde edilen küme sayısı tahminleri Tablo 6.5’de verilmiştir.

Tablo 6.5. İncelenen veri setlerinde Öklid uzaklığının kullanıldığı “Tek bağlantı” ve “Ward” kümeleme algoritmalarında farklı c sabiti değerleri için üst kuyruk yöntemi ile elde edilen küme sayısı tahminleri

Veri Setleri (Küme Sayısı)	Kümeleme Algoritması	Uzaklık Ölçüsü	Küme Sayısı Tahminleri		
			$c = 1.25$	$c = 2.75$	$c = 3.5$
Diyabet (3/2)	Tek bağlantı	Öklid	15	3	1
	Ward	Öklid	4	3	1
E.coli (8/4)	Tek bağlantı	Öklid	22	8	6
	Ward	Öklid	12	5	4
Gayzer (3)	Tek bağlantı	Öklid	54	5	3
	Ward	Öklid	6	4	2
Cam (6/2)	Tek bağlantı	Öklid	19	6	3
	Ward	Öklid	11	5	4
İris (3)	Tek bağlantı	Öklid	11	2	1
	Ward	Öklid	4	2	2
Şarap (3)	Tek bağlantı	Öklid	11	5	2
	Ward	Öklid	6	3	2
Araç (4)	Tek bağlantı	Öklid	38	13	9
	Ward	Öklid	11	5	5
Maya (10/8)	Tek bağlantı	Öklid	119	33	20
	Ward	Öklid	56	22	17

Tablo 6.5 incelendiğinde üst kuyruk tahmini ile aynı veri seti için farklı kümeleme algoritmalarında farklı küme sayısı tahminlerinin yapıldığını ve gerçekleşen küme sayısı tahminlerinin de genelde kötü olduğu görülmektedir. Aynı kümeleme algoritmasında farklı uzaklık ölçülerinin kullanılması durumunda ve farklı kümeleme algoritmalarında aynı uzaklık ölçüsünün kullanılmasıyla elde edilen kümelemeler için Marriott indeksi, varyans açıklama oranı ve Silhouette indeksine göre gerçekleştirilen küme sayısı tahminleri Tablo 6.6’da verilmiştir.

Tablo 6.6. Marriott indeksi, varyans açıklama oranı ve Silhouette indeksine göre gerçekleştirilen küme sayısı tahminleri

Veri Seti (Küme Sayısı)	Kümeleme Algoritması	Uzaklık Ölçüsü	Marriott İndeksine Göre	Varyans Açıklama Oranına Göre	Silhouette İndeksine Göre
Diyabet (3/2)	<i>k</i> -Ortalamlar	Karesel Öklid	18	2	2
	<i>k</i> -Ortalamlar	City-Block	19	2	2
	Tek Bağlantı	Öklid	4	4	3
	Tam Bağlantı	Öklid	20	2	2
E.coli (8/4)	<i>k</i> -Ortalamlar	Karesel Öklid	7	3	3
	<i>k</i> -Ortalamlar	City-Block	19	3	3
	Tek Bağlantı	Öklid	1	1	2
	Tam Bağlantı	Öklid	6	3	3
Gayzer (3)	<i>k</i> -Ortalamlar	Karesel Öklid	2	18	2
	<i>k</i> -Ortalamlar	City-Block	2	19	2
	Tek Bağlantı	Öklid	5	5	4
	Tam Bağlantı	Öklid	16	18	2
Cam (6/2)	<i>k</i> -Ortalamlar	Karesel Öklid	18	2	4
	<i>k</i> -Ortalamlar	City-Block	14	3	3
	Tek Bağlantı	Öklid	20	18	2
	Tam Bağlantı	Öklid	20	5	8
İris (3)	<i>k</i> -Ortalamlar	Karesel Öklid	19	4	2
	<i>k</i> -Ortalamlar	City-Block	19	3	2
	Tek Bağlantı	Öklid	19	2	2
	Tam Bağlantı	Öklid	20	4	4
Şarap (3)	<i>k</i> -Ortalamlar	Karesel Öklid	19	19	2
	<i>k</i> -Ortalamlar	City-Block	16	16	2
	Tek Bağlantı	Öklid	8	8	2
	Tam Bağlantı	Öklid	20	20	2
Araç (4)	<i>k</i> -Ortalamlar	Karesel Öklid	20	2	2
	<i>k</i> -Ortalamlar	City-Block	16	2	2
	Tek Bağlantı	Öklid	3	7	2
	Tam Bağlantı	Öklid	20	6	4
Maya (10/8)	<i>k</i> -Ortalamlar	Karesel Öklid	19	2	3
	<i>k</i> -Ortalamlar	City-Block	14	2	2
	Tek Bağlantı	Öklid	5	3	3
	Tam Bağlantı	Öklid	15	8	2

Tibshirani ve ark. (2001), tarafından küme sayısı tahmini için önerilen Gap yönteminin genellikle aşamalı kümeleme tekniklerinde daha iyi performans göstermektedir (Martinez ve Martinez 2005). Bu nedenden dolayı, ilgili veri setleri için Gap istatistiği ile küme sayısı tahminleri karesel Öklid uzaklığının kullanıldığı tam bağlantı ve Ward kümeleme teknikleri ile gerçekleştirilecektir. Gap istatistiğinin hesaplanmasında kullanılan referans dağılımı Uniform dağılımından üretilecektir. Her kümelemede $B = 100$ alınacaktır. Gap istatistiği ile elde edilen küme sayısı tahminleri Tablo 6.7’de verilmiştir.

Tablo 6.7. İncelenen veri setlerinde karesel Öklid uzaklığının kullanıldığı “Tam bağlantı” ve “Ward” kümeleme algoritmalarında Gap istatistiği ile elde edilen küme sayısı tahminleri

Veri Setleri (Küme Sayısı)	Kümeleme Algoritması	Uzaklık Ölçüsü	Küme Sayısı Tahminleri ($\hat{k}$)
Diyabet (3/2)	Tam bağlantı	Öklid	7
	Ward	Öklid	10
E.coli (8/4)	Tam bağlantı	Öklid	3
	Ward	Öklid	12
Gayzer (3)	Tam bağlantı	Öklid	2
	Ward	Öklid	2
Cam (6/2)	Tam bağlantı	Öklid	6
	Ward	Öklid	15
İris (3)	Tam bağlantı	Öklid	8
	Ward	Öklid	9
Şarap (3)	Tam bağlantı	Öklid	2
	Ward	Öklid	2
Araç (4)	Tam bağlantı	Öklid	1
	Ward	Öklid	15
Maya (10/8)	Tam bağlantı	Öklid	6
	Ward	Öklid	9

Tablo 6.7 incelendiğinde Gap istatistiği ile aynı veri seti için farklı kümeleme algoritmalarında farklı küme sayısı tahminlerinin yapıldığını ve gerçekleşen küme sayısı tahminlerinin de iyi olmadığı görülmektedir. Gap istatistiği ile küme sayısı tahmininde kullanılan referans dağılımının üretiminde ki farklı yaklaşımlar küme sayısı tahminini etkilemektedir. Yan ve Ye (2007) İris verisi için referans dağılımı için Uniform dağılımı ve temel bileşenleri kullanarak gerçekleştirdikleri küme sayısı tahminleri sırasıyla 6 ve 8 olmuştur.

6.5. Sonuçlar

Küme sayısı tahmini için önerdiğimiz, $p \times 1$ boyutlu gözlem vektörünün tek bir değerle temsil edilmesi esasına dayanan ardışık farklar yöntemi literatürde iyi bilinen küme sayısı tahmin yöntemlerinden birkaçı ile karşılaştırılmıştır. Karşılaştırma sonrası önerdiğimiz yöntemin küme sayısı tahmin performansının diğer yöntemlere göre daha iyi olduğu görülmüştür. Uygulamalarımızda iç içe kümelemenin olmaması durumunda iki küme merkezi arasındaki beklenen minimum farkın hesaplanmasında $\gamma = 3$, iç içe kümelemede kümelerin kendi içerisinde bir kez parçalanması durumunda $\gamma = 2.5$ ve kümelerin kendi içerisinde birden fazla kez parçalanması durumunda $\gamma = 1.2$ alındığında küme sayısı tahminleri doğru olarak yapılmıştır.

Küme sayısı tahmini için literatürde mevcut olan ve iyi bilinen yöntemlerin hesaplanmasındaki zorluklar, birçoğunun deneme yanılmaya dayalı olması ve aynı veri seti için farklı küme sayısı tahmini vermeleri bu yöntemlerin dezavantajları olarak görünmektedir. Buna karşın önerdiğimiz yöntemin uygulamadaki pratikliği, küme sayısı tahminindeki etkinliği ve deneme yanılmaya bağlı olmaması dolayısıyla aynı veri seti için farklı küme sayısı tahminlerinin elde edilmemesi yöntemin avantajları olarak ortaya çıkmaktadır.

7. KÜME BAŞLANGIÇ MERKEZLERİ İÇİN YENİ YÖNTEM

k -ortalamalar algoritması basit ve farklı veri setleri için yaygın bir kullanıma sahip olsa da küme başlangıç merkezlerine karşı duyarlıdır. Elde edilen küme merkezleri konum bakımından optimum bir sonuca yakınsayabileceği gibi optimal olmayabilir. Ayrıca bir kümeye herhangi bir birim atanmaması durumunda boş küme durumu ile de karşılaşılabilir. Bu yüzden k -ortalamalar algoritmasında iyi başlangıç merkezlerinin seçimi önemlidir.

Literatürde k -ortalamalar algoritmasında küme başlangıç merkezlerinin seçimi için önerilen çok sayıda algoritma vardır. Bu bölümde, son dönemde önerilen algoritmalarından Bradley ve Fayyad (1998), Likas ve ark (2003), Khan ve Ahmad (2004) ile Deelers ve Auwatanamongkol (2007) tarafından önerilen algoritmalar incelenecektir.

Bradley ve Fayyad (1998) veri seti içerisinde rastgele oluşturulan küçük alt örneklerin k -ortalamalar tekniği ile kümelendirmesi sonucu oluşan merkezlerden başlangıç merkezlerinin oluşturulduğu arıtma algoritmasını önerdiler. Bu başlangıç algoritmasının kullanılması durumunda verinin k -ortalamalar algoritması ile kümeleneceği sonucu oluşan kümeleme hatasının minimum olacağını gösterdiler.

Likas ve ark (2003) $k - d$ ağaç yöntemiyle veri uzayını alt uzaylara parçalayarak başlangıç merkezlerini oluşturan global(küresel veya genel) k -ortalamalar algoritmasını önermişlerdir. Global k -ortalamalar algoritmasında, temel bileşenler analizinden elde edilen en büyük varyanslı veri eksenine dik düzlemde, hiper düzlem kesmesi kullanılır. Parçalanma birimlerin her birinin önceden belirlenen merkez sayısından az olmamak koşuluyla belirlenen noktalara bağlanmasına kadar gerçekleştirilir. Parçalanma sonucu oluşan düğümlerin merkezleri k -ortalamalar algoritmasında başlangıç küme merkezleri olarak alınır.

S.S. Khan ve A.Ahmad (2004), veri setinde verilen her değişkeni göz önünde bulunduran küme merkezi başlangıç algoritmasını (Cluster Center Initialization Algorithm) önermişlerdir. Bu algoritmada iyi küme başlangıç merkezi için bazı bilgiler gereklidir. Algoritma aşağıdaki gibi tanımlanır.

- i. Her değişken için ortalama ve standart sapma hesaplanır.
- ii. $-\infty$ ile $(2s-1)/2k$, $(s=1,...,k)$ aralığında $Z_1,...,Z_k$ için normal eğri alanları altında kalan yüzdelikleri hesaplanır.
- iii. Her bir değişken için hesaplanan ortalama ve standart sapmayla elde edilen yüzdelikler kullanılarak her değişken için $j=1,...,p$ olmak üzere $x_{sj} = z_{sj}\sigma_j + \mu_j$ değeri hesaplanır.
- iv. Belirlenen x_{sj} başlangıçlarına göre j . değişken için k -ortalamlar tekniği kullanılarak her birimin küme üyelikleri tanımlanır.
- v. iii. ve iv. adımlar tüm değişkenler için tekrarlanır.
- vi. Her değişken için her birimin küme üyeliklerini gösteren $(G_{x_1} G_{x_2} ... G_{x_p})$ oluşturulur.
- vii. $(G_{x_1} G_{x_2} ... G_{x_p})$ göre k' kümeleme gerçekleştirilir. Eğer $k' > k$ DBMSDC(Çok Ölçekli Verinin İndirgenmesine Dayalı Yoğunluk) uygulanarak k kümeleme k kümelemeye dönüştürülür.
- viii. Belirlenen k küme merkezi k -ortalamlar tekniğinde başlangıç küme merkezi alınarak orijinal veri k -ortalamlar tekniğiyle kümelendirilir.

Deelters ve Auwatanamongkol (2007) çalışmalarında, k -ortalamlar algoritmasında küme başlangıç merkezlerinin seçimi için en büyük varyansa sahip değişkenin yayıldığı eksen boyunca veriyi parçalamaya dayalı bir yaklaşım önermişlerdir. En büyük varyansa sahip eksen de verinin parçalanması için dikey kesme algoritmasını kullanmışlardır. Çalışmalarında önerilen yöntemin özellikle renk ölçüm verilerinde daha başarılı olduğunu belirtmişlerdir.

Bahsedilen algoritmalar iyi başlangıç küme merkezlerinin bulunmasında yararlanılabilir olmalarına rağmen, oldukça karmaşık ve başlangıç merkezlerinin oluşturulmasında rassal başlangıçlı k -ortalamlar algoritmasına ihtiyaç duyan tekniklerdir.

7.1. Küme Başlangıç Merkezlerinin Seçimi İçin Önerilen Yeni Algoritma

Bu bölümde, k - ortalamalar algoritmasına başlarken küme merkezlerinin seçimi için yeni bir algoritma önerilmiştir. Önerilen algoritmanın temeli p değişken içerisinde iki temel eksenin belirlenmesi esasına dayanır. Eksenlerin seçimi için öncelikle her değişken için (7.1) nolu eşitlikte verilen değişim katsayısı hesaplanır.

$$vc_j = \frac{s(x_j)}{\bar{x}_j}, \quad j = 1, 2, \dots, p \quad (7.1)$$

(7.1) nolu eşitlikte yer alan $s(x_j)$, j . değişkene ait standart sapmayı ve $\bar{x}_j$, j . değişkene ait ortalamayı göstermektedir. Birinci temel eksen olarak mutlak değerce en büyük değişim katsayısına sahip değişken seçilir. Değişim katsayısı ölçü birimi ve terim büyüklükleri için invariant olduğundan dolayı değişkenliğin ölçümünde değişim katsayısının kullanılmasıyla ölçü birimi ve terim büyüklüklerinin etkisi arındırılmış olmaktadır.

Birinci temel eksenin belirlenmesinden sonra ikinci eksen için, temel eksen seçilen değişken ile diğer değişkenler arasındaki Pearson korelasyon katsayıları

$$r_{jj'} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^n (x_{ij'} - \bar{x}_{j'})^2} \quad (7.2)$$

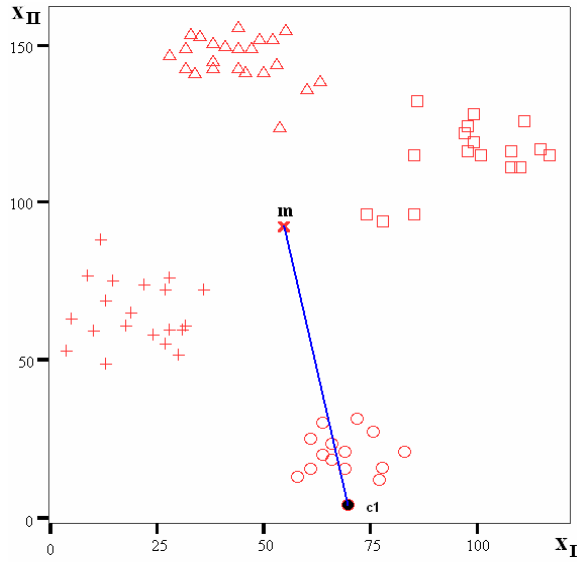
eşitliği ile hesaplanır. Temel eksen seçilen değişken ile mutlak değerce minimum korelasyona sahip değişken ikinci eksen olarak belirlenir. Temel eksen ile ikinci eksen seçilecek değişken arasındaki ilişki ne kadar küçük ise ikinci eksenin toplam değişime yaptığı yalın katkı daha fazla olacağı için ikinci eksenin seçiminde minimum korelasyon göz önünde bulundurulmuştur. Veri setini temsil edecek iki eksenin belirlenmesinden sonra bu iki değişken için merkez nokta yani ortalamalar belirlenir.

$$\mathbf{m} = [\bar{x}_I \quad \bar{x}_{II}] \quad (7.3)$$

(7.3) nolu eşitlikte yer alan $\bar{x}_I$ birinci temel eksen seçilen değişken için ortalamayı gösterirken $\bar{x}_{II}$ ikinci eksen seçilen değişken için ortalamayı göstermektedir. Seçilen iki temel eksene göre merkezin belirlenmesinden sonra her birimin merkeze olan Öklid uzaklığı (7.4) nolu eşitlik yardımıyla hesaplanır.

$$d_{im} = \sqrt{(x_{iI} - \bar{x}_I)^2 + (x_{iII} - \bar{x}_{II})^2}, \quad i = 1, 2, \dots, n \quad (7.4)$$

Birimlerin merkeze olan uzaklıkları arasından merkeze uzaklığı maksimum olan birinci aday küme başlangıç merkezi olarak belirlenir. Birinci aday küme başlangıç merkezinin seçimi örnek bir veri seti için grafiksel olarak Şekil 7.1.'de gösterilmiştir.

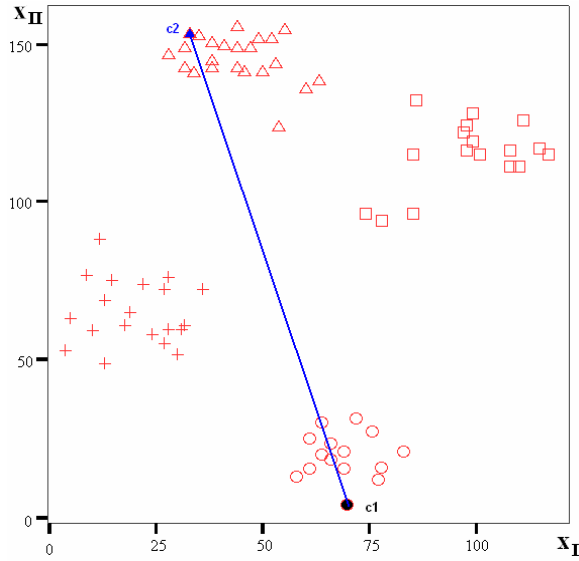


Şekil 7.1. Birinci aday küme başlangıç merkezinin seçimi

İkinci aday küme başlangıç merkezinin seçimi için bu kez tüm birimlerin birinci aday küme başlangıç merkezine olan Öklid uzaklıkları (7.5) nolu eşitlik ile hesaplanır.

$$d_{ic_1} = \sqrt{(x_{iI} - x_{c_1I})^2 + (x_{iII} - x_{c_1II})^2}, \quad i = 1, 2, \dots, n \quad (7.5)$$

Birinci aday küme merkezine en uzakta bulunan birim ikinci aday küme başlangıç merkezi olarak seçilir. İkinci aday küme başlangıç merkezinin seçiminin grafiksel olarak Şekil 7.2.'de verilmiştir.



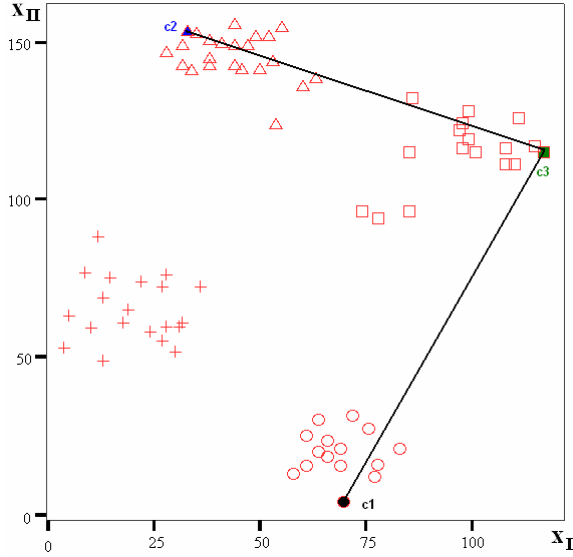
Şekil 7.2. İkinci aday küme başlangıç merkezinin seçimi

Geriye kalan başlangıç küme merkezleri belirlenirken r . iterasyonda tüm birimlerin $r-1$ tane küme merkezine olan uzaklıklarının toplamı ile ilgilenilir. Belirlenen toplam uzaklıklar içerisinde maksimum uzaklığa sahip olan birim r . iterasyonda r . aday küme başlangıç merkezi kabul edilir. Örneğin üçüncü aday küme başlangıç merkezinin seçimi için tüm birimlerin birinci ve ikinci aday küme başlangıç merkezlerine olan uzaklıkları toplanarak her birim için toplam uzaklık Sd_{i3} hesaplanır.

$$Sd_{i3} = d_{ic_1} + d_{ic_2}, \quad i = 1, 2, \dots, n \quad (7.6)$$

Hesaplanan toplam uzaklıklar arasından, önceden belirlenen iki aday küme başlangıç merkezlerine olan uzaklıkları toplamı maksimum olan birim üçüncü aday küme

başlangıç merkezi olur. Üçüncü aday küme başlangıç merkezinin belirlenmesi Şekil 7.3.'de gösterilmiştir.



Şekil 7.3. Üçüncü aday küme başlangıç merkezinin seçimi

Aday küme başlangıç merkezlerinin oluşturulması daha önceden tanımlanan küme sayısı k 'ya ulaşınca kadar devam eder. Küme başlangıç merkezleri için k tane aday merkez belirlendikten sonra birimler iki temel eksene göre kümelenilerek birimlerin küme üyelikleri oluşturulur. Elde edilen küme üyelikleri, n birimin p değişken bakımından gözlenmesi sonucu oluşan veri setinin kümelenilmesinde başlangıç küme üyeliklerini oluşturur.

7.2. Karşılaştırma Kriterleri

k -ortalamalar algoritmasında küme başlangıç merkezlerinin belirlenmesi için önerilen algoritmanın rassal başlangıçlı k -ortalamalar algoritmasına göre performansı değerlendirilirken karşılaştırma kriteri olarak hatalı sınıflandırma yüzdesi, Rand indeks değeri ve Wilks'in lambda test istatistiği kullanılacaktır. Rand indeks değeri ve Wilks'in lambda test istatistiği daha önce sırasıyla (3.15) ve (3.21) eşitlikleri ile tanımlandıkları gibi elde edilir.

Gerçek küme üyeliklerinin bilinmesi durumunda gerçekleştirilen kümeleme sonuçlarının karşılaştırılmasında hatalı sınıflandırma yüzdesi yaygın olarak kullanılan bir kriterdir. Doğru sınıflandırılan birim sayısı r ve toplam birim sayısı n olmak üzere hatalı sınıflandırma yüzdesi

$$Hata(\%) = \left(1 - \frac{r}{n}\right) \times 100 \quad (7.7)$$

eşitliği ile elde edilir. Karşılaştırılan kümeleme sonuçlarından hangisinin hatalı sınıflandırma yüzdesi düşük ise o kümeleme sonucu daha başarılı kabul edilmektedir.

7.3. Uygulama

k -ortalamalar algoritmasında küme başlangıç merkezleri için önerilen algoritmanın pratikte uygulanabilirliğini ve rassal başlangıçlı k -ortalamalar algoritması karşısındaki performansını görebilmek için İris, Şarap, Harf ve Ruspini gibi gerçek veri setlerine uygulanacaktır.

İlk defa Fisher tarafından 1936 yılında kullanılan İris veri seti kümeleme algoritmalarının performanslarının ele alındığı çalışmalarda yaygın olarak kullanılan ve sınıflama problemlerinde dünyada en çok bilinen veri setidir. Veri seti içinde İris çiçeğinin üç farklı çeşidi İris Setosa, İris Versicolor ve İris Virginica'yı temsil eden üç farklı küme mevcuttur. İris veri setinde 150 iris çiçeği dört değişken çanak yaprağı uzunluğu, çanak yaprağı genişliği, taç yaprağı uzunluğu ve taç yaprağı genişliği bakımından gözlemlenmiştir.

Şarap veri seti İtalya'da aynı bölgede yetiştirilen ancak üç farklı çeşit türetilmiş şarapların kimyasal analiz sonuçlarından oluşmaktadır. Veri seti 13 değişken bakımından gözlemlenen 178 şarabı içermektedir. Veri seti içinde 1. çeşit şaraptan 48 birim, ikinci çeşit şaraptan 59 ve üçüncü çeşit şaraptan 71 birim şarap bulunmaktadır.

Harf veri seti, İngiliz alfabesindeki 26 harfin siyah beyaz olarak 20 farklı sitede 20000 görüntüsünün 16 sayısal özelliğinden oluşmaktadır. Bu çalışmada 805

7. KÜME BAŞLANGIÇ MERKEZLERİ İÇİN YENİ YÖNTEM Murat ERİŞOĞLU

A harfi ile 789 D harfine ait veriler alınarak iki küme, 1594 birim ve 16 değişkenden oluşan bir veri seti kullanılacaktır.

Kümeleme algoritmalarının gösteriminde yaygın kullanılan bir diğer veri seti ise 1970 yılında Ruspini tarafından oluşturulan 75 birim, iki değişken ve dört kümeden oluşan Ruspini veri setidir. Bu 75 birimin 23'ü birinci kümeyi, 20'si ikinci kümeyi, 17'si üçüncü kümeyi ve geriye kalan 15'i dördüncü kümeyi oluşturmaktadır.

İris veri setindeki değişkenler için hesaplanan ortalama, standart sapma ve değişim katsayısı Tablo 7.1'de verilmiştir.

Tablo 7.1. İris verisindeki değişkenler için aritmetik ortalama, standart sapma ve değişim katsayısı değerleri

İstatistik	Çanak Yapağı Uzunluğu	Çanak Yapağı Genişliği	Taç Yapağı Uzunluğu	Taç Yapağı Genişliği
$\bar{x}_j$	5.8433	3.0573	3.7580	1.1993
$s(\bar{x}_j)$	0.8281	0.4359	1.7653	0.7622
vc_j	0.1417	0.1426	0.4697	0.6355

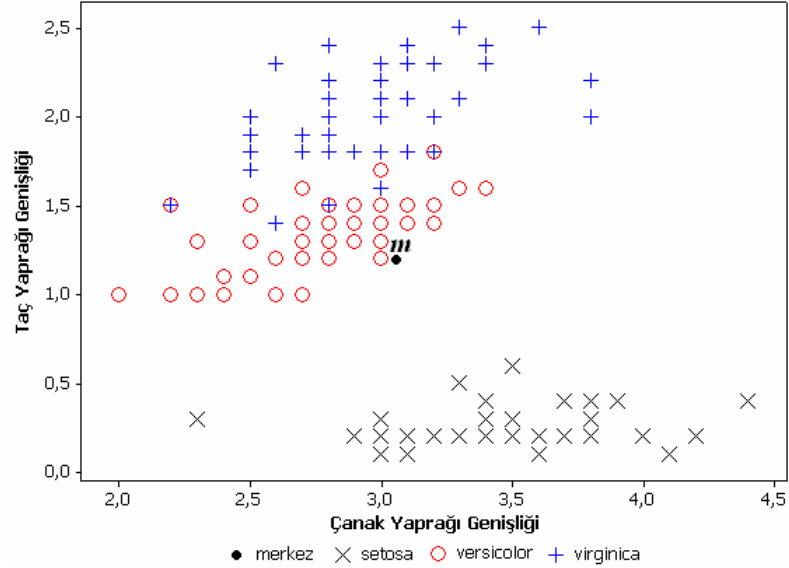
k -ortalamalar algoritmasında küme başlangıç merkezleri için önerilen algoritmaya göre en büyük değişim katsayısına sahip taç yapağı genişliği değişkeni temel eksen seçilir. İkinci eksenin belirlenmesi için temel eksen seçilen taç yapağı genişliği değişkeni ile diğer değişkenler arasındaki korelasyonlar Tablo 7.2'de verilmiştir.

Tablo 7.2. Taç yapağı genişliği değişkeni ile diğer değişkenler arasındaki korelasyonlar

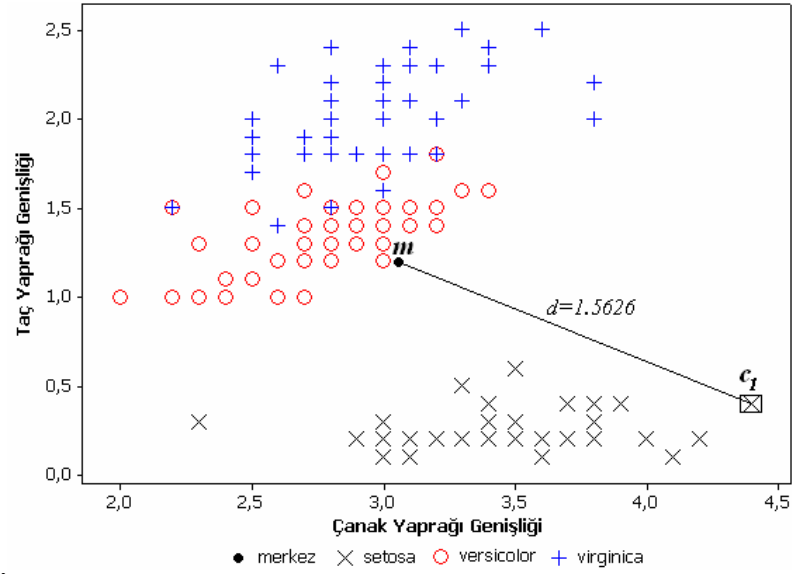
	Taç Yapağı Genişliği
Çanak Yapağı Uzunluğu	0.818
Çanak Yaprak Genişliği	-0.366
Taç Yapağı Uzunluğu	0.963

Tablo 7.2'deki korelasyonlar mutlak değerce incelendiğinde taç yapağı genişliği ile minimum korelasyona sahip değişken çanak yapağı genişliğidir. Böylece temel eksen olarak taç yapağı genişliği, ikinci eksen olarak da çanak yapağı genişliği seçilmiştir. İki eksen seçildikten sonra seçilen iki eksene göre veri

merkezi $\mathbf{m} = [1.1993 \ 3.0573]$ olarak belirlenir. Taç yaprağı genişliği ile çanak yaprağı genişliğine göre gözlemlerin saçılımı ve iki eksene göre veri merkezi Şekil 7.4.'de verilmiştir.

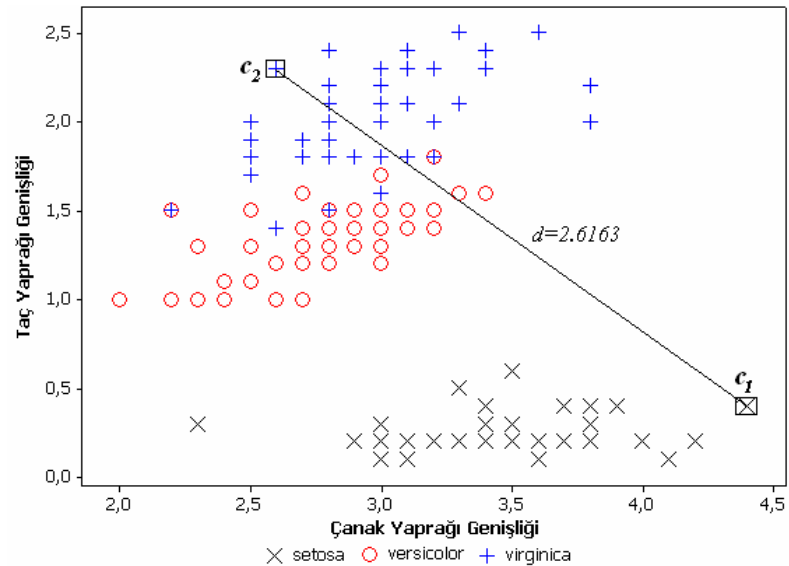


İris verisi için önerilen yöntemle göre ilk aday küme başlangıç merkezinin seçimi için tüm gözlemlerin merkeze olan Öklid uzaklığı hesaplanmış ve 1.5626 maksimum uzaklığa sahip $[0.4 \ 4.4]$ noktası birinci aday küme başlangıç merkezi c_1 olarak belirlenmiştir. İris verisinde birinci aday küme başlangıç merkezi c_1 olarak belirlenmesi Şekil 7.5'de verilmiştir.



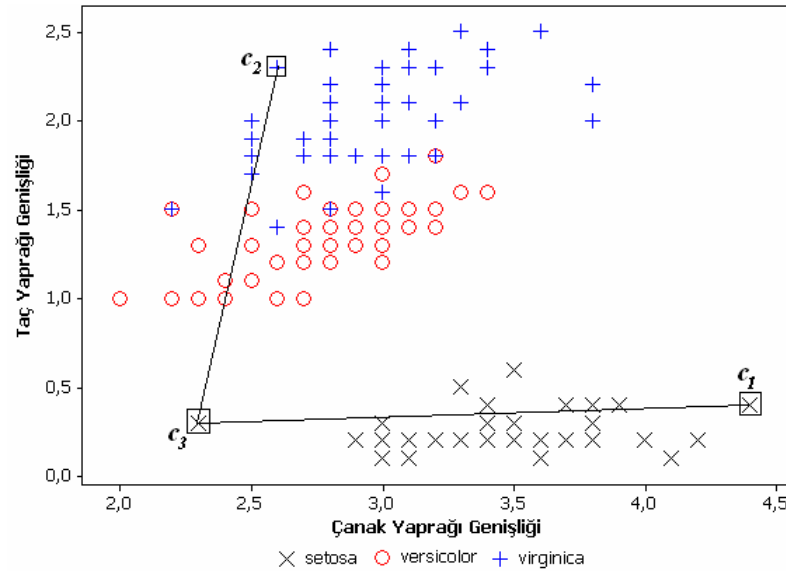
Şekil 7.5. İris verisinde önerilen yöntemle birinci aday küme başlangıç merkezinin seçimi

İkinci aday küme başlangıç merkezinin seçimi için bu kez tüm gözlemlerin $c_1 = [0.4 \quad 4.4]$ noktasına olan uzaklıkları hesaplanmış ve c_1 noktasına en uzak nokta 2.6163 Öklid uzaklığı ile $c_2 = [2.3 \quad 2.6]$ olarak belirlenmiştir. İris verisinde ikinci aday küme başlangıç merkezi c_2 olarak belirlenmesi Şekil 7.6'da verilmiştir.



Şekil 7.6. İris verisinde önerilen yöntemle ikinci aday küme başlangıç merkezinin seçimi

İris veri seti içerisinde üç küme olduğundan üçüncü aday küme başlangıç merkezinin seçimi için öncelikle tüm birimlerin $c_2 = [2.3 \ 2.6]$ noktasına olan uzaklığı hesaplanmış ve ardından tüm birimler için ilk iki aday küme başlangıç noktalarına olan uzaklık toplamaları Sd_{i3} hesaplanmıştır. Hesaplanan toplam uzaklıklar içerisinde 4.1248 toplam uzaklık ile maksimum değere sahip $c_3 = [0.3 \ 2.3]$ noktası üçüncü aday küme başlangıç merkezi olarak belirlenmiştir. Önerilen yöntemle göre üçüncü aday küme başlangıç merkezinin seçimi Şekil 7.7.'de gösterilmiştir.



Şekil 7.7. İris verisinde önerilen yöntemle üçüncü aday küme başlangıç merkezinin seçimi

k -ortalamalar algoritmasında küme başlangıç merkezleri için önerilen algoritmaya ile belirlenen iki eksen için oluşturulan aday küme başlangıç merkezlerine göre k -ortalamalar algoritması uygulanarak küme üyelikleri belirlenir. Belirlenen küme üyelikleri p değişken için verinin kümelenmesinde başlangıç küme üyelikleri olarak kabul edilir ve bu küme üyeliklerine göre başlangıç küme merkezleri oluşturulur. Küme başlangıç merkezlerinin seçimi için önerilen algoritmaya göre İris, Şarap, Harf ve Ruspini verilerinin kümelenmesi sonucunda elde edilen hatalı sınıflandırma yüzdesi, Rand indeksi ve Wilks'in lambda test

istatistiği değerleri ile rassal başlangıçlı k -ortalamlar algoritmasıyla elde edilen sonuçlar Tablo 7.3’de verilmiştir.

Tablo 7.3. Önerilen algoritma ile rassal başlangıçlı k - ortalamlar algoritmasının hatalı sınıflandırma yüzdesi, Rand indeks ve Wilks’in lambda istatistiğine göre karşılaştırılması

Veri seti	Algoritma	Hatalı Sınıflandırma Yüzdesi	Rand İndeks	Wilks'in Lambdası
İris	Önerilen algoritma	10.7000	0.8797	0.0322
	Rassal başlangıçlı	13.8300	0.8639	0.0376
Şarap	Önerilen algoritma	3.4000	0.9543	0.0196
	Rassal başlangıçlı	10.5800	0.9018	0.0329
Harf	Önerilen algoritma	7.9046	0.8543	0.0877
	Rassal başlangıçlı	9.7380	0.6364	0.1071
Ruspini	Önerilen algoritma	0.0000	1.0000	0.0034
	Rassal başlangıçlı	21.8667	0.8887	0.0160

Tablo 7.3’de yer alan rassal başlangıçlı k -ortalamlar algoritması sonuçları 10 farklı rassal başlangıçlı k -ortalamlar algoritmasından elde edilen sonuçlarının ortalamasıdır. Önerilen algoritma ile rassal başlangıçlı k -ortalamlar algoritmasının hatalı sınıflandırma yüzdesi, Rand indeks ve Wilks’in lambda test istatistiğine göre karşılaştırılması sonucunda dört veri setinde tüm kriterlere göre de önerilen algoritma rassal başlangıçlı k -ortalamlar algoritmasına göre daha başarılı bulunmuştur.

Khan ve Ahmad(2004)’in çalışmalarında İris, Şarap, Harf ve Ruspini verileri için CCIA algoritması ile hatalı sınıflandırma yüzdeleri sırasıyla 11.33, 5.05, 8.55 ve 4.00 elde edilmiştir. Ayrıca Deelers ve Auwatanamongkol(2007)’un çalışmalarında da aynı veri setleri için en büyük varyanslı eksene göre parçalama sonucunda da Khan ve Ahmad(2004)’in sonuçlarına paralel sonuçlar elde edilmiştir. Hatalı sınıflandırma yüzdesi bakımından önerilen yeni algoritma ile daha iyi sonuçlar elde edilmiştir. k -ortalamlar algoritmasında küme başlangıç merkezlerinin oluşturulması için önerdiğimiz algoritma uygulanabilirlik açısından literatürdeki mevcut algoritmalara göre oldukça basit olmasına karşın diğer yöntemlere göre kümeleme performansı bakımından da oldukça etkilidir.

7.4. Sonuçlar

Bu bölümde k -ortalamalar algoritmasında küme başlangıç merkezlerinin oluşturulması için yeni bir algoritma önerilmiştir. Önerilen algoritmanın temeli en büyük değişim katsayısına sahip değişken ve bu değişken ile mutlak değerce minimum korelasyona sahip değişken tarafından iki eksenin oluşturulmasıdır. Önerilen algoritma ile sınıflama problemlerinde yaygın bir kullanıma sahip olan gerçek veri setleri İris, Şarap, Harf ve Ruspini verilerinde rassal başlangıçlı k -ortalamalar algoritmasına göre daha başarılı kümelemeler elde edilmiştir. Ayrıca hatalı sınıflandırma yüzdesi bakımından Khan ve Ahmad(2004)'in önerdikleri CCIA algoritması ile Deelers ve Auwatanamongkol(2007) tarafından önerilen algoritmalara göre de önerilen yeni algoritma daha başarılı bulunmuştur. Literatürdeki mevcut algoritmaların çoğuna göre daha basit olan bu yeni algoritma küme başlangıç merkezlerinin oluşturulmasında öncelikle veri setini iki boyuta indirgediği için değişken sayısı çok fazla olan veri setlerinde de kolaylıkla uygulanabilir.

8. UZAKLIK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ

Bu bölümde, uzaklık ölçülerinin aşamalı ve aşamalı olmayan kümeleme tekniklerinde kümelemeye olan etkisi incelenecektir. Birimler arasında hesaplanan uzaklık matrisi ile ilgilenen aşamalı kümeleme tekniklerinde ve birimlerin küme merkezlerine olan uzaklığını göz önünde bulunduran aşamalı olmayan kümeleme tekniklerinden k -ortalamlar tekniğinde seçilen uzaklık ölçüsünün etkisi genelde küme geçerliliği için kullanılan farklı birkaç kritere göre değerlendirilecektir. Uzaklık ölçülerinin kümelemeye olan etkisi hem gerçek veri setleri üzerinde hem de simülasyon ile incelenecektir.

8.1. Aşamalı Kümeleme Tekniklerinde Uzaklık Ölçülerinin Etkisi

8.1.1. Aşamalı Kümeleme Tekniklerinde Uzaklık Ölçülerinin Etkisinin Gerçek Veri Setleri Üzerinde İncelenmesi

Aşamalı kümeleme tekniklerinde seçilen uzaklık ölçüsün kümelemeye etkisi incelenirken öncelikle gerçek veri setleri kullanılacaktır. Çalışma kapsamında Bölüm 3’de verilen aşamalı kümeleme tekniklerinden tek bağlantı, tam bağlantı, ortalama bağlantı, ağırlıklı ortalama bağlantı, merkez bağlantı, medyan bağlantı ve Ward bağlantı kümeleme teknikleri göz önünde bulundurulacaktır. Kümelemeye etkisi incelenecek uzaklık ölçüleri ise,

Manhattan (City-block) uzaklığı,

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^p |x_{ik} - x_{jk}| \quad (8.1)$$

Öklid uzaklığı,

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^p (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}} \quad (8.2)$$

Standartlaştırılmış Öklid uzaklığı,

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^p \frac{1}{s_k^2} (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}} \quad (8.3)$$

Minkowski uzaklığı ($\lambda = 4$),

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^p |x_{ik} - x_{jk}|^4 \right)^{\frac{1}{4}} \quad (8.4)$$

Mahalanobis uzaklığı,

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left[(\mathbf{x}_i - \mathbf{x}_j) \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)' \right]^{\frac{1}{2}} \quad (8.5)$$

Geometrik uzaklık,

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\prod_{k=1}^p (|x_{ik} - x_{jk}| + 1) \right)^{\frac{1}{p}} - 1 \quad (8.6)$$

Karma uzaklık

$$d(\mathbf{x}_i, \mathbf{x}_j) = w \left(\sum_{k=1}^p \frac{(x_{ik} - x_{jk})^2}{\bar{x}_{..k}^2} \right)^{\frac{1}{2}} + (1 - w) (1 - |\rho(\mathbf{x}_i, \mathbf{x}_j)|) \quad (8.7)$$

olarak belirlenmiştir. Çalışmada kullanılacak gerçek veri setleri, veri setlerinin örneklem hacmi, değişken sayısı ve küme sayısı ile ilgili bilgiler Tablo 8.1’de verilmiştir.

Tablo 8.1. Aşamalı kümeleme tekniklerinde uzaklık ölçülerinin etkisinin inceleneceği veri setlerinin özellikleri

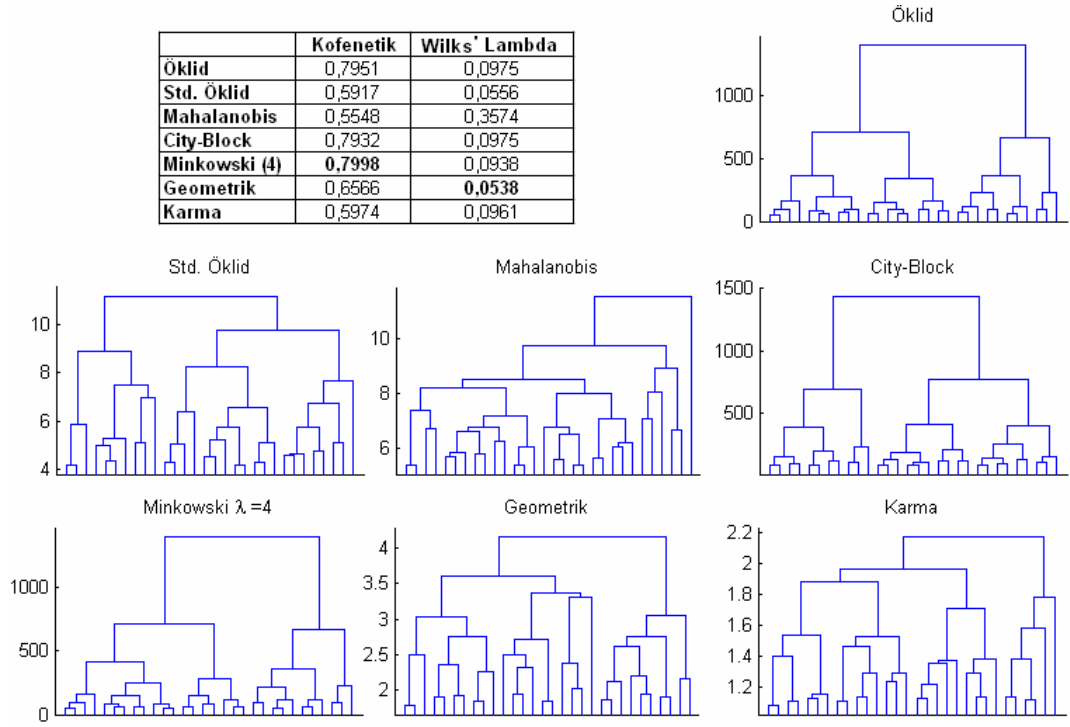
Veri Seti	Değişken Sayısı	Birim Sayısı	Küme Sayısı
İris	4	150	3
Şarap	13	178	3
Ecoli	5	336	8
Liver	6	345	2
Araç	18	325	4
Ruspini	2	75	4
Diyabet	4	145	3
Geyser	2	272	3
Harf	16	150	2
Yengeç	5	100	2
Cam	9	214	6
Landsat	36	180	6
Maya	6	220	10

Araç, Harf, Landsat ve Maya veri setlerinin gerçek büyüklükleri sırasıyla 846, 1594, 6435 ve 1484 birim iken uygulamamızda bu veri setleri içersinden her bir kümeyi barındıracak şekilde rassal olarak örneklemeler çekilmiştir. Ayrıca gerçekte Ecoli veri seti 7 değişkenden, Maya verisi de 8 değişkenden oluşmakta iken, uygulamamızda değişkenliğin az olduğu ve kümelemeyi bozduğu gerekçesi ile Ecoli ve Maya veri setlerinden 2'şer değişken kümelemede göz önünde bulundurulmamıştır.

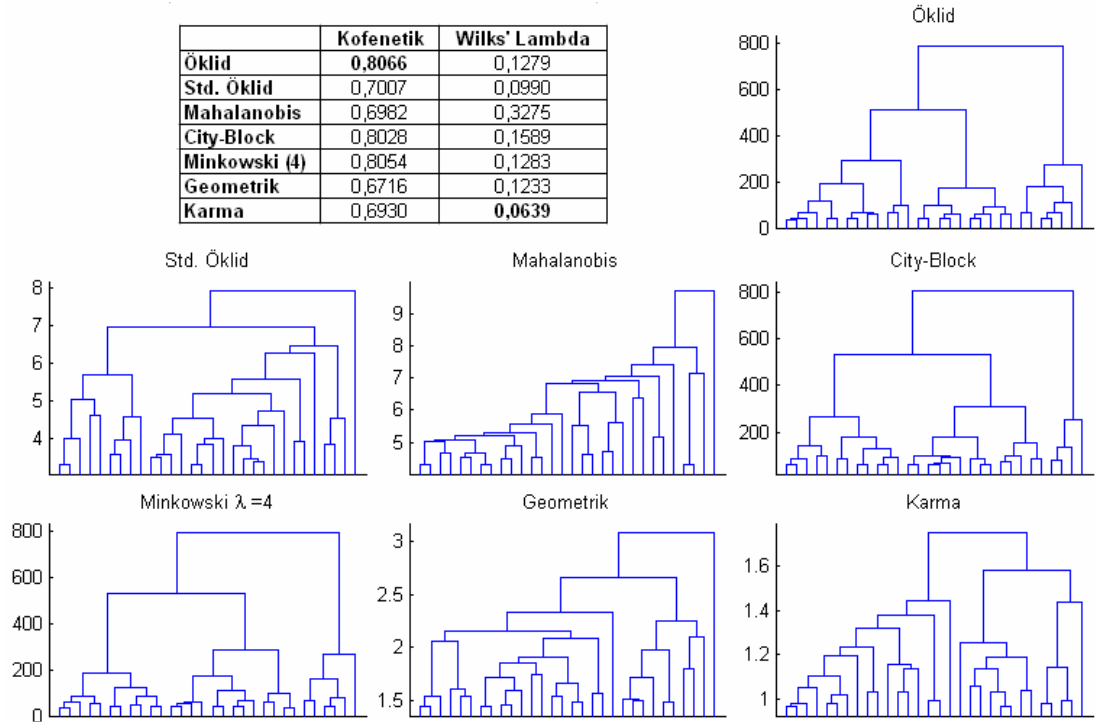
Aşamalı kümeleme analizi tekniklerinde uzaklık ölçülerinin kümelemeye etkisi incelenirken performans kriteri olarak içsel kriterlerden eşitlik (3.20) ve (3.21) ile verilen kofenetik korelasyon katsayısı ile Wilks'in lambda test istatistiği kullanılacaktır. Kofenetik korelasyon katsayısı kümeleme öncesi birimler arasındaki uzaklıklar ile kümeleme sonrası elde edilen uzaklıklar arasındaki ilişkinin bir ölçüsü olup, katsayı ne kadar büyük ise kümeleme o kadar başarılıdır. Çok değişkenli varyans analizinde gruplar arasındaki farklılığın test edilmesinde kullanılan Wilks'in lambda test istatistiğinin değeri ne kadar küçük ise kümeleme o kadar başarılıdır.

Şarap verisinin tam bağlantı kümeleme tekniğinde farklı uzaklık ölçüleri kullanılarak elde edilen ağaç grafikleri ile kofenetik korelasyon katsayısı ve Wilks'in lambda değerleri Şekil 8.1'de verilmiştir.

Kofenetik korelasyon katsayısına göre en iyi kümelenmenin elde edildiği uzaklık ölçüsü $\lambda = 4$ olan Minkowski uzaklığıdır. En kötü kümelemenin elde edildiği uzaklık ise Mahalanobis uzaklığıdır. Wilks'in lambda test istatistiğine göre değerlendirme yapabilmek için birimler üç kümeye atanacak şekilde tek bağlantı kümeleme tekniği uygulanmıştır. Wilks'in lambda test istatistiğine göre en başarılı kümeleme geometrik uzaklık ile elde edilmiştir. Wilks'in lambda test istatistiğine göre de en kötü kümeleme performansı kofenetik korelasyon katsayısındaki gibi Mahalanobis uzaklığı ile elde edilmiştir. Şekil 8.1'de yer alan ağaç grafikleri incelendiğinde, kümelenmenin seçilen uzaklık ölçüsüne göre değiştiği açıkça görülmektedir.



Şekil 8.1. Şarap verisinin tam bağlantı kümeleme tekniğinde farklı uzaklık ölçülerine göre elde edilen ağaç grafikleri



Şekil 8.2. Şarap verisinin ağırlıklı ortalama bağlantı kümeleme tekniğinde farklı uzaklık ölçülerine göre elde edilen ağaç grafikleri

Şarap verisinin ağırlıklı ortalama bağlantı kümeleme tekniğinde farklı uzaklık ölçüleri kullanılarak elde edilen ağaç grafikleri ile kofenetik korelasyon katsayısı ve Wilks'in lambda değerleri Şekil 8.2'de verilmiştir.

Kofenetik korelasyon katsayısına göre en iyi kümelenmenin elde edildiği uzaklık ölçüsü Öklid uzaklığı iken Wilks'in lambda test istatistiğine göre en başarılı kümeleme geometrik uzaklık ile elde edilmiştir.

İlgilenilen veri setlerinde farklı aşamalı kümeleme tekniği ve farklı uzaklık ölçülerine göre elde edilen kofenetik korelasyon katsayıları Tablo 8.2'de verilmiştir.

Tablo 8.2. İlgilenilen veri setlerinde farklı aşamalı kümeleme tekniği ve farklı uzaklık ölçülerine göre elde edilen kofenetik korelasyon katsayı değerleri

VERİ	Uzaklık	Tek	Tam	Ortalama	Ağırlıklı	Merkez	Medyan	Ward
İRİS	Öklid	0.8639	0.7276	0.8770	0.8680	0.8768	0.7537	0.8728
	Std. Öklid	0.8300	0.7515	0.8544	0.7159	0.8537	0.7249	0.8226
	Mahalanobis	0.6451	0.4721	0.6768	0.5518	0.6516	0.4980	0.5250
	City-Block	0.8529	0.6843	0.8698	0.8527	0.8681	0.8612	0.8660
	Minkowski (4)	0.8736	0.7091	0.8886	0.8524	0.8886	0.8846	0.8811
	Geometrik	0.8427	0.6907	0.8659	0.8314	0.8671	0.8596	0.8616
	Karma	0.8113	0.7294	0.8479	0.8268	0.8114	0.6884	0.8421
ŞARAP	Öklid	0.7765	0.7951	0.8023	0.8066	0.8023	0.7678	0.7964
	Std. Öklid	0.5436	0.5917	0.7591	0.7007	0.7565	0.6796	0.6623
	Mahalanobis	0.7171	0.5548	0.7567	0.6982	0.7459	0.7083	0.3152
	City-Block	0.7675	0.7932	0.7705	0.8028	0.8031	0.7723	0.7972
	Minkowski (4)	0.7756	0.7998	0.7682	0.8054	0.8025	0.7656	0.7594
	Geometrik	0.3068	0.6566	0.6939	0.6716	0.6971	0.6040	0.6604
	Karma	0.5147	0.5974	0.7461	0.6930	0.7004	0.4046	0.6843
ECOLİ	Öklid	0.4063	0.7566	0.7895	0.7196	0.7921	0.7590	0.7576
	Std. Öklid	0.5053	0.6853	0.7531	0.7303	0.7640	0.4900	0.7004
	Mahalanobis	0.6973	0.6469	0.7560	0.5619	0.7496	0.5565	0.5431
	City-Block	0.4938	0.7524	0.7850	0.7407	0.7816	0.5930	0.7542
	Minkowski (4)	0.4275	0.7428	0.7544	0.7418	0.7646	0.6808	0.7335
	Geometrik	0.5010	0.7443	0.7789	0.5655	0.7768	0.6561	0.7509
	Karma	0.3993	0.2884	0.4608	0.3743	0.4146	0.3842	0.2560
LİVER	Öklid	0.9089	0.7796	0.9158	0.8328	0.9098	0.8732	0.7036
	Std. Öklid	0.8827	0.7369	0.8736	0.7040	0.8755	0.6662	0.6602
	Mahalanobis	0.8608	0.6216	0.8498	0.6436	0.8630	0.7805	0.3714
	City-Block	0.9100	0.8261	0.9057	0.8036	0.9032	0.7945	0.7625
	Minkowski (4)	0.8922	0.8617	0.8945	0.8309	0.8939	0.8909	0.7108
	Geometrik	0.7818	0.6809	0.7999	0.6041	0.7784	0.7092	0.6834
	Karma	0.8920	0.8653	0.8982	0.8797	0.8889	0.5776	0.6363

Tablo 8.2 (devam). İlgilenilen veri setlerinde farklı aşamalı kümeleme tekniği ve farklı uzaklık ölçülerine göre elde edilen kofenetik korelasyon katsayı değerleri

VERİ	Uzaklık	Tek	Tam	Ortalama	Ağırlıklı	Merkez	Medyan	Ward
ARAÇ	Öklid	0.5417	0.8264	0.8145	0.8042	0.8313	0.8065	0.7909
	Std. Öklid	0.5850	0.7407	0.7875	0.7795	0.8230	0.7353	0.5938
	Mahalanobis	0.7906	0.6883	0.8226	0.7848	0.8169	0.7289	0.3674
	City-Block	0.4546	0.8184	0.7820	0.8003	0.8061	0.7889	0.7819
	Minkowski (4)	0.5492	0.7922	0.8475	0.8389	0.8486	0.8507	0.8326
	Geometrik	0.4707	0.7285	0.7694	0.6970	0.7729	0.7532	0.7262
	Karma	0.7222	0.7114	0.8266	0.7960	0.7923	0.4772	0.5575
RUSPİNİ	Öklid	0.8730	0.8892	0.9008	0.8845	0.8985	0.8826	0.8914
	Std. Öklid	0.8475	0.8433	0.8762	0.8645	0.8581	0.8595	0.8662
	Mahalanobis	0.8392	0.8533	0.8737	0.8580	0.8548	0.8496	0.8607
	City-Block	0.8709	0.8858	0.8990	0.8764	0.8935	0.8795	0.8885
	Minkowski (4)	0.8611	0.8887	0.8899	0.8816	0.8889	0.8753	0.8816
	Geometrik	0.7776	0.8107	0.8195	0.7866	0.8101	0.8130	0.8098
	Karma	0.2721	0.3444	0.4177	0.3710	0.3854	0.3646	0.3557
DİYABET	Öklid	0.8431	0.8751	0.8795	0.8688	0.8787	0.8776	0.8656
	Std. Öklid	0.7867	0.8223	0.8343	0.7732	0.8260	0.7068	0.8243
	Mahalanobis	0.7988	0.7522	0.8628	0.7850	0.8476	0.7113	0.5520
	City-Block	0.8482	0.8606	0.8761	0.8744	0.8754	0.8588	0.8672
	Minkowski (4)	0.8338	0.8663	0.8754	0.8464	0.8733	0.8384	0.8614
	Geometrik	0.8162	0.8374	0.8418	0.8167	0.8415	0.8364	0.8304
	Karma	0.8007	0.8449	0.8610	0.6279	0.8413	0.6088	0.7645
GEYSER	Öklid	0.8168	0.8210	0.8535	0.8216	0.8191	0.8315	0.8470
	Std. Öklid	0.9152	0.8842	0.9207	0.8989	0.9184	0.9016	0.9141
	Mahalanobis	0.7200	0.6594	0.7744	0.5975	0.7691	0.7060	0.7445
	City-Block	0.8085	0.7987	0.8658	0.8279	0.8658	0.8446	0.8604
	Minkowski (4)	0.8117	0.8269	0.8522	0.5838	0.8184	0.8417	0.8454
	Geometrik	0.8213	0.8819	0.8611	0.8510	0.8609	0.8964	0.9003
	Karma	0.9087	0.9083	0.9318	0.8994	0.9278	0.8000	0.9285
HARF	Öklid	0.6597	0.5365	0.6976	0.6629	0.6842	0.5092	0.6003
	Std. Öklid	0.6968	0.5518	0.7333	0.7050	0.7029	0.5803	0.6266
	Mahalanobis	0.7442	0.6893	0.7810	0.6505	0.7566	0.7194	0.3963
	City-Block	0.6785	0.6375	0.6990	0.5803	0.6955	0.5126	0.6342
	Minkowski (4)	0.5837	0.5838	0.6540	0.6001	0.6258	0.5779	0.5805
	Geometrik	0.6281	0.5964	0.6789	0.6427	0.6749	0.5557	0.6271
	Karma	0.6286	0.5926	0.6936	0.5486	0.6474	0.4567	0.6213
YENGEÇ	Öklid	0.5668	0.6891	0.6948	0.6878	0.6944	0.6443	0.6780
	Std. Öklid	0.5350	0.6895	0.6968	0.6694	0.6945	0.6684	0.6787
	Mahalanobis	0.6380	0.4444	0.6677	0.5150	0.7021	0.4175	0.4195
	City-Block	0.5471	0.6998	0.6996	0.6408	0.6985	0.6311	0.6678
	Minkowski (4)	0.5581	0.6907	0.6916	0.6349	0.6929	0.6806	0.6766
	Geometrik	0.5273	0.7013	0.7006	0.6240	0.7005	0.6216	0.6675
	Karma	0.5262	0.6939	0.6985	0.6492	0.6589	0.3032	0.6797

Tablo 8.2 (devam). İlgilenilen veri setlerinde farklı aşamalı kümeleme tekniği ve farklı uzaklık ölçülerine göre elde edilen kofenetik korelasyon katsayı değerleri

VERİ	Uzaklık	Tek	Tam	Ortalama	Ağırlıklı	Merkez	Medyan	Ward
CAM	Öklid	0.8294	0.8909	0.9479	0.8722	0.9512	0.8324	0.8356
	Std. Öklid	0.8768	0.7222	0.9264	0.8359	0.9234	0.8087	0.6996
	Mahalanobis	0.9012	0.7829	0.9272	0.8649	0.9208	0.8849	0.5059
	City-Block	0.8291	0.8954	0.9265	0.8393	0.9324	0.8412	0.8357
	Minkowski (4)	0.7932	0.8938	0.9491	0.8011	0.9461	0.8646	0.8035
	Geometrik	0.8457	0.8573	0.9251	0.8351	0.9161	0.8836	0.8094
	Karma	0.8309	0.8797	0.9245	0.9464	0.9470	0.9375	0.8130
LANDSAT	Öklid	0.7444	0.6449	0.8763	0.8141	0.8758	0.8241	0.7790
	Std. Öklid	0.7396	0.6845	0.8734	0.8072	0.8709	0.7904	0.7629
	Mahalanobis	0.7895	0.6895	0.8156	0.7777	0.8146	0.7703	0.4698
	City-Block	0.7136	0.7079	0.8606	0.7979	0.8281	0.8062	0.7443
	Minkowski (4)	0.7654	0.7991	0.8614	0.8171	0.8813	0.8192	0.7762
	Geometrik	0.6245	0.7642	0.8142	0.6364	0.8138	0.6952	0.6982
	Karma	0.6976	0.7238	0.8463	0.8240	0.8394	0.7313	0.7156
MAYA	Öklid	0.5232	0.6211	0.7195	0.6650	0.7098	0.6253	0.5610
	Std. Öklid	0.6671	0.4637	0.7059	0.6101	0.6959	0.5797	0.4429
	Mahalanobis	0.7103	0.5897	0.7064	0.5417	0.7365	0.5947	0.4530
	City-Block	0.5252	0.5892	0.6992	0.6205	0.7018	0.6374	0.5856
	Minkowski (4)	0.5004	0.6414	0.7013	0.6649	0.7023	0.6492	0.5895
	Geometrik	0.5284	0.6025	0.6855	0.6108	0.6882	0.6284	0.5337
	Karma	0.3558	0.5896	0.7193	0.6803	0.6135	0.5077	0.4983

Tablo 8.2 incelendiğinde kofenetik korelasyon katsayısına göre en iyi kümelemeler, İris verisinde Minkowski ($\lambda = 4$) uzaklığı ile ortalama bağlantı ve merkez bağlantı kümeleme tekniklerinde, Şarap verisinde Öklid uzaklığı ile ağırlıklı ortalama bağlantı tekniğinde, Ecoli verisinde Öklid uzaklığı ile merkez bağlantı tekniğinde, Liver verisinde Öklid uzaklığı ile ortalama bağlantı tekniğinde, Araç verisinde Minkowski ($\lambda = 4$) uzaklığı ile medyan bağlantı tekniğinde, Ruspini verisinde Öklid uzaklığı ile ortalama bağlantı tekniğinde, Diyabet verisinde Öklid uzaklığı ile ortalama bağlantı tekniğinde, Geyser verisinde karma uzaklığı ile ortalama bağlantı tekniğinde, Harf verisinde Mahalanobis uzaklığı ile ortalama bağlantı tekniğinde, Yengeç verisinde Mahalanobis uzaklığı ile merkez bağlantı tekniğinde, Cam verisinde Öklid uzaklığı ile merkez bağlantı tekniğinde, Landsat verisinde Öklid uzaklığı ile merkez bağlantı tekniğinde ve Maya verisinde Mahalanobis uzaklığı ile merkez bağlantı tekniğinde elde edilmiştir.

İlgilenilen 13 gerçek veri setinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen kofenetik korelasyon katsayı değerlerinin ortalaması ve standart hatası Tablo 8.3’de verilmiştir.

Tablo 8.3. İlgilenilen 13 gerçek veri setinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen kofenetik korelasyon katsayı değerlerinin ortalaması ve standart hatası

Uzaklık	T.İstatistik	Tek	Tam	Ortalama	Ağırlıklı	Merkez	Medyan	Ward
Öklit	Ortalama	0.7195	0.7579	0.8284	0.7929	0.8249	0.7682	0.7676
	Std. Hata	0.0162	0.0110	0.0084	0.0081	0.0086	0.0113	0.0105
Std. Öklit	Ortalama	0.7239	0.7052	0.8150	0.7534	0.8125	0.7070	0.7119
	Std. Hata	0.0146	0.0117	0.0080	0.0082	0.0082	0.0117	0.0127
Mahalanobis	Ortalama	0.7579	0.6496	0.7900	0.6793	0.7868	0.6866	0.5018
	Std. Hata	0.0081	0.0116	0.0078	0.0124	0.0074	0.0135	0.0154
City-Block	Ortalama	0.7154	0.7653	0.8184	0.7737	0.8195	0.7555	0.7727
	Std. Hata	0.0160	0.0096	0.0083	0.0099	0.0081	0.0120	0.0096
Minkowski 4	Ortalama	0.7097	0.7766	0.8175	0.7615	0.8175	0.7861	0.7640
	Std. Hata	0.0161	0.0097	0.0093	0.0104	0.0095	0.0104	0.0102
Geometrik	Ortalama	0.6517	0.7348	0.7873	0.7056	0.7845	0.7317	0.7353
	Std. Hata	0.0175	0.0093	0.0079	0.0103	0.0077	0.0116	0.0104
Karma	Ortalama	0.6431	0.6746	0.7594	0.7013	0.7283	0.5570	0.6425
	Std. Hata	0.0212	0.0193	0.0164	0.0187	0.0181	0.0189	0.0190

Kofenetik korelasyon katsayı değerlerinin ortalaması incelendiğinde Öklid uzaklığının en başarılı olduğu aşamalı kümeleme teknikleri ortalama bağlantı, ağırlıklı ortalama bağlantı ve merkez bağlantı kümeleme teknikleri olmuştur. Tam bağlantı ve medyan bağlantı kümeleme tekniklerinde en başarılı uzaklık ölçüsü Minkowski ($\lambda = 4$) iken Ward bağlantı kümeleme tekniğinde City-Block uzaklığı, tek bağlantı kümeleme tekniğinde Mahalanobis uzaklığı en başarılı uzaklık ölçüleri olarak belirlenmiştir. Yeni tanımlanan geometrik ve karma uzaklık ölçülerinin standart hata değerleri diğer uzaklık ölçülerine göre oldukça yüksek bulunmuştur. Uzaklık ölçülerinin ilgili gerçek veri setlerinde performansı tek tek incelendiğinde yeni tanımlanan uzaklık ölçülerinin çoğu veri setinde diğer uzaklık ölçüleri ile benzer sonuçlar elde ettiği gözlemlenirken özellikle Ruspini verisinde karma uzaklık ölçüsünün performansı çok kötü bulunmuştur.

İlgilenilen veri setlerinde farklı aşamalı kümeleme tekniği ve farklı uzaklık ölçülerine göre elde edilen Wilks'in lambda test istatistiği değerleri Tablo 8.4'de verilmiştir.

Tablo 8.4. İlgilenilen veri setlerinde farklı aşamalı kümeleme tekniği ve farklı uzaklık ölçülerine göre elde edilen Wilks'in lambda test istatistiği değerleri

VERİ	Uzaklık	Tek	Tam	Ortalama	Ağırlıklı	Merkez	Medyan	Ward
İRİS	Öklid	0.0766	0.0409	0.0313	0.0310	0.0313	0.0465	0.0312
	Std. Öklid	0.0826	0.0515	0.0761	0.1575	0.0761	0.1830	0.0463
	Mahalanobis	0.7544	0.2277	0.7525	0.1764	0.7525	0.2313	0.0335
	City-Block	0.0859	0.0318	0.0315	0.0255	0.0756	0.0465	0.0320
	Minkowski (4)	0.0766	0.0406	0.0441	0.0406	0.0454	0.0441	0.0313
	Geometrik	0.0859	0.0310	0.0315	0.0353	0.0756	0.0310	0.0312
	Karma	0.0766	0.0231	0.0228	0.0633	0.0862	0.3395	0.0333
ŞARAP	Öklid	0.7037	0.0975	0.1575	0.1279	0.1575	0.1411	0.0946
	Std. Öklid	0.3879	0.0556	0.5638	0.0990	0.4267	0.5456	0.0255
	Mahalanobis	0.5211	0.3574	0.5211	0.3275	0.5211	0.5211	0.0722
	City-Block	0.7049	0.0975	0.1194	0.1589	0.1532	0.1271	0.0986
	Minkowski (4)	0.7037	0.0938	0.1194	0.1283	0.1575	0.1602	0.1192
	Geometrik	0.6910	0.0538	0.0847	0.1233	0.5417	0.6044	0.0347
	Karma	0.4359	0.0961	0.1250	0.0639	0.5835	0.4359	0.0586
ECOLİ	Öklid	0.5327	0.0099	0.0118	0.0082	0.0136	0.0150	0.0041
	Std. Öklid	0.5748	0.0092	0.0095	0.0077	0.0114	0.0108	0.0044
	Mahalanobis	0.5477	0.0272	0.2113	0.0270	0.3562	0.0879	0.0059
	City-Block	0.4008	0.0057	0.0126	0.0080	0.0129	0.0363	0.0066
	Minkowski (4)	0.5327	0.0069	0.0147	0.0073	0.0170	0.0139	0.0039
	Geometrik	0.4008	0.0072	0.0125	0.0051	0.0136	0.0408	0.0066
	Karma	0.6443	0.2087	0.6388	0.3446	0.7510	0.6612	0.1487
LİVER	Öklid	0.7994	0.7994	0.4674	0.7994	0.4144	0.4506	0.3322
	Std. Öklid	0.7920	0.3904	0.6017	0.9030	0.5561	0.9030	0.3448
	Mahalanobis	0.7920	0.4307	0.7920	0.7920	0.7920	0.7920	0.4835
	City-Block	0.6017	0.5513	0.5192	0.3571	0.4455	0.3571	0.3084
	Minkowski (4)	0.7994	0.4506	0.4078	0.4185	0.4078	0.4506	0.3208
	Geometrik	0.8524	0.4139	0.5651	0.5651	0.5171	0.5651	0.3891
	Karma	0.7994	0.7182	0.4532	0.8084	0.8409	0.7994	0.3484
ARAÇ	Öklid	0.0102	0.0137	0.0019	0.0019	0.0104	0.0157	0.0136
	Std. Öklid	0.0266	0.0017	0.0031	0.0029	0.0047	0.0023	0.0023
	Mahalanobis	0.0200	0.0313	0.0200	0.0200	0.0200	0.0200	0.0028
	City-Block	0.0102	0.0094	0.0021	0.0021	0.0020	0.0020	0.0158
	Minkowski (4)	0.0135	0.0106	0.0121	0.0117	0.0101	0.0134	0.0090
	Geometrik	0.1419	0.0020	0.0015	0.0023	0.0017	0.0019	0.0126
	Karma	0.0169	0.0111	0.0131	0.0096	0.0328	0.0117	0.0081

Tablo 8.4 (devam). İlgilenilen veri setlerinde farklı aşamalı kümeleme tekniği ve farklı uzaklık ölçülerine göre elde edilen Wilks'in lambda test istatistiği değerleri

VERİ	Uzaklık	Tek	Tam	Ortalama	Ağırlıklı	Merkez	Medyan	Ward
RUSPINİ	Öklid	0.0034	0.0051	0.0034	0.0034	0.0034	0.0034	0.0034
	Std. Öklid	0.0034	0.0034	0.0034	0.0034	0.0034	0.0034	0.0034
	Mahalanobis	0.0034	0.0034	0.0034	0.0034	0.0034	0.0034	0.0034
	City-Block	0.0034	0.0051	0.0034	0.0034	0.0034	0.0034	0.0034
	Minkowski (4)	0.0034	0.0034	0.0034	0.0034	0.0034	0.0034	0.0034
	Geometrik	0.0239	0.0034	0.0041	0.0034	0.0034	0.0034	0.0034
	Karma	0.9078	0.0710	0.0900	0.0882	0.1162	0.1169	0.1085
DİYABET	Öklid	0.1191	0.0683	0.0820	0.0607	0.0577	0.0683	0.0581
	Std. Öklid	0.6048	0.0609	0.0625	0.1067	0.0686	0.0924	0.0530
	Mahalanobis	0.3784	0.1955	0.3784	0.2627	0.3784	0.3784	0.1050
	City-Block	0.1191	0.0577	0.1191	0.0577	0.1191	0.0584	0.0581
	Minkowski (4)	0.1191	0.0683	0.0820	0.0583	0.0577	0.0583	0.0577
	Geometrik	0.0657	0.0707	0.0676	0.0676	0.0577	0.0707	0.0686
	Karma	0.7069	0.0860	0.0820	0.0861	0.0879	0.0728	0.0645
GEYSER	Öklid	0.9652	0.1245	0.0781	0.1172	0.0881	0.0913	0.0727
	Std. Öklid	0.0884	0.0547	0.0629	0.0449	0.0624	0.0565	0.0511
	Mahalanobis	0.9615	0.1504	0.0577	0.1930	0.0785	0.2646	0.0555
	City-Block	0.9652	0.1011	0.0736	0.0940	0.0736	0.0913	0.0742
	Minkowski (4)	0.9652	0.1313	0.0781	0.1011	0.0890	0.1029	0.0727
	Geometrik	0.9652	0.0736	0.0804	0.0804	0.0804	0.0736	0.0727
	Karma	0.0880	0.0460	0.0675	0.0447	0.0871	0.0852	0.0480
HARF	Öklid	0.3511	0.2921	0.3511	0.0361	0.3511	0.2951	0.2970
	Std. Öklid	0.6736	0.3004	0.3511	0.2248	0.7978	0.4283	0.0646
	Mahalanobis	0.7381	0.3024	0.6328	0.6328	0.7288	0.7556	0.1626
	City-Block	0.7588	0.2704	0.0361	0.2228	0.3511	0.3571	0.1109
	Minkowski (4)	0.7588	0.2658	0.3198	0.2893	0.3511	0.2186	0.2554
	Geometrik	0.8255	0.2850	0.0361	0.1109	0.3511	0.2289	0.0361
	Karma	0.3511	0.3104	0.1124	0.2970	0.7575	0.6686	0.0397
YENGEÇ	Öklid	0.8779	0.2806	0.2739	0.2758	0.2806	0.3500	0.2758
	Std. Öklid	0.8779	0.2837	0.2758	0.2837	0.2837	0.2837	0.2758
	Mahalanobis	0.8433	0.5820	0.6933	0.4987	0.7059	0.6933	0.3220
	City-Block	0.8779	0.2758	0.2806	0.3594	0.2806	0.3500	0.3846
	Minkowski (4)	0.8779	0.2758	0.2739	0.3500	0.2739	0.2739	0.2758
	Geometrik	0.8779	0.2758	0.2806	0.3500	0.2806	0.3594	0.3846
	Karma	0.8779	0.2758	0.2758	0.3846	0.8779	0.8454	0.2758
CAM	Öklid	0.0280	0.0015	0.0144	0.0077	0.0164	0.0211	0.0028
	Std. Öklid	0.0470	0.0109	0.0137	0.0117	0.0178	0.0154	0.0020
	Mahalanobis	0.0230	0.0091	0.0337	0.0241	0.0337	0.0241	0.0015
	City-Block	0.0303	0.0022	0.0183	0.0090	0.0201	0.0090	0.0015
	Minkowski (4)	0.0384	0.0022	0.0035	0.0092	0.0171	0.0205	0.0011
	Geometrik	0.0303	0.0025	0.0183	0.0090	0.0183	0.0151	0.0018
	Karma	0.0265	0.0005	0.0071	0.0069	0.0071	0.0069	0.0003

Tablo 8.4 (devam). İlgilenilen veri setlerinde farklı aşamalı kümeleme tekniği ve farklı uzaklık ölçülerine göre elde edilen Wilks'in lambda test istatistiği değerleri

VERİ	Uzaklık	Tek	Tam	Ortalama	Ağırlıklı	Merkez	Medyan	Ward
LANDSAT	Öklid	0.00169	0.00006	0.00004	0.00006	0.00009	0.00012	0.00004
	Std. Öklid	0.00133	0.00011	0.00004	0.00004	0.00005	0.00004	0.00002
	Mahalanobis	0.00697	0.00334	0.00697	0.00424	0.02110	0.00697	0.00005
	City-Block	0.00147	0.00003	0.00004	0.00012	0.00004	0.00047	0.00004
	Minkowski (4)	0.00089	0.00004	0.00004	0.00004	0.00005	0.00004	0.00003
	Geometrik	0.03330	0.00016	0.00003	0.00014	0.00003	0.00065	0.00002
	Karma	0.00086	0.00013	0.00010	0.00018	0.00073	0.00134	0.00007
MAYA	Öklid	0.2379	0.0070	0.0151	0.0087	0.0185	0.0078	0.0041
	Std. Öklid	0.3467	0.0074	0.0129	0.0058	0.0618	0.0169	0.0046
	Mahalanobis	0.2222	0.0076	0.0199	0.0103	0.0907	0.0454	0.0047
	City-Block	0.2131	0.0087	0.0168	0.0091	0.0535	0.0110	0.0053
	Minkowski (4)	0.2379	0.0072	0.0207	0.0098	0.0241	0.0184	0.0046
	Geometrik	0.2131	0.0089	0.0131	0.0074	0.0370	0.0123	0.0048
	Karma	0.5760	0.0202	0.0679	0.0681	0.3829	0.4164	0.0102

Tablo 8.4 incelendiğinde Wilks'in lambda test istatistiğine göre en iyi kümelemeler, İris verisinde karma uzaklığı ile ortalama bağlantı kümeleme tekniğinde, Şarap verisinde standartlaştırılmış Öklid uzaklığı ile Ward bağlantı tekniğinde, Ecoli verisinde Minkowski ($\lambda = 4$) uzaklığı ile Ward bağlantı tekniğinde, Liver verisinde City-Block uzaklığı ile Ward bağlantı tekniğinde, Araç verisinde geometrik uzaklığı ile ortalama bağlantı tekniğinde, Diyabet verisinde standartlaştırılmış Öklid uzaklığı ile Ward bağlantı tekniğinde, Geyser verisinde karma uzaklığı ile ağırlıklı ortalama bağlantı tekniğinde, Harf verisinde Öklid uzaklığı ile ağırlıklı ortalama bağlantı tekniğinde, Cam verisinde karma uzaklığı ile Ward bağlantı tekniğinde, Landsat verisinde geometrik uzaklığı ile Ward bağlantı tekniğinde ve Maya verisinde Öklid uzaklığı ile Ward bağlantı tekniğinde elde edilmiştir. Wilks'in lambda test istatistiğine göre aşamalı kümeleme teknikleri içerisinde Ward bağlantı kümeleme tekniği en iyi sonuçları vermiştir.

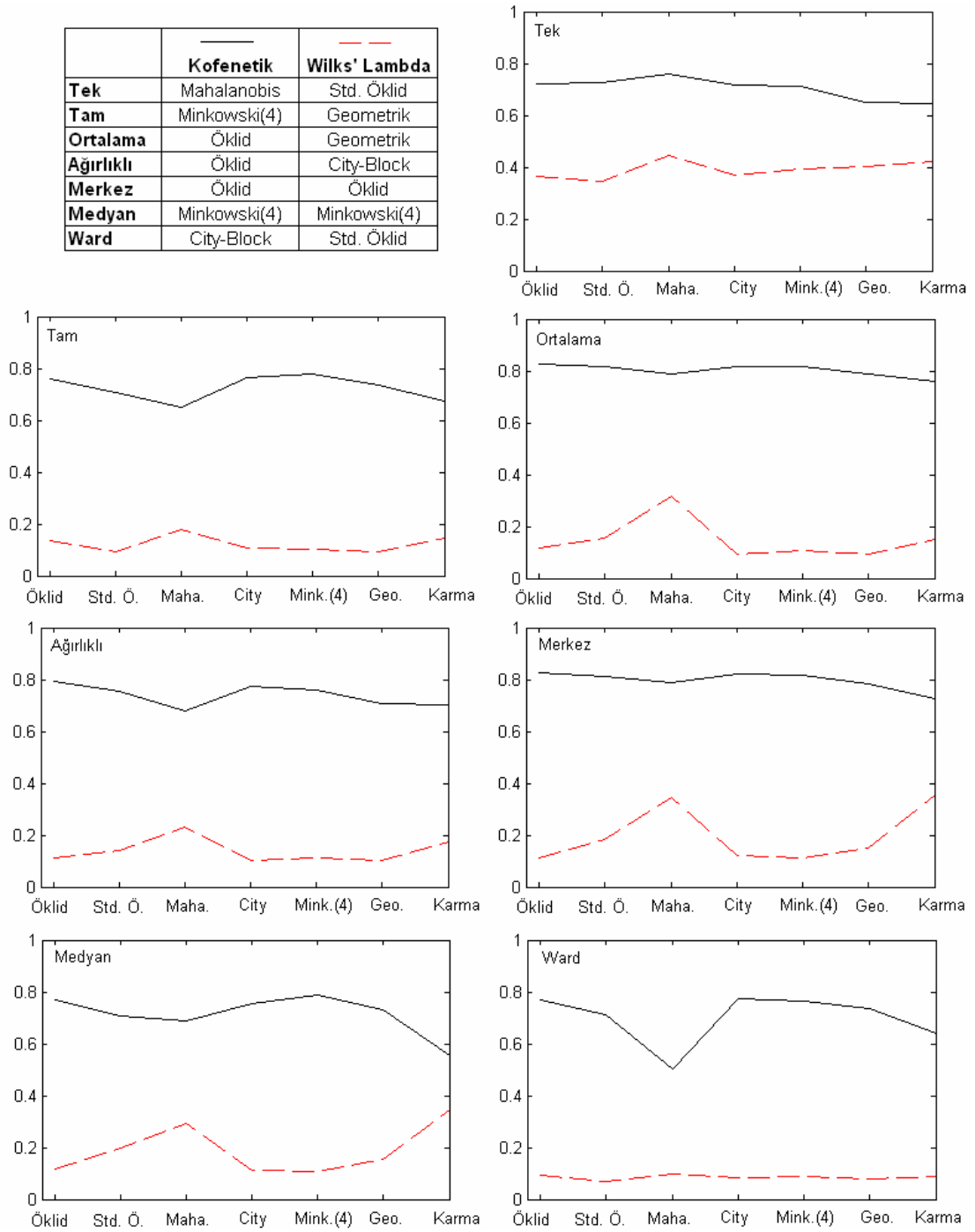
İlgilenilen 13 gerçek veri setinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen Wilks'in lambda test istatistiği değerlerinin ortalaması ve standart hatası Tablo 8.5'de verilmiştir.

Tablo 8.5. İlgilenilen 13 gerçek veri setinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen Wilks'in lambda test istatistiği değerlerinin ortalaması ve standart hatası

Uzaklık	T.İstatistik	Tek	Tam	Ortalama	Ağırlıklı	Merkez	Medyan	Ward
Öklit	Ortalama	0.3620	0.1339	0.1145	0.1137	0.1110	0.1159	0.0915
	Std. Hata	0.0367	0.0224	0.0154	0.0221	0.0145	0.0151	0.0124
Std. Öklit	Ortalama	0.3467	0.0946	0.1567	0.1424	0.1824	0.1955	0.0675
	Std. Hata	0.0326	0.0135	0.0219	0.0247	0.0257	0.0278	0.0111
Mahalanobis	Ortalama	0.4471	0.1791	0.3172	0.2286	0.3448	0.2942	0.0964
	Std. Hata	0.0359	0.0192	0.0320	0.0265	0.0321	0.0303	0.0148
City-Block	Ortalama	0.3671	0.1090	0.0948	0.1006	0.1224	0.1115	0.0846
	Std. Hata	0.0367	0.0164	0.0150	0.0133	0.0147	0.0144	0.0123
Minkowski 4	Ortalama	0.3944	0.1043	0.1061	0.1098	0.1119	0.1060	0.0889
	Std. Hata	0.0382	0.0142	0.0138	0.0146	0.0142	0.0136	0.0118
Geometrik	Ortalama	0.4005	0.0945	0.0920	0.1046	0.1522	0.1544	0.0805
	Std. Hata	0.0381	0.0138	0.0161	0.0168	0.0200	0.0218	0.0138
Karma	Ortalama	0.4237	0.1436	0.1504	0.1743	0.3547	0.3432	0.0880
	Std. Hata	0.0351	0.0202	0.0193	0.0232	0.0354	0.0318	0.0109

Wilks'in lambda test istatistiği değerlerinin ortalaması incelendiğinde aşamalı kümeleme tekniklerinde en başarılı uzaklıklar tek bağlantı tekniğinde standartlaştırılmış Öklid uzaklığı, tam bağlantı tekniğinde geometrik uzaklık, ortalama bağlantı tekniğinde geometrik uzaklık, ağırlıklı ortalama bağlantı tekniğinde City-Block uzaklığı, merkez bağlantı tekniğinde Öklid uzaklığı, medyan bağlantı tekniğinde Minkowski ($\lambda = 4$) uzaklığı ve Ward bağlantı tekniğinde standartlaştırılmış Öklid uzaklığı olmuştur. Wilks'in lambda test istatistiğine göre en başarılı aşamalı kümeleme tekniği Ward kümeleme tekniği iken en başarısız kümeleme tekniği tek bağlantı kümeleme tekniğidir.

Gerçek veri setlerinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçülerine göre kümelenmesi sonrası elde edilen ortalama kofenetik korelasyon katsayısı ve ortalama Wilks'in lambda test istatistiğine göre grafiksel olarak değerlendirilmesi Şekil 8.3'de verilmiştir.



Şekil 8.3. Gerçek veri setlerinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçülerine göre kümelendiği ile elde edilen ortalama kofenetik korelasyon katsayısı ve ortalama Wilks'in lambda test istatistiğine göre değerlendirilmesi

8.1.2. Aşamalı Kümeleme Tekniklerinde Uzaklık Ölçülerinin Etkisinin Waveform Simülasyon Veri Setlerinde İncelenmesi

Aşamalı kümeleme tekniklerinde uzaklık ölçülerinin kümelemeye etkisini göstermek için literatürde oldukça bilinen ve simülasyonla elde edilen Waveform veri seti kullanılacaktır (Breiman ve ark 1984). Üç farklı alt kümenin ve her bir alt kümede 21 değişkenin bulunduğu Waveform veri kümesi özellikle ayırıştırma analizi için zor bir veri setidir. Waveform veri kümesindeki her bir alt küme için gözlemler

$$x_i = uh_1(i) + (1-u)h_2(i) + \varepsilon_i \quad (1. \text{ Alt Küme}) \quad (8.7)$$

$$x_i = uh_1(i) + (1-u)h_3(i) + \varepsilon_i \quad (2. \text{ Alt Küme}) \quad (8.8)$$

$$x_i = uh_2(i) + (1-u)h_3(i) + \varepsilon_i \quad (3. \text{ Alt Küme}) \quad (8.9)$$

eşitlikleri ile elde edilir. Burada u , $(0,1)$ aralığında düzgün dağılıma sahip, ε_i $i = 1,2,...,21$ olmak üzere standart normal dağılıma sahip rassal değişkenlerdir. Eşitlik (8.7), (8.8) ve (8.9)'da yer alan h_1 , h_2 ve h_3 Waveform veri kümesine üçgensel şekli veren değişkenlerdir ve

$$h_1(i) = \max(6 - |i - 11|, 0), \quad i = 1,2,...,21 \quad (8.10)$$

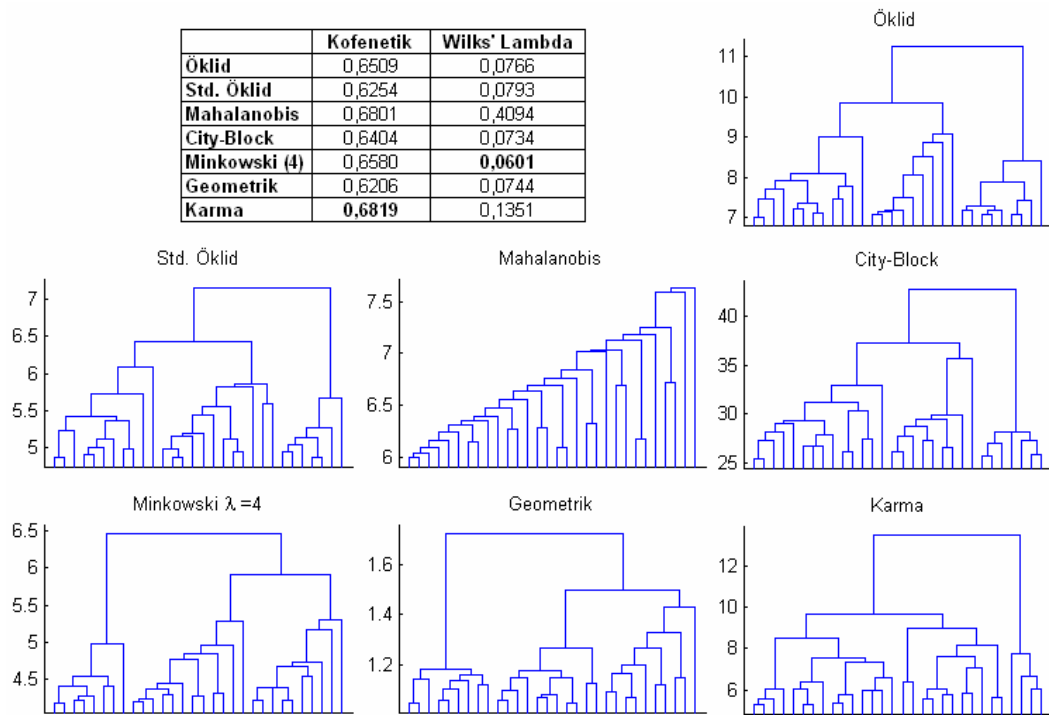
$$h_2(i) = h_1(i - 4), \quad i = 1,2,...,21 \quad (8.11)$$

$$h_3(i) = h_1(i + 4), \quad i = 1,2,...,21 \quad (8.12)$$

şeklinde tanımlanır. Simülasyon çalışmamızda her bir alt kümede eşit gözlem sayısı olacak şekilde sırasıyla 75, 150 ve 300 birimlik Waveform veri setleri oluşturulmuştur. Oluşturulan her veri seti tek bağlantı, tam bağlantı, ortalama bağlantı, ağırlıklı ortalama bağlantı, merkez bağlantı, medyan bağlantı ve Ward bağlantı kümeleme tekniklerinde ilgilenilen 7 farklı uzaklık ölçüsü ile kümelendirilmiştir. Elde edilen kümelemeler sonrası kofenetik korelasyon katsayısı ve Wilks'in lambda test istatistiği değerleri elde edilmiştir. Her bir örneklem

hacminde 100 tekrar yapılarak oluşturulan 300 Waveform veri setinin her biri farklı aşamalı kümele tekniği ve farklı uzaklık ölçüleri ile kümelendirilmiştir. Simülasyon çalışmasında 14700 adet kümeleme sonucu incelenmiştir.

Rassal olarak üretilen 300 birimlik Waveform simülasyon verisinin ortalama bağlantı kümeleme tekniğinde farklı uzaklık ölçüleri kullanılarak elde edilen ağaç grafikleri ile kofenetik korelasyon katsayısı ve Wilks'in lambda değerleri Şekil 8.4'de verilmiştir.



Şekil 8.4. Rassal olarak üretilen 300 birimlik Waveform simülasyon verisinin ortalama bağlantı kümeleme tekniğinde farklı uzaklık ölçülerine göre elde edilen ağaç grafikleri

Şekil 8.4 incelendiğinde Waveform simülasyon verisinin farklı uzaklık ölçüleri ile kümelmesi sonucu elde edilen ağaç grafiklerinin farklı yapılar da olduğu görülmektedir. Üç kümeye sahip olan Waveform simülasyon verisi için elde edilen ağaç grafikleri içersinden Mahalanobis uzaklığı ile elde edilen ağaç grafiğinin diğer uzaklık ölçüleri ile elde edilen ağaç grafiklerine göre yorumlanması oldukça güçtür. Dolayısıyla kümeleme analizinde uzaklık ölçüsünün seçiminin, kümelemenin geçerliliğinde önemli bir etken olduğu Şekil 8.4'de açıkça görülmektedir.

Simülasyon çalışmasında rassal olarak üretilen 75 birimlik üretilen Waveform veri setlerinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen kofenetik korelasyon katsayısı değerlerinin ortalama ve standart hataları Tablo 8.6’da verilmiştir.

Tablo 8.6. Wafeform simülasyon veri setlerinin ($n = 75$) farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen kofenetik korelasyon katsayısı değerlerinin ortalama ve standart hataları

Uzaklık	T.İstatistik	Tek	Tam	Ortalama	Ağırlıklı	Merkez	Medyan	Ward
Öklid	Ortalama	0.3904	0.6761	0.6938	0.6718	0.6787	0.5993	0.6599
	Std. Hata	0.0070	0.0047	0.0043	0.0045	0.0043	0.0096	0.0048
Std. Öklid	Ortalama	0.4118	0.6472	0.6736	0.6484	0.6277	0.5638	0.6233
	Std. Hata	0.0066	0.0045	0.0041	0.0045	0.0053	0.0088	0.0052
Mahalanobis	Ortalama	0.8601	0.8446	0.8895	0.8847	0.8657	0.8564	0.8487
	Std. Hata	0.0017	0.0012	0.0009	0.0009	0.0009	0.0010	0.0011
City-Block	Ortalama	0.3772	0.6635	0.6858	0.6638	0.6704	0.5915	0.6536
	Std. Hata	0.0072	0.0047	0.0043	0.0046	0.0043	0.0098	0.0045
Minkowski ($\lambda = 4$)	Ortalama	0.4072	0.6617	0.6845	0.6665	0.6639	0.5931	0.6461
	Std. Hata	0.0067	0.0051	0.0042	0.0048	0.0049	0.0087	0.0049
Geometrik	Ortalama	0.3571	0.6453	0.6685	0.6419	0.6476	0.5698	0.6321
	Std. Hata	0.0073	0.0049	0.0045	0.0046	0.0045	0.0106	0.0047
Karma	Ortalama	0.6876	0.6620	0.7525	0.6927	0.7487	0.6844	0.6576
	Std. Hata	0.0096	0.0085	0.0056	0.0082	0.0053	0.0084	0.0073

Tablo 8.6’deki ortalama kofenetik korelasyon değerleri incelendiğinde örneklem hacmi 75 birim olan Waveform veri setlerinin kümelenmesinde en başarılı uzaklık ölçüsü, tüm aşamalı kümeleme tekniklerinde Mahalanobis uzaklığı olduğu görülmektedir. Yeni tanımlanan karma uzaklık ölçüsü, Mahalanobis uzaklığından sonra örneklem hacmi 75 birim olan Waveform veri setlerinin kümelenmesinde en başarılı ikinci uzaklık ölçüsüdür.

Örneklem hacmi 150 birim olarak rassal üretilen Waveform veri setlerinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen kofenetik korelasyon katsayısı değerlerinin ortalama ve standart hataları Tablo 8.7’de verilmiştir.

Tablo 8.7. Wafeform simülasyon veri setlerinin ($n = 150$) farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen kofenetik korelasyon katsayısı değerlerinin ortalama ve standart hataları

Uzaklık	T.İstatistik	Tek	Tam	Ortalama	Ağırlıklı	Merkez	Medyan	Ward
Öklid	Ortalama	0.2961	0.6442	0.6777	0.6521	0.6685	0.5463	0.6507
	Std. Hata	0.0064	0.0037	0.0031	0.0033	0.0034	0.0103	0.0035
Std. Öklid	Ortalama	0.3386	0.6128	0.6509	0.6218	0.6181	0.4963	0.6061
	Std. Hata	0.0057	0.0033	0.0033	0.0037	0.0044	0.0104	0.0039
Mahalanobis	Ortalama	0.6990	0.5795	0.7483	0.7232	0.7332	0.6572	0.5817
	Std. Hata	0.0020	0.0025	0.0014	0.0016	0.0017	0.0028	0.0017
City-Block	Ortalama	0.2875	0.6373	0.6685	0.6459	0.6606	0.5384	0.6409
	Std. Hata	0.0065	0.0034	0.0031	0.0036	0.0033	0.0117	0.0037
Minkowski ($\lambda = 4$)	Ortalama	0.3097	0.6265	0.6658	0.6420	0.6530	0.5429	0.6334
	Std. Hata	0.0065	0.0042	0.0034	0.0038	0.0037	0.0105	0.0038
Geometrik	Ortalama	0.2843	0.6130	0.6496	0.6293	0.6404	0.5264	0.6232
	Std. Hata	0.0059	0.0038	0.0034	0.0036	0.0032	0.0103	0.0033
Karma	Ortalama	0.6454	0.6293	0.7167	0.6501	0.7206	0.6344	0.6134
	Std. Hata	0.0082	0.0085	0.0055	0.0075	0.0052	0.0089	0.0071

Tablo 8.7'deki ortalama kofenetik korelasyon değerleri incelendiğinde örneklem hacmi 150 birim olan Waveform veri setlerinin kümelenmesinde tam bağlantı ve Ward bağlantı aşamalı kümeleme tekniklerinde en başarılı uzaklık ölçüsü Öklid uzaklığı iken, diğer aşamalı kümeleme tekniklerinde en başarılı uzaklık ölçüsü Mahalanobis uzaklığıdır. Yeni tanımlanan karma uzaklık ölçüsü örneklem hacmi 150 olduğunda da, Waveform veri setlerinin kümelenmesinde en başarılı ikinci uzaklık ölçüsüdür.

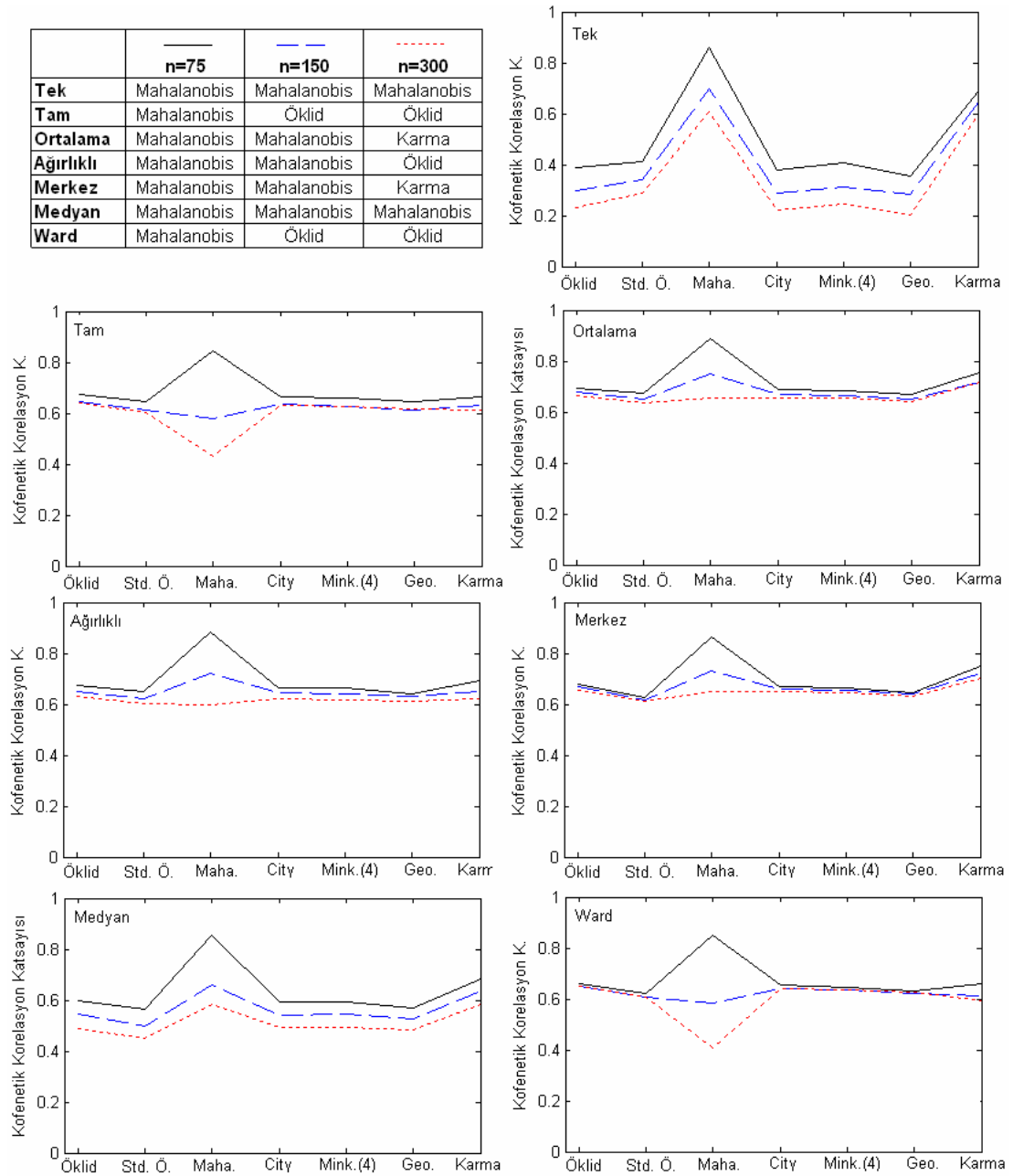
Örneklem hacmi 300 birim olarak rassal üretilen Waveform veri setlerinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen kofenetik korelasyon katsayısı değerlerinin ortalama ve standart hataları Tablo 8.8'de verilmiştir.

Tablo 8.8. Wafeform simülasyon veri setlerinin ($n = 300$) farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen kofenetik korelasyon katsayısı değerlerinin ortalama ve standart hataları

Uzaklık	T.İstatistik	Tek	Tam	Ortalama	Ağırlıklı	Merkez	Medyan	Ward
Öklid	Ortalama	0.2295	0.6410	0.6627	0.6319	0.6550	0.4866	0.6484
	Std. Hata	0.0049	0.0028	0.0027	0.0034	0.0024	0.0109	0.0025
Std. Öklid	Ortalama	0.2878	0.6037	0.6361	0.6028	0.6113	0.4523	0.6056
	Std. Hata	0.0044	0.0028	0.0027	0.0036	0.0030	0.0102	0.0026
Mahalanobis	Ortalama	0.6074	0.4293	0.6531	0.5973	0.6518	0.5830	0.4051
	Std. Hata	0.0025	0.0032	0.0021	0.0032	0.0024	0.0030	0.0016
City-Block	Ortalama	0.2198	0.6321	0.6544	0.6197	0.6480	0.4909	0.6401
	Std. Hata	0.0050	0.0026	0.0027	0.0044	0.0025	0.0108	0.0025
Minkowski ($\lambda = 4$)	Ortalama	0.2451	0.6246	0.6525	0.6177	0.6432	0.4946	0.6360
	Std. Hata	0.0046	0.0027	0.0027	0.0037	0.0026	0.0098	0.0025
Geometrik	Ortalama	0.2013	0.6155	0.6387	0.6099	0.6302	0.4838	0.6245
	Std. Hata	0.0046	0.0026	0.0027	0.0035	0.0024	0.0107	0.0024
Karma	Ortalama	0.5970	0.6096	0.7146	0.6193	0.7047	0.5826	0.5945
	Std. Hata	0.0072	0.0080	0.0048	0.0060	0.0052	0.0088	0.0076

Tablo 8.7'deki ortalama kofenetik korelasyon değerleri incelendiğinde örneklem hacmi 300 birim olan Waveform veri setlerinin kümelenmesinde tam, ağırlıklı ortalama ve Ward bağlantı tekniklerinde en başarılı uzaklık ölçüsü Öklid uzaklığı, tek ve medyan bağlantı tekniklerinde en başarılı uzaklık ölçüsü Mahalanobis uzaklığı, ortalama ve merkez bağlantı tekniklerinde ise en başarılı uzaklık ölçüsü karma uzaklık ölçüsüdür.

Rassal olarak farklı örneklem hacimlerinde üretilen Waveform simülasyon verilerinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçülerine göre kümelenmesi ile elde edilen ortalama kofenetik korelasyon katsayısına göre grafiksel olarak değerlendirilmesi Şekil 8.5'de verilmiştir.



Şekil 8.5. Rassal olarak farklı örneklem hacimlerinde üretilen Waveform simülasyon verilerinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçülerine göre kümelenmesi ile elde edilen ortalama kofenetik korelasyon katsayısına göre değerlendirilmesi

Şekil 8.5 incelendiğinde kofenetik korelasyon katsayısına göre Waveform veri setlerinin kümelenmesindeki en başarısız aşamalı kümeleme tekniği, tek bağlantı kümeleme tekniğidir.

Simülasyon çalışmasında rassal olarak üretilen 75 birimlik üretilen Waveform veri setlerinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen Wilks'in lambda test istatistiği değerlerinin ortalama ve standart hataları Tablo 8.9'da verilmiştir.

Tablo 8.9. Wafeform simülasyon veri setlerinin ($n = 75$) farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen Wilks'in lambda test istatistiği değerlerinin ortalama ve standart hataları

Uzaklık	T.İstatistik	Tek	Tam	Ortalama	Ağırlıklı	Merkez	Medyan	Ward
Öklid	Ortalama	0.1927	0.0345	0.0273	0.0336	0.0702	0.0838	0.030370
	Std. Hata	0.0121	0.0016	0.0013	0.0020	0.0080	0.0069	0.001502
Std. Öklid	Ortalama	0.1487	0.0341	0.0254	0.0333	0.1298	0.1011	0.029420
	Std. Hata	0.0106	0.0017	0.0014	0.0020	0.0124	0.0087	0.001349
Mahalanobis	Ortalama	0.0035	0.0028	0.0026	0.0021	0.0019	0.0027	0.000020
	Std. Hata	0.0003	0.0003	0.0002	0.0002	0.0001	0.0002	0.000002
City-Block	Ortalama	0.2063	0.0361	0.0300	0.0338	0.0586	0.0737	0.030410
	Std. Hata	0.0132	0.0017	0.0014	0.0019	0.0052	0.0056	0.001504
Minkowski ($\lambda = 4$)	Ortalama	0.2042	0.0342	0.0284	0.0345	0.0776	0.0869	0.029980
	Std. Hata	0.0135	0.0017	0.0016	0.0020	0.0082	0.0069	0.001496
Geometrik	Ortalama	0.2194	0.0372	0.0290	0.0397	0.0687	0.0882	0.033290
	Std. Hata	0.0128	0.0016	0.0013	0.0027	0.0066	0.0083	0.001569
Karma	Ortalama	0.0438	0.0153	0.0102	0.0120	0.0166	0.0207	0.009300
	Std. Hata	0.0055	0.0019	0.0017	0.0013	0.0028	0.0029	0.001289

Tablo 8.9'daki ortalama Wilks'in lambda test istatistiği değerleri incelendiğinde örneklem hacmi 75 birim olan Waveform veri setlerinin kümelenmesinde en başarılı uzaklık ölçüsü kofenetik korelasyon katsayısındaki gibi tüm aşamalı kümeleme tekniklerinde Mahalanobis uzaklığı olmuştur. Wilks'in lambda test istatistiğine göre de yeni tanımlanan karma uzaklık ölçüsü, Mahalanobis uzaklığından sonra en başarılı ikinci uzaklık ölçüsü olmuştur.

Simülasyon çalışmasında rassal olarak üretilen 150 birimlik üretilen Waveform veri setlerinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen Wilks'in lambda test istatistiği değerlerinin ortalama ve standart hataları Tablo 8.10'da verilmiştir.

Tablo 8.10. Wafeform simülasyon veri setlerinin ($n = 150$) farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen Wilks'in lambda test istatistiği değerlerinin ortalama ve standart hataları

Uzaklık	T.İstatistik	Tek	Tam	Ortalama	Ağırlıklı	Merkez	Medyan	Ward
Öklid	Ortalama	0.4402	0.0710	0.0567	0.0721	0.1243	0.2047	0.0587
	Std. Hata	0.0118	0.0025	0.0024	0.0033	0.0118	0.0154	0.0019
Std. Öklid	Ortalama	0.4085	0.0701	0.0680	0.0860	0.3287	0.2477	0.0595
	Std. Hata	0.0119	0.0024	0.0036	0.0053	0.0191	0.0158	0.0018
Mahalanobis	Ortalama	0.1662	0.0776	0.1394	0.1201	0.1541	0.1668	0.0240
	Std. Hata	0.0033	0.0023	0.0032	0.0031	0.0024	0.0032	0.0005
City-Block	Ortalama	0.4718	0.0713	0.0589	0.0737	0.1061	0.1891	0.0587
	Std. Hata	0.0110	0.0021	0.0025	0.0038	0.0078	0.0130	0.0018
Minkowski ($\lambda = 4$)	Ortalama	0.4479	0.0691	0.0591	0.0776	0.1712	0.1995	0.0579
	Std. Hata	0.0116	0.0025	0.0027	0.0037	0.0159	0.0144	0.0018
Geometrik	Ortalama	0.5098	0.0725	0.0594	0.0787	0.1261	0.2069	0.0606
	Std. Hata	0.0122	0.0023	0.0022	0.0035	0.0098	0.0138	0.0018
Karma	Ortalama	0.2392	0.0820	0.1175	0.0988	0.1614	0.1144	0.0696
	Std. Hata	0.0089	0.0031	0.0066	0.0047	0.0091	0.0060	0.0025

Tablo 8.10'daki ortalama Wilks'in lambda test istatistiği değerleri incelendiğinde örneklem hacmi 150 birim olan Waveform veri setlerinin kümelenmesinde ortalama ve ağırlıklı ortalama bağlantı tekniklerinde en başarılı uzaklık ölçüsü Öklid uzaklığı, tek ve Ward bağlantı tekniklerinde en başarılı uzaklık ölçüsü Mahalanobis uzaklığı, tam bağlantı tekniğinde en başarılı uzaklık ölçüsü Minkowski ($\lambda = 4$) uzaklığı, merkez bağlantı tekniğinde en başarılı uzaklık ölçüsü City-Block uzaklığı ve medyan bağlantı tekniğinde en başarılı uzaklık ölçüsü karma uzaklık ölçüsüdür.

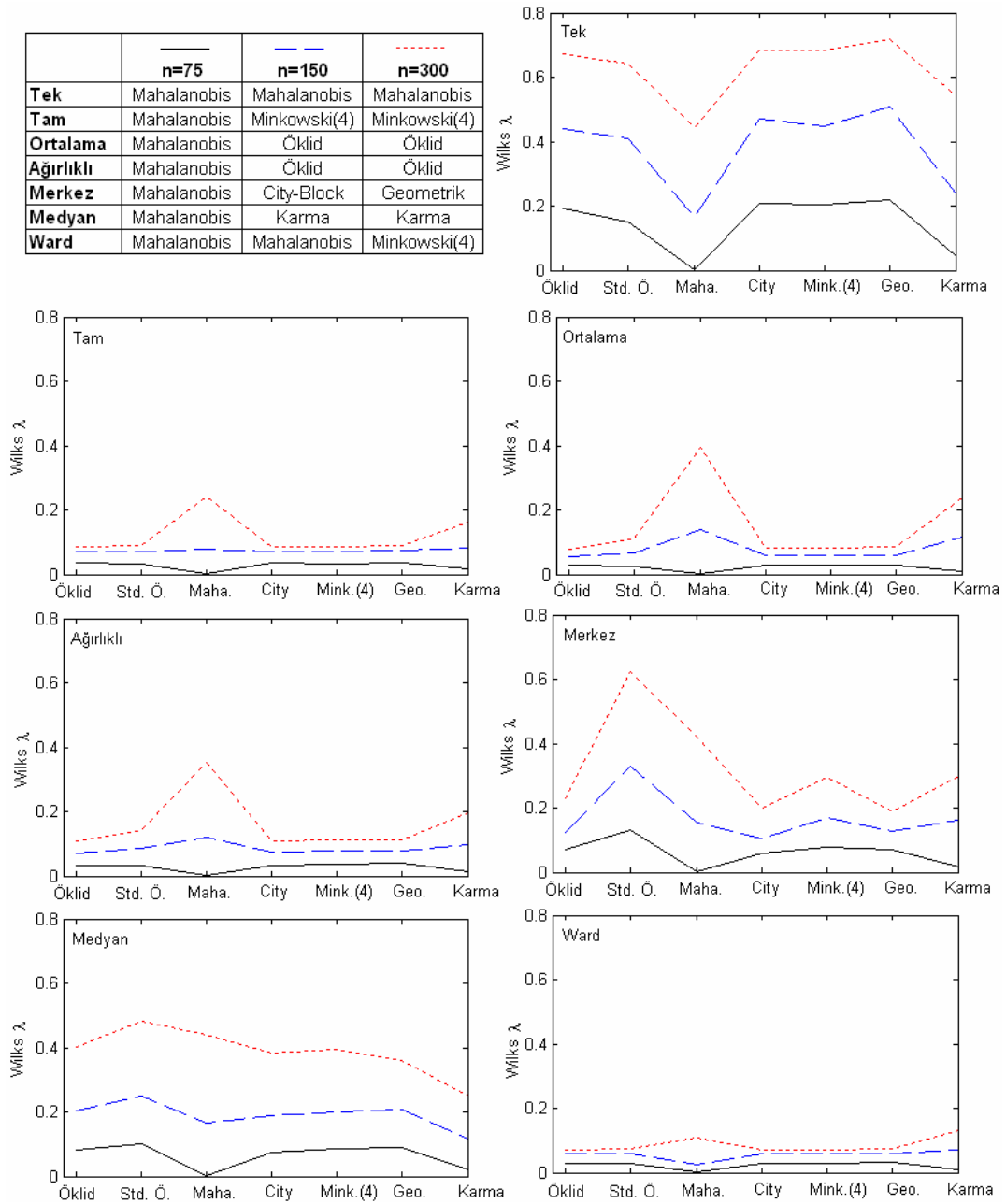
Örneklem hacmi 300 birim olarak rassal üretilen Waveform veri setlerinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen Wilks'in lambda test istatistiği değerlerinin ortalama ve standart hataları Tablo 8.11'de verilmiştir.

Tablo 8.11. Wafeform simülasyon veri setlerinin ($n = 300$) farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen Wilks'in lambda test istatistiği değerlerinin ortalama ve standart hataları

Uzaklık	T. İstatistik	Tek	Tam	Ortalama	Ağırlıklı	Merkez	Medyan	Ward
Öklid	Ortalama	0.6713	0.0852	0.0785	0.1070	0.2298	0.4023	0.0718
	Std. Hata	0.0076	0.0023	0.0028	0.0054	0.0203	0.0215	0.0014
Std. Öklid	Ortalama	0.6424	0.0884	0.1139	0.1445	0.6242	0.4832	0.0740
	Std. Hata	0.0084	0.0022	0.0067	0.0100	0.0171	0.0194	0.0014
Mahalanobis	Ortalama	0.4455	0.2406	0.3958	0.3531	0.4195	0.4385	0.1078
	Std. Hata	0.0040	0.0048	0.0049	0.0062	0.0033	0.0042	0.0012
City-Block	Ortalama	0.6836	0.0857	0.0825	0.1089	0.1996	0.3827	0.0721
	Std. Hata	0.0083	0.0022	0.0031	0.0065	0.0172	0.0214	0.0014
Minkowski ($\lambda = 4$)	Ortalama	0.6855	0.0851	0.0835	0.1128	0.2956	0.3951	0.0699
	Std. Hata	0.0085	0.0023	0.0035	0.0061	0.0228	0.0214	0.0014
Geometrik	Ortalama	0.7166	0.0887	0.0847	0.1119	0.1896	0.3619	0.0749
	Std. Hata	0.0074	0.0023	0.0033	0.0051	0.0135	0.0211	0.0014
Karma	Ortalama	0.5426	0.1637	0.2390	0.1951	0.2980	0.2507	0.1314
	Std. Hata	0.0102	0.0046	0.0119	0.0065	0.0155	0.0122	0.0025

Tablo 8.11'deki ortalama Wilks'in lambda test istatistiği değerleri incelendiğinde örneklem hacmi 300 birim olan Waveform veri setlerinin kümelenmesinde ortalama ve ağırlıklı ortalama bağlantı tekniklerinde en başarılı uzaklık ölçüsü Öklid uzaklığı, tek bağlantı tekniğinde en başarılı uzaklık ölçüsü Mahalanobis uzaklığı, tam ve Ward bağlantı tekniklerinde en başarılı uzaklık ölçüsü Minkowski ($\lambda = 4$) uzaklığı, merkez bağlantı tekniğinde en başarılı uzaklık ölçüsü geometrik uzaklık ve medyan bağlantı tekniğinde en başarılı uzaklık ölçüsü karma uzaklık ölçüsüdür.

Rassal olarak farklı örneklem hacimlerinde üretilen Waveform simülasyon verilerinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçülerine göre kümelenmesi ile elde edilen ortalama Wilks'in lambda test istatistiği değerlerine göre grafiksel olarak değerlendirilmesi Şekil 8.6'da verilmiştir.



Şekil 8.6. Rassal olarak farklı örneklem hacimlerinde üretilen Waveform simülasyon verilerinin farklı aşamalı kümeleme tekniklerinde farklı uzaklık ölçülerine göre kümelenmesi ile elde edilen ortalama Wilks'in lambda test istatistiğine göre değerlendirilmesi

Şekil 8.6 incelendiğinde Wilks'in lambda test istatistiği değerlerine göre Waveform veri setlerinin kümelenmesindeki en başarılı aşamalı kümeleme tekniği, Ward bağlantı kümeleme tekniği, en başarısız aşamalı kümeleme tekniği tek bağlantı kümeleme tekniğidir.

8.2. Aşamalı Olmayan Kümeleme Tekniklerinde Uzaklık Ölçülerinin Etkisi

Aşamalı olmayan kümeleme teknikleri içersinde basitliği ve etkinliği açısından en yaygın kullanıma sahip olan k -ortalamlar tekniğinde birimlerin küme merkezlerine olan uzaklığın hesaplanmasında seçilen uzaklık ölçüsünün kümelemeye olan etkisi Tablo 8.1'de yer alan gerçek veri setleri ve çok değişkenli normal dağılımdan üretilecek olan simülasyon verileri üzerinde incelenecektir. Seçilen uzaklık ölçülerinin kümelemeye etkisi Wilks'in lambda test istatistiği ile değerlendirilecektir.

k -ortalamlar tekniğinde seçilen uzaklık ölçülerinin etkisini yalın olarak ortaya koyabilmek için, küme başlangıç merkezleri ve küme sayıları her uzaklık ölçüsü için sabit tutulmuştur. Böylece k -ortalamlar tekniğinde küme başlangıç merkezlerinin seçimi ve küme sayısının kümelemeye olan etkileri her uzaklık ölçüsü için sabitlenmiştir.

8.2.1. k – Ortalamalar Tekniğinde Uzaklık Ölçülerinin Etkisinin Gerçek Veri Setleri Üzerinde İncelenmesi

k -ortalamlar tekniğinde farklı uzaklık ölçülerinin etkisinin gerçek veri setleri üzerinde incelemesinde, bir önceki kısımdan farklı olarak Araç, Harf ve Maya veri setleri gerçek örneklem büyüklüklerinde çalışmaya dahil edilmiştir.

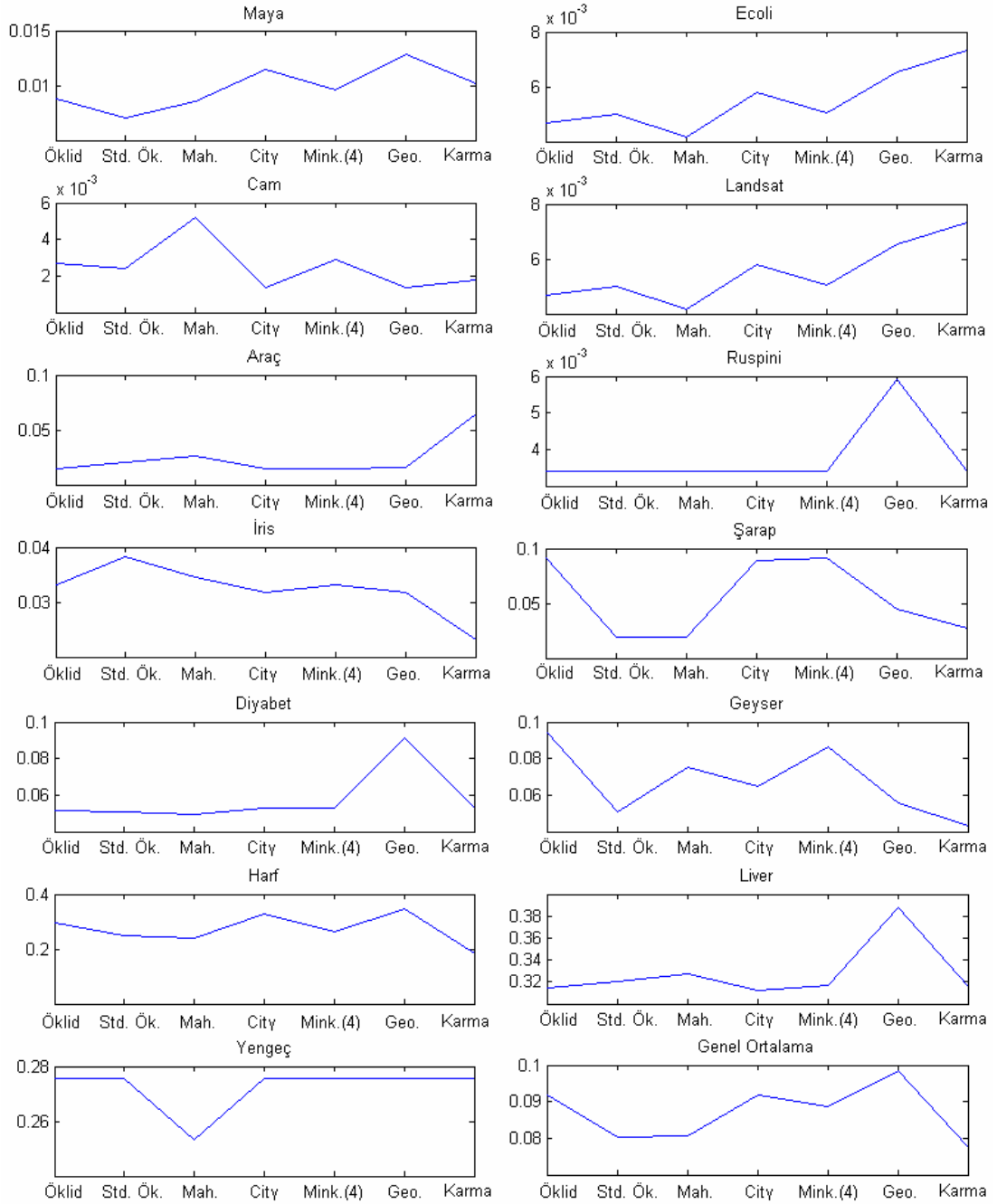
Farklı yapılarıdaki gerçek veri setlerinin k -ortalamlar tekniğinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen Wilks'in lambda test istatistiği değerleri ve genel değerlendirme için hesaplanan ortalama Wilks'in lambda test istatistiği değerleri ve standart hataları Tablo 8.12'de verilmiştir.

Tablo 8.12. Farklı yapılarıdaki gerçek veri setlerinin k -ortalamalar tekniğinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen Wilks'in lambda test istatistiği değerleri

Veri Seti	Öklid	Std. Ök	Maha	City	Min(4)	Geo	Karma
Maya(10)	0.00881	0.00705	0.0086	0.01144	0.00954	0.01288	0.01017
Ecoli(8)	0.00469	0.00502	0.00418	0.00579	0.00503	0.00653	0.00732
Cam(6)	0.00267	0.00239	0.00523	0.00138	0.00293	0.00138	0.00179
Landsat(6)	0.00002	0.00001	0.00013	0.00025	0.00003	0.00022	0.00001
Araç(4)	0.01500	0.02066	0.02568	0.01431	0.01478	0.01538	0.06500
Ruspini(4)	0.00339	0.00339	0.00339	0.00339	0.00339	0.00590	0.00339
İris(3)	0.03318	0.03825	0.03449	0.03184	0.03317	0.03184	0.02314
Şarap(3)	0.09128	0.01958	0.01959	0.08952	0.09128	0.04465	0.02710
Diyabet(3)	0.05155	0.05104	0.04976	0.05295	0.05273	0.09128	0.05306
Geyser(3)	0.09403	0.05088	0.07536	0.06459	0.08613	0.05577	0.04285
Harf(2)	0.29673	0.24748	0.24035	0.33018	0.26135	0.34838	0.18289
Liver(2)	0.31426	0.32069	0.32676	0.31216	0.31685	0.38782	0.31544
Yengeç(2)	0.27578	0.27578	0.25315	0.27578	0.27578	0.27578	0.27578
Genel	Öklid	Std. Ök.	Maha	City	Min(4)	Geo	Karma
Ortalama	0.09165	0.08017	0.08051	0.09181	0.08869	0.09829	0.07753
Std. Hata	0.03349	0.03244	0.03151	0.03484	0.03227	0.03899	0.03008

Tablo 8.12'de incelemeye dahil edilen veri setlerinin farklı uzaklık ölçüleri ile kümelenmesi sonrasında elde edilen Wilks'in lambda test istatistiği değerlerinin ortalaması incelendiğinde en başarılı uzaklık ölçüsü yeni önerilen karma uzaklık ölçüsü olmuştur. Ayrıca karma uzaklık ölçüsü, Landsat, Geyser ve Harf veri setlerinin kümelenmesinde en başarılı bulunan uzaklık ölçüsüdür. Ortalama Wilks'in lambda değerlerine göre karma uzaklık ölçüsünden sonraki en başarılı uzaklık ölçüleri standartlaştırılmış Öklid ve Mahalanobis uzaklık ölçüleridir. Cam ve İris veri setlerinin kümelenmesinde en başarılı uzaklık ölçüsü olan geometrik uzaklık ölçüsü ortalama Wilks'in lambda değerine göre en başarısız uzaklık ölçüsü olmuştur.

Farklı yapılarıdaki gerçek veri setlerinin k -ortalamalar tekniğinde farklı uzaklık ölçülerine göre Wilks'in lambda test istatistiğinin değişimini gösteren çizgi grafikleri Şekil 8.7'de verilmiştir.



Şekil 8.7. Farklı yapılarıdaki gerçek veri setlerinin k -ortalamalar tekniğinde farklı uzaklık ölçüleri kullanılarak kümelenmesi sonrası elde edilen Wilks'in lambda test istatistiğine göre değerlendirilmesi

Şekil 8.7 incelendiğinde geometrik uzaklık ölçüsünün Wilks'in lambda test istatistiğine göre en kötü performans gösterdiği veri setleri Maya, Diyabet, Ruspini ve Liver veri setleridir.

8.2.2. k – Ortalamalar Tekniğinde Uzaklık Ölçülerinin Etkisinin Simülasyonla İncelenmesi

Bu başlık altında, k -ortalamalar tekniğinde farklı uzaklık ölçülerinin kümelemeye etkisi simülasyon çalışması ile ortaya konmaya çalışılacaktır. Bu amaçla gerçekleştirilecek simülasyon çalışmasında değişken sayısı 2 farklı, küme sayısı 3 farklı, örneklem hacmi 3 farklı ve kovaryans matris yapısı 4 farklı şekilde alınarak çok değişkenli normal dağılımdan rassal olarak örneklemeler üretilecektir. İlgilenilen 7 farklı uzaklık ölçüsü ile 72 farklı koşulda 1000'er tekrarlar üretilen 72000 verinin k -ortalamalar tekniğiyle kümeleneceği ile 504000 kümeleme sonucu incelenecektir.

Simülasyon çalışmasında değişken sayısı p sırasıyla 4 ve 6, küme sayısı k sırasıyla 2, 3 ve 4, örneklem hacmi n 100, 200 ve 500 alınacaktır. Değişken sayısının 4 olduğu simülasyon çalışmasında üretilecek rassal örneklemeler için ortalama vektörü $\mu = [1 \ 3 \ 5 \ 7]$ iken değişken sayısının 6 olduğu durumda ortalama vektörü $\mu = [1 \ 10 \ 5 \ 1000 \ 7 \ 0.5]$ olarak alınacaktır. Simülasyon çalışmasında farklı değişken sayısı, farklı küme sayısı ve farklı örneklem büyüklüklerinde uzaklık ölçülerinin kümelemeye etkisi incelenirken ayrıca dört farklı kovaryans matris yapısı göz önünde bulundurulacaktır. Rassal örneklemelerin çok değişkenli normal dağılımdan üretilmesi esnasında kovaryans matrisinin $\Sigma = \mathbf{I}$ olması Durum I, kovaryans matrisinin $\Sigma = \sigma_i^2 \mathbf{I}$ olması Durum II, kovaryans matrisinde farklı varyansa sahip değişkenler arasındaki korelasyonların aynı olması Durum III ve kovaryans matrisinde farklı varyansa sahip değişkenler arasındaki korelasyonların farklı olması Durum IV olarak ele alınacaktır. Simülasyon çalışmasında çok değişkenli normal dağılımdan rassal olarak üretilecek örneklemelerin oluşturulmasında kullanılacak kovaryans matrisleri Tablo 8.13'de verilmiştir.

Tablo 8.13. Simülasyon çalışmasında çok değişkenli normal dağılımdan rassal olarak üretilecek örneklemelerin oluşturulmasında kullanılacak kovaryans matrisleri

DURUM I	$\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$	$\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$
DURUM II	$\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 16 \end{bmatrix}$	$\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 36 & 0 & 0 & 0 & 0 \\ 0 & 0 & 16 & 0 & 0 & 0 \\ 0 & 0 & 0 & 100 & 0 & 0 \\ 0 & 0 & 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$
DURUM III	$\Sigma = \begin{bmatrix} 1 & 1 & 1.5 & 2 \\ 1 & 4 & 3 & 4 \\ 1.5 & 3 & 9 & 6 \\ 2 & 4 & 6 & 16 \end{bmatrix}$	$\Sigma = \begin{bmatrix} 1 & 1.8 & 1.2 & 3 & 0.9 & 0.3 \\ 1.8 & 36 & 7.2 & 18 & 5.4 & 1.8 \\ 1.2 & 7.2 & 16 & 12 & 3.6 & 1.2 \\ 3 & 18 & 12 & 100 & 9 & 3 \\ 0.9 & 5.4 & 3.6 & 9 & 9 & 0.9 \\ 0.3 & 1.8 & 1.2 & 3 & 0.9 & 1 \end{bmatrix}$
DURUM IV	$\Sigma = \begin{bmatrix} 1 & 1 & 2.4 & 0.8 \\ 1 & 4 & 2.4 & 5.6 \\ 2.4 & 2.4 & 9 & 6 \\ 0.8 & 5.6 & 6 & 16 \end{bmatrix}$	$\Sigma = \begin{bmatrix} 1 & 4.8 & 1.2 & 6 & 0.3 & 0.3 \\ 4.8 & 36 & 2.4 & 48 & 9 & 1.2 \\ 1.2 & 2.4 & 16 & 8 & 4.8 & 2 \\ 6 & 48 & 8 & 100 & 18 & 2 \\ 0.3 & 9 & 4.8 & 18 & 9 & 2.4 \\ 0.3 & 1.2 & 2 & 2 & 2.4 & 1 \end{bmatrix}$

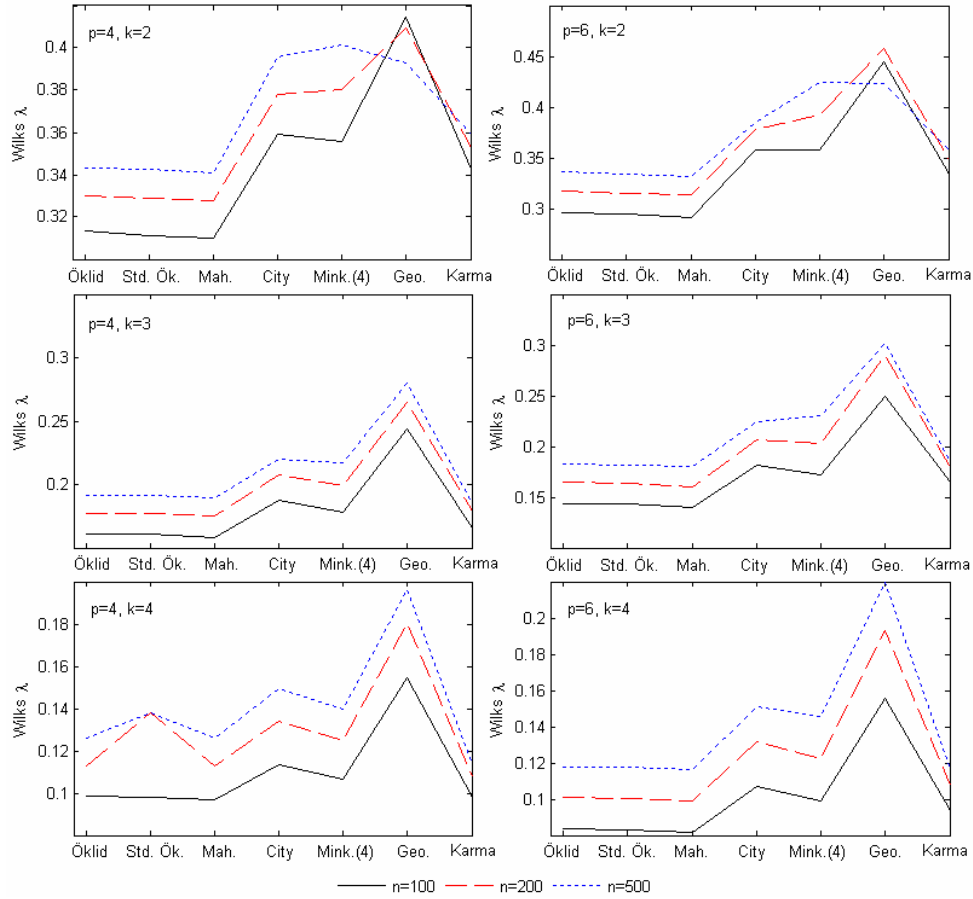
İlgilenilen 7 farklı uzaklık ölçüsü ile kovaryans matrisinin birim matrisi olması durumunda 18 farklı koşulda üretilen 1000'er veri setinin kümelenmesi sonucu elde edilen Wilks'in lambda test istatistiği değerlerine ait ortalama ve standart hatalar Tablo 8.14'de verilmiştir.

Tablo 8.14. Durum I için ilgilenilen 7 uzaklık ölçüsü ile elde edilen Wilks'in lambda test istatistiği değerlerine ait ortalama ve standart hatalar

p	k	n	T. İstatistik	Öklid	Std. Ök.	Maha.	City	Mink.(4)	Geo.	Karma
4	2	100	Ortalama	0.3137	0.3110	0.3102	0.3590	0.3556	0.4147	0.3425
			S.Hata	0.0010	0.0010	0.0010	0.0016	0.0013	0.0026	0.0010
		200	Ortalama	0.3301	0.3288	0.3278	0.3782	0.3804	0.4097	0.3531
			S.Hata	0.0007	0.0007	0.0008	0.0016	0.0009	0.0025	0.0008
		500	Ortalama	0.3429	0.3425	0.3410	0.3956	0.4013	0.3926	0.3592
			S.Hata	0.0005	0.0005	0.0005	0.0017	0.0006	0.0022	0.0005
	3	100	Ortalama	0.1609	0.1608	0.1584	0.1878	0.1783	0.2438	0.1657
			S.Hata	0.0007	0.0007	0.0007	0.0009	0.0008	0.0014	0.0007
		200	Ortalama	0.1776	0.1773	0.1753	0.2074	0.1992	0.2651	0.1786
			S.Hata	0.0005	0.0005	0.0005	0.0008	0.0006	0.0012	0.0005
		500	Ortalama	0.1913	0.1910	0.1900	0.2201	0.2175	0.2803	0.1850
			S.Hata	0.0003	0.0003	0.0003	0.0006	0.0004	0.0009	0.0003
	4	100	Ortalama	0.0986	0.0980	0.0968	0.1136	0.1066	0.1548	0.0978
			S.Hata	0.0005	0.0005	0.0005	0.0006	0.0005	0.0010	0.0005
		200	Ortalama	0.1129	0.1380	0.1127	0.1341	0.1248	0.1804	0.1077
			S.Hata	0.0003	0.0011	0.0004	0.0005	0.0004	0.0009	0.0004
		500	Ortalama	0.1260	0.1382	0.1258	0.1497	0.1398	0.1965	0.1132
			S.Hata	0.0002	0.0005	0.0002	0.0004	0.0003	0.0007	0.0003
6	2	100	Ortalama	0.2963	0.2951	0.2912	0.3582	0.3576	0.4444	0.3342
			S.Hata	0.0010	0.0010	0.0010	0.0016	0.0013	0.0024	0.0010
		200	Ortalama	0.3173	0.3151	0.3135	0.3776	0.3924	0.4582	0.3482
			S.Hata	0.0007	0.0007	0.0007	0.0014	0.0010	0.0026	0.0008
		500	Ortalama	0.3360	0.3342	0.3320	0.3854	0.4241	0.4237	0.3573
			S.Hata	0.0004	0.0004	0.0005	0.0015	0.0007	0.0027	0.0005
	3	100	Ortalama	0.1436	0.1430	0.1402	0.1820	0.1725	0.2494	0.1653
			S.Hata	0.0006	0.0006	0.0006	0.0009	0.0008	0.0015	0.0008
		200	Ortalama	0.1645	0.1637	0.1605	0.2069	0.2031	0.2895	0.1786
			S.Hata	0.0004	0.0005	0.0005	0.0008	0.0006	0.0013	0.0006
		500	Ortalama	0.1826	0.1817	0.1799	0.2244	0.2298	0.3014	0.1854
			S.Hata	0.0003	0.0003	0.0003	0.0006	0.0005	0.0011	0.0004
	4	100	Ortalama	0.0836	0.0831	0.0815	0.1068	0.0989	0.1555	0.0934
			S.Hata	0.0004	0.0004	0.0004	0.0006	0.0005	0.0010	0.0005
		200	Ortalama	0.1010	0.1006	0.0990	0.1318	0.1226	0.1931	0.1076
			S.Hata	0.0003	0.0003	0.0004	0.0005	0.0004	0.0009	0.0004
		500	Ortalama	0.1179	0.1174	0.1161	0.1513	0.1456	0.2197	0.1163
			S.Hata	0.0002	0.0002	0.0002	0.0004	0.0003	0.0008	0.0002

Tablo 8.14 incelendiğinde değişken sayısı, küme sayısı veya örneklem hacminin değişmesi durumunda ortalama Wilks'in lambda test istatistiğine göre uzaklık ölçülerinin kümeleme performansının değiştiği görülmektedir. Göz önünde tutulan 18 farklı koşuldan 12'sinde Mahalanobis uzaklığı, 3'ünde karma uzaklık,

2'sinde standartlaştırılmış Öklid uzaklığı ve 1'inde Öklid uzaklığı en başarılı uzaklık ölçüsü olmuştur. Tablo 8.14'de verilen ortalama Wilks'in lambda değerlerinin uzaklık ölçüsüne göre değişimini gösteren grafikler Şekil 8.8'de verilmiştir.



Şekil 8.8. Durum I'de ilgilenilen 7 farklı uzaklık ölçüsüne göre elde edilen ortalama Wilks'in lambda değerlerinin değişimi

Şekil 8.8'de yer alan grafikler incelendiğinde çok değişkenli normal dağılıma uyan ve kovaryans matrisi birim matris olan veri setlerinin k -ortalamlar tekniği ile kümelendiğinde Mahalanobis uzaklığı en başarılı uzaklık ölçüsüdür.

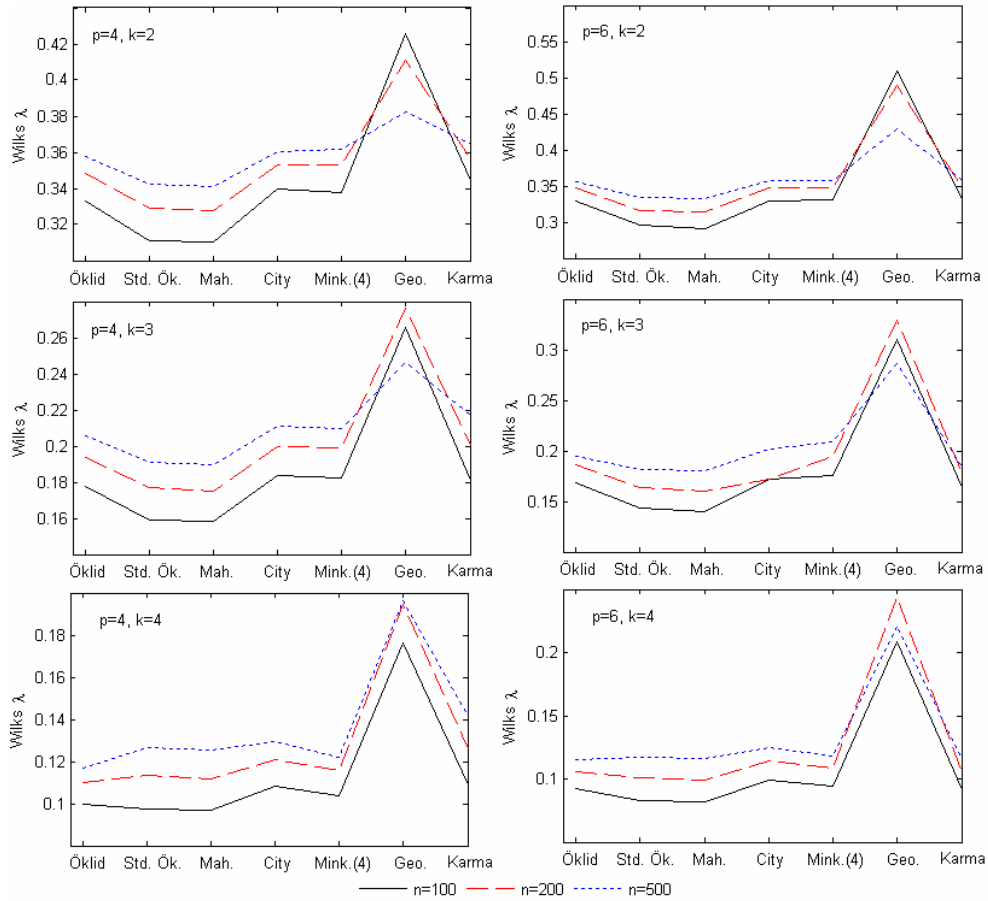
Durum II'de ilgilenilen 7 farklı uzaklık ölçüsü 18 farklı koşulda üretilen 1000'er veri setinin kümelendiği sonucu elde edilen Wilks'in lambda test istatistiği değerlerine ait ortalama ve standart hatalar Tablo 8.15'de verilmiştir.

Tablo 8.15. Durum II için ilgilenilen 7 uzaklık ölçüsü ile elde edilen Wilks'in lambda test istatistiği değerlerine ait ortalama ve standart hatalar

<i>p</i>	<i>k</i>	<i>n</i>	T. İstatistik	Öklid	Std. Ök.	Maha.	City	Mink.(4)	Geo.	Karma
4	2	100	Ortalama	0.3328	0.3110	0.3102	0.3399	0.3378	0.4258	0.3442
			S.Hata	0.0010	0.0010	0.0010	0.0011	0.0011	0.0032	0.0012
		200	Ortalama	0.3484	0.3288	0.3278	0.3527	0.3527	0.4107	0.3566
			S.Hata	0.0008	0.0007	0.0008	0.0008	0.0008	0.0028	0.0009
		500	Ortalama	0.3579	0.3425	0.3410	0.3602	0.3615	0.3822	0.3644
			S.Hata	0.0005	0.0005	0.0005	0.0005	0.0005	0.0017	0.0007
	3	100	Ortalama	0.1775	0.1592	0.1582	0.1835	0.1824	0.2657	0.1809
			S.Hata	0.0007	0.0007	0.0007	0.0008	0.0008	0.0020	0.0008
		200	Ortalama	0.1935	0.1773	0.1753	0.1999	0.1988	0.2761	0.2006
			S.Hata	0.0005	0.0005	0.0005	0.0007	0.0006	0.0019	0.0006
		500	Ortalama	0.2060	0.1910	0.1900	0.2110	0.2099	0.2464	0.2173
			S.Hata	0.0004	0.0003	0.0003	0.0005	0.0004	0.0020	0.0004
	4	100	Ortalama	0.0996	0.0974	0.0966	0.1083	0.1037	0.1765	0.1088
			S.Hata	0.0005	0.0005	0.0005	0.0006	0.0005	0.0014	0.0005
		200	Ortalama	0.1101	0.1134	0.1115	0.1206	0.1158	0.1947	0.1259
			S.Hata	0.0004	0.0004	0.0004	0.0005	0.0005	0.0013	0.0004
		500	Ortalama	0.1171	0.1263	0.1255	0.1294	0.1222	0.1963	0.1412
			S.Hata	0.0002	0.0002	0.0003	0.0004	0.0004	0.0013	0.0003
6	2	100	Ortalama	0.3295	0.2951	0.2912	0.3299	0.3309	0.5093	0.3324
			S.Hata	0.0010	0.0010	0.0010	0.0011	0.0010	0.0036	0.0010
		200	Ortalama	0.3476	0.3151	0.3135	0.3478	0.3482	0.4885	0.3472
			S.Hata	0.0008	0.0007	0.0007	0.0008	0.0008	0.0037	0.0008
		500	Ortalama	0.3565	0.3342	0.3320	0.3572	0.3570	0.4292	0.3568
			S.Hata	0.0005	0.0004	0.0005	0.0005	0.0005	0.0029	0.0005
	3	100	Ortalama	0.1681	0.1430	0.1402	0.1724	0.1754	0.3098	0.1635
			S.Hata	0.0008	0.0006	0.0006	0.0009	0.0008	0.0023	0.0007
		200	Ortalama	0.1860	0.1637	0.1605	0.1724	0.1945	0.3287	0.1776
			S.Hata	0.0006	0.0005	0.0005	0.0009	0.0006	0.0023	0.0006
		500	Ortalama	0.1946	0.1817	0.1799	0.2019	0.2091	0.2868	0.1844
			S.Hata	0.0005	0.0003	0.0003	0.0006	0.0004	0.0024	0.0004
	4	100	Ortalama	0.0922	0.0831	0.0815	0.0993	0.0944	0.2083	0.0915
			S.Hata	0.0005	0.0004	0.0004	0.0005	0.0005	0.0017	0.0005
		200	Ortalama	0.1054	0.1006	0.0990	0.1144	0.1090	0.2441	0.1051
			S.Hata	0.0004	0.0003	0.0004	0.0005	0.0004	0.0019	0.0004
		500	Ortalama	0.1149	0.1174	0.1161	0.1250	0.1177	0.2212	0.1153
			S.Hata	0.0002	0.0002	0.0002	0.0004	0.0003	0.0020	0.0002

Tablo 8.15 incelendiğinde göz önünde tutulan 18 farklı koşuldan 15'inde Mahalanobis uzaklığı ve 3'ünde Öklid uzaklığı en başarılı uzaklık ölçüsü olmuştur. Simülasyon çalışması sonucuna göre çok değişkenli normal dağılıma uyan ve kovaryans matrisi $\Sigma = \sigma_i^2 \mathbf{I}$ şeklinde olan veri setlerinin *k*-ortalamlar tekniği ile

kümelenmesinde en başarılı uzaklık ölçüsü Mahalanobis uzaklığıdır. Tablo 8.15’de verilen ortalama Wilks’in lambda değerlerinin uzaklık ölçüsüne göre değişimini gösteren grafikler Şekil 8.9’da verilmiştir.



Şekil 8.9. Durum II’de ilgilenilen 7 farklı uzaklık ölçüsüne göre elde edilen ortalama Wilks’in lambda değerlerinin değişimi

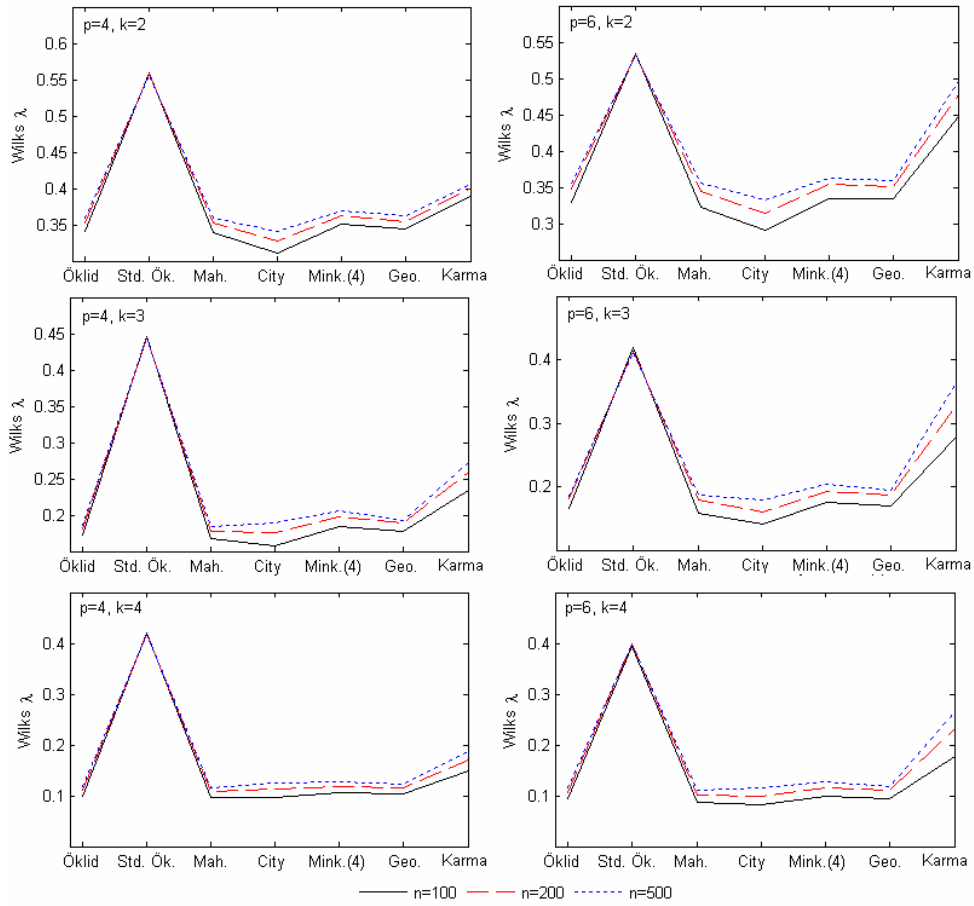
Şekil 8.9’de yer alan grafikler incelendiğinde örneklem hacminin 100 olduğu durumda çok değişkenli normal dağılıma uyan ve kovaryans matrisi $\Sigma = \sigma_i^2 \mathbf{I}$ şeklinde olan veri setlerinin k -ortalamlar tekniği ile kümelenmesinde Mahalanobis ve standartlaştırılmış Öklid uzaklıklarının diğer uzaklık ölçülerine göre daha başarılı olduğu görülmektedir. Örneklem hacminin artması durumunda geometrik uzaklık haricindeki diğer uzaklık ölçüleri arasındaki performans farklılığının azaldığı görülmektedir.

İlgilenilen 7 farklı uzaklık ölçüsü ile Durum III'de 18 farklı koşulda üretilen 1000'er veri setinin kümelenmesi sonucu elde edilen Wilks'in lambda test istatistiği değerlerine ait ortalama ve standart hatalar Tablo 8.16'da verilmiştir.

Tablo 8.16. Durum III için ilgilenilen 7 uzaklık ölçüsü ile elde edilen Wilks'in lambda test istatistiği değerlerine ait ortalama ve standart hatalar

<i>p</i>	<i>k</i>	<i>n</i>	T. İstatistik	Öklid	Std. Ök.	Maha.	City	Mink.(4)	Geo.	Karma
4	2	100	Ortalama	0.3416	0.5586	0.3389	0.3102	0.3501	0.3440	0.3885
			S.Hata	0.0011	0.0011	0.0010	0.0010	0.0011	0.0011	0.0014
		200	Ortalama	0.3520	0.5569	0.3517	0.3278	0.3617	0.3545	0.4013
			S.Hata	0.0008	0.0008	0.0007	0.0008	0.0008	0.0008	0.0011
		500	Ortalama	0.3593	0.5560	0.3591	0.3410	0.3695	0.3618	0.4063
			S.Hata	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0007
	3	100	Ortalama	0.1717	0.4459	0.1669	0.1582	0.1836	0.1771	0.2348
			S.Hata	0.0008	0.0013	0.0007	0.0007	0.0008	0.0008	0.0012
		200	Ortalama	0.1809	0.4434	0.1774	0.1753	0.1974	0.1884	0.2594
			S.Hata	0.0005	0.0009	0.0005	0.0005	0.0006	0.0006	0.0009
		500	Ortalama	0.1854	0.4442	0.1847	0.1900	0.2056	0.1922	0.2733
			S.Hata	0.0003	0.0005	0.0003	0.0003	0.0004	0.0005	0.0006
	4	100	Ortalama	0.0988	0.4214	0.0960	0.0967	0.1059	0.1028	0.1491
			S.Hata	0.0005	0.0013	0.0005	0.0005	0.0005	0.0005	0.0009
		200	Ortalama	0.1107	0.4182	0.1076	0.1124	0.1187	0.1150	0.1709
			S.Hata	0.0004	0.0009	0.0004	0.0004	0.0004	0.0004	0.0008
		500	Ortalama	0.1179	0.4180	0.1157	0.1257	0.1279	0.1224	0.1871
			S.Hata	0.0002	0.0006	0.0002	0.0002	0.0003	0.0003	0.0006
6	2	100	Ortalama	0.3300	0.5347	0.3228	0.2912	0.3349	0.3335	0.4469
			S.Hata	0.0011	0.0013	0.0010	0.0010	0.0011	0.0011	0.0020
		200	Ortalama	0.3473	0.5333	0.3435	0.3135	0.3534	0.3507	0.4776
			S.Hata	0.0008	0.0008	0.0007	0.0007	0.0008	0.0008	0.0015
		500	Ortalama	0.3563	0.5315	0.3553	0.3320	0.3632	0.3596	0.4960
			S.Hata	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0010
	3	100	Ortalama	0.1655	0.4184	0.1588	0.1402	0.1744	0.1702	0.2785
			S.Hata	0.0007	0.0012	0.0007	0.0006	0.0008	0.0008	0.0016
		200	Ortalama	0.1802	0.4138	0.1783	0.1605	0.1929	0.1866	0.3270
			S.Hata	0.0006	0.0009	0.0005	0.0005	0.0007	0.0007	0.0014
		500	Ortalama	0.1855	0.4106	0.1864	0.1799	0.2033	0.1938	0.3633
			S.Hata	0.0004	0.0005	0.0004	0.0003	0.0006	0.0005	0.0011
	4	100	Ortalama	0.0931	0.3949	0.0870	0.0815	0.0989	0.0950	0.1777
			S.Hata	0.0004	0.0013	0.0004	0.0004	0.0006	0.0005	0.0012
		200	Ortalama	0.1066	0.3984	0.1017	0.0990	0.1156	0.1099	0.2318
			S.Hata	0.0004	0.0009	0.0004	0.0004	0.0005	0.0004	0.0012
		500	Ortalama	0.1159	0.3981	0.1104	0.1161	0.1280	0.1190	0.2652
			S.Hata	0.0002	0.0006	0.0002	0.0002	0.0003	0.0003	0.0009

Tablo 8.16 incelendiğinde göz önünde tutulan 18 farklı koşuldan 13'ünde City-Blok uzaklığı ve 5'inde Mahalanobis uzaklığı en başarılı uzaklık ölçüsü olmuştur. Simülasyon çalışması sonucuna göre çok değişkenli normal dağılıma uyan ve kovaryans matrisinde farklı varyansa sahip değişkenler arasındaki ilişkilerin aynı olması durumunda veri setlerinin k -ortalamalar tekniği ile kümelenmesinde en başarılı uzaklık ölçüsü City-Blok uzaklığıdır. Tablo 8.16'de verilen ortalama Wilks'in lambda değerlerinin uzaklık ölçüsüne göre değişimini gösteren grafikler Şekil 8.10'da verilmiştir.



Şekil 8.10. Durum III'de ilgilenilen 7 farklı uzaklık ölçüsüne göre elde edilen ortalama Wilks'in lambda değerlerinin değişimi

Şekil 8.10 incelendiğinde Durum III'de kümeleme performansı en kötü olan uzaklık ölçüsünün standartlaştırılmış Öklid uzaklığı olduğu ve ayrıca geometrik

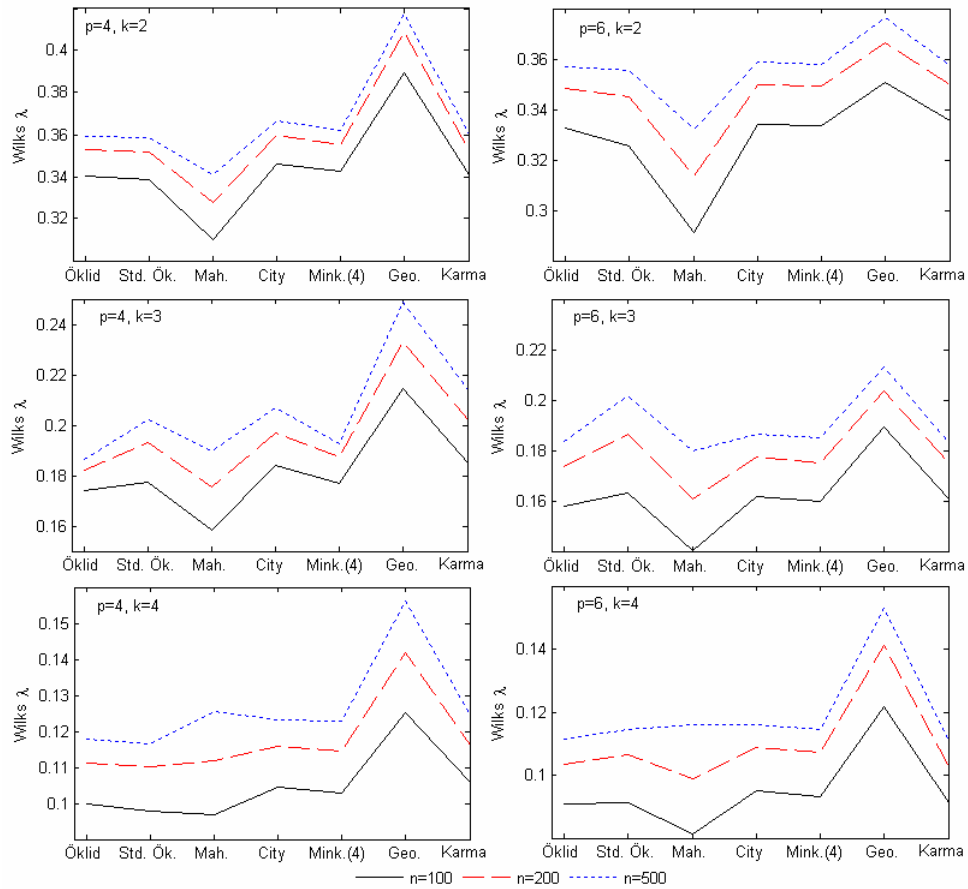
uzaklık ölçüsün önceki iki duruma göre Durum III'de daha başarılı bir kümeleme performansına sahip olduğu görülmüştür.

Simülasyon çalışmasında Durum IV için elde edilen Wilks'in lambda test istatistiği değerlerine ait ortalama ve standart hatalar Tablo 8.17'de verilmiştir.

Tablo 8.17. Durum IV için ilgilenilen 7 uzaklık ölçüsü ile elde edilen Wilks'in lambda test istatistiği değerlerine ait ortalama ve standart hatalar

p	k	n	T. İstatistik	Öklid	Std. Ök.	Maha.	City	Mink.(4)	Geo.	Karma
4	2	100	Ortalama	0.3400	0.3386	0.3102	0.3462	0.3425	0.3896	0.3411
			S.Hata	0.0011	0.0011	0.0010	0.0011	0.0011	0.0015	0.0011
		200	Ortalama	0.3526	0.3515	0.3278	0.3598	0.3554	0.4087	0.3525
			S.Hata	0.0008	0.0007	0.0008	0.0008	0.0008	0.0011	0.0007
		500	Ortalama	0.3592	0.3587	0.3410	0.3666	0.3620	0.4177	0.3601
			S.Hata	0.0005	0.0005	0.0005	0.0005	0.0005	0.0008	0.0005
	3	100	Ortalama	0.1740	0.1773	0.1582	0.1839	0.1768	0.2145	0.1852
			S.Hata	0.0008	0.0008	0.0007	0.0008	0.0008	0.0011	0.0008
		200	Ortalama	0.1821	0.1929	0.1753	0.1970	0.1874	0.2333	0.2022
			S.Hata	0.0006	0.0006	0.0005	0.0006	0.0006	0.0009	0.0006
		500	Ortalama	0.1866	0.2020	0.1900	0.2069	0.1924	0.2487	0.2142
			S.Hata	0.0004	0.0005	0.0003	0.0005	0.0004	0.0007	0.0005
	4	100	Ortalama	0.0999	0.0980	0.0967	0.1045	0.1027	0.1251	0.1057
			S.Hata	0.0005	0.0005	0.0005	0.0005	0.0005	0.0008	0.0005
		200	Ortalama	0.1112	0.1101	0.1120	0.1159	0.1145	0.1418	0.1161
			S.Hata	0.0004	0.0004	0.0004	0.0004	0.0004	0.0007	0.0004
		500	Ortalama	0.1177	0.1164	0.1255	0.1233	0.1227	0.1562	0.1245
			S.Hata	0.0002	0.0002	0.0003	0.0003	0.0003	0.0005	0.0003
6	2	100	Ortalama	0.3326	0.3253	0.2912	0.3340	0.3338	0.3507	0.3354
			S.Hata	0.0011	0.0010	0.0010	0.0011	0.0011	0.0013	0.0010
		200	Ortalama	0.3483	0.3448	0.3135	0.3499	0.3493	0.3664	0.3498
			S.Hata	0.0008	0.0007	0.0007	0.0007	0.0008	0.0008	0.0007
		500	Ortalama	0.3570	0.3555	0.3320	0.3587	0.3579	0.3764	0.3573
			S.Hata	0.0005	0.0005	0.0005	0.0005	0.0005	0.0006	0.0005
	3	100	Ortalama	0.1579	0.1631	0.1402	0.1616	0.1600	0.1895	0.1607
			S.Hata	0.0007	0.0007	0.0006	0.0007	0.0007	0.0010	0.0007
		200	Ortalama	0.1736	0.1865	0.1605	0.1775	0.1752	0.2038	0.1749
			S.Hata	0.0005	0.0005	0.0005	0.0006	0.0005	0.0007	0.0005
		500	Ortalama	0.1836	0.2018	0.1799	0.1866	0.1851	0.2129	0.1831
			S.Hata	0.0003	0.0004	0.0003	0.0004	0.0003	0.0005	0.0003
	4	100	Ortalama	0.0907	0.0911	0.0815	0.0949	0.0933	0.1217	0.0911
			S.Hata	0.0005	0.0005	0.0004	0.0005	0.0005	0.0007	0.0005
		200	Ortalama	0.1036	0.1065	0.0990	0.1088	0.1072	0.1410	0.1027
			S.Hata	0.0004	0.0004	0.0004	0.0004	0.0004	0.0006	0.0004
		500	Ortalama	0.1116	0.1146	0.1161	0.1160	0.1145	0.1531	0.1106
			S.Hata	0.0003	0.0003	0.0002	0.0003	0.0003	0.0004	0.0002

Tablo 8.17 incelendiğinde göz önünde tutulan 18 farklı koşuldan 13'ünde Mahalanobis uzaklığı ve 2'sinde Öklid uzaklığı, 2'sinde standartlaştırılmış Öklid uzaklığı ve 1'inde karma uzaklık en başarılı uzaklık ölçüsü olmuştur. Tablo 8.17'de verilen ortalama Wilks'in lambda değerlerinin uzaklık ölçüsüne göre değişimini gösteren grafikler Şekil 8.11'de verilmiştir.



Şekil 8.11. Durum IV'de ilgilenilen 7 farklı uzaklık ölçüsüne göre elde edilen ortalama Wilks'in lambda değerlerinin değişimi

Şekil 8.11 incelendiğinde Durum IV'de kümeleme performansı Wilks'in lambda istatistiğine göre en iyi olan uzaklık ölçüsünün Mahalanobis uzaklık ölçüsü olduğu görülmektedir.

Yeni önerilen uzaklık ölçülerinden geometrik uzaklık ölçüsünün Wilks'in lambda test istatistiğine göre kümeleme performansı sadece Durum III'de diğer uzaklık ölçüleri kadar iyi olmuştur. Yeni önerilen karma uzaklık ölçüsü kovaryans

matrisinin $\Sigma = I$ olduğu Durum I'deki farklı 18 koşuldan 3'ünde, kovaryans matrisinde farklı varyansa sahip değişkenler arasındaki ilişkilerin farklı olduğu Durum IV'deki farklı 18 koşuldan 1'inde kümeleme performansı en iyi olan uzaklık ölçüsü olmuştur.

8.3. Sonuçlar

Uzaklık ölçülerinin aşamalı ve aşamalı olmayan kümeleme tekniklerinde kümelemeye olan etkisi hem gerçek veri setlerinde hem de simülasyon çalışması ile incelenmiştir. Aşamalı kümeleme tekniklerinde uzaklık ölçülerinin etkisi incelenirken karşılaştırma kriteri olarak kofenetik korelasyon katsayısı ve Wilks'in lambda test istatistiği kullanılmıştır.

Gerçek veri setlerinin farklı uzaklık ölçülerine göre kümelenmesi sonrası elde edilen kofenetik korelasyon katsayı değerleri incelendiğinde ortalama bağlantı, ağırlıklı ortalama bağlantı ve merkez bağlantı kümeleme tekniklerinde Öklid uzaklık ölçüsü en başarılı uzaklık ölçüsü olmuştur. Wilks'in lambda test istatistiği değerlerinin ortalamasına göre tam ve ortalama bağlantı kümeleme tekniklerinde en başarılı uzaklık ölçüsü geometrik uzaklık ölçüsü olmuştur.

Aşamalı kümeleme tekniklerinde uzaklık ölçülerinin etkisi ayrıca farklı örneklem hacimlerinde rassal olarak üretilen 21 değişkene ve 3 kümeye sahip olan Waveform veri setleri ile de incelenmiştir. Kofenetik korelasyon katsayısı ve Wilks'in lambda test istatistiği değerlerine göre örneklem hacminin düşük olduğu ($n = 75$) olduğu durumda tüm aşamalı kümeleme tekniklerinde Mahalanobis uzaklığı en başarılı uzaklık ölçüsü olmuştur. Örneklem hacminin artması durumunda en başarılı uzaklık ölçülerinin tekniklere göre farklılık gösterdiği gözlemlenmiştir.

k -ortalamalar tekniğinde incelemeye dahil edilen veri setlerinin farklı uzaklık ölçüleri ile kümelenmesi sonrasında elde edilen Wilks'in lambda test istatistiği değerlerinin ortalamasına göre en başarılı uzaklık ölçüsü yeni önerilen karma uzaklık ölçüsü olmuştur.

k -ortalamalar tekniğinde uzaklık ölçülerinin kümelemeye etkisinin farklı değişken sayısı, farklı küme sayısı, farklı örneklem hacimleri ve farklı kovaryans

yapılarında Wilks'in lambda test istatistiğine göre incelendiği simülasyon çalışmasında çok değişkenli normal dağılımdan üretilen verilerin kümelendirilmesinde genel olarak en başarılı uzaklık ölçüsü Mahalanobis uzaklık ölçüsü olmuştur.

9. SONUÇLAR VE ÖNERİLER

Kümeleme analizi, küme yapısı hakkında bir ön bilgiye sahip olunmayan birimlerin p değişken bakımından ölçülmesi ile oluşan $n \times p$ boyutlu X veri matrisindeki farklı birimleri farklı kümelerde, benzer birimleri aynı kümede bir araya getirerek veri seti içerisindeki gizli kümeleri ortaya çıkaran çok değişkenli istatistiksel bir analizdir. Kümelemede asıl olan tanımdan da anlaşılabilir gibi birimler arasındaki benzerliğin belirlenmesidir. Birimler arasındaki benzerliklerin belirlenmesinde ya ilişki türü benzerlik ölçüleri ya da uzaklık türü benzerlik ölçüleri kullanılmaktadır. Özellikle sayısal (nicel) verilerde birimler arasındaki benzerliğin saptanmasında genellikle uzaklık ölçüleri kullanılmaktadır. Kümeleme analizinde genel amaç benzer birimleri bir araya getirerek küme içi değişimi minimum, kümeler arası değişimi ise maksimum yapmaktır. Bundan dolayı birimler arasındaki benzerliğin oluşturulması için seçilecek olan uzaklık ölçüsü kümelemeyi doğrudan etkileyen önemli bir faktördür. Bu tez çalışmasında uzaklık ölçülerinin kümelemeye olan etkisi hem gerçek hem de simülasyon verileri kullanılarak detaylı bir biçimde incelenmiştir. Ayrıca boyut indirgeme ve standartlaştırmanın kümelemeye olan etkisi, uygun küme sayısının belirlenmesi ve k -ortalamlar tekniğinde küme başlangıç merkezlerinin seçimi gibi kümeleme analizinin en önemli problemleri üzerinde çalışılmıştır.

Bölüm 3’de kümeleme analizi ile ilgili temel tanım ve gösterimlere yer verilmiştir. Farklı birçok disiplinde farklı amaçlar doğrultusunda gelişimini sürdüren kümeleme analizi içerisinde çok sayıda kümeleme tekniği mevcuttur. Kümeleme analizi içerisinde yer alan bu kümeleme tekniklerinin çok farklı sınıflandırılması söz konusu olsa da en genel anlamda aşamalı ve aşamalı olmayan kümeleme teknikleri olarak iki sınıfa ayrılmaktadır. Buna uygun olarak Bölüm 3’de kümeleme analizinde yer alan teknikler aşamalı ve aşamalı olmayan kümeleme teknikleri olmak üzere iki alt başlık altında incelenmiştir. Aşamalı kümeleme teknikleri arasından eklemeli teknikler olarak bilinen tek bağlantı, tam bağlantı, ortalama bağlantı, ağırlıklı ortalama bağlantı, merkez bağlantı, medyan bağlantı, Ward bağlantı ve esnek beta teknikleri ele alınmıştır. Aşamalı olmayan teknikler içerisinde basitliği ve

kümelemedeki etkinliği dolayısıyla en fazla bilinen ve kullanılan k -ortalamalar tekniği anlatılmıştır. Bölüm 3’de son olarak küme geçerliliği kavramı üzerinde durulmuştur. Küme geçerliliği için önerilen indeksler dışsal, içsel ve oransal kriterler olmak üzere üç başlık altında ele alınmıştır. Dışsal kriterlerden Rand indeksi, Jaccard katsayısı, Fowlkes ve Mallows indeksi, Γ istatistiği ve düzeltilmiş Rand indeksi incelenmiştir. İçsel kriterlerden kofenetik korelasyon katsayısı ve Wiks’in lambda katsayısı incelenirken oransal kriterlerden Calinsky ve Harabasz indeksi, Davies-Bouldin indeksi ve Dunn indeksi incelenmiştir.

Bölüm 4’de uzaklık ölçüsü ile ilgili tanım ve özellikler verilmiştir. Sayısal veri durumunda birimler arasındaki benzerliklerin saptanmasında yaygın bir kullanıma sahip olan City-Blok (Manhattan) uzaklığı, Chebyshev uzaklığı, Öklid uzaklığı, standartlaştırılmış Öklid uzaklığı, Minkowski uzaklığı ve Mahalanobis uzaklığı incelenmiştir. Bölüm 4’de ayrıca aykırı değerlerin gözlem vektörleri arasındaki uzaklığa olan etkisini azaltmak için temeli geometrik ortalamaya dayanan geometrik uzaklık ölçüsü ve ağırlıklandırılmış Öklid uzaklığı ile Pearson korelasyon uzaklığının karmasından oluşan karma uzaklık ölçüsü tanımlanmıştır.

Değişken sayısının fazla olduğu yüksek boyutlu verilerde kümeleme analizi öncesi boyut indirgeme yöntemi ile veri indirgenerek daha az sayıda bileşen ile kümeleme analizi uygulanmaktadır. Kümelemede orijinal değişkenler yerine boyut indirgeme yöntemi ile elde edilen bileşenlerin kullanımının kümelemeye olan etkisi, kümelemede üzerinde çalışılan önemli problemlerden biridir. Bazı araştırmacılar boyut indirgemenin kümelemeye önemsenecek kadar önemli bir olumsuz etkisi olmadığını savunurken, bazı araştırmacılar da boyut indirgeme ile yaşanan bilgi kaybının kümelemeyi olumsuz olarak etkilediğini ve bundan dolayı boyut indirgemenin yapılmaması gerektiğini savunmaktadır. Bölüm 5’de yüksek boyutlu veri matrislerinde kümeleme analizi öncesi uygulanacak temel bileşenler analizinden elde edilen bileşenler içerisinde tercih edilen bileşen sayısının kümelemeye etkisi incelenmiştir. Karşılaştırmada doğru sınıflandırma yüzdesi ve eşleştirme katsayısı kullanılmıştır. Doğru sınıflandırma yüzdesi gerçek küme üyelikleri ile kümeleme sonrası elde edilen küme üyelikleri arasındaki uyumun bir ölçüsü iken, eşleştirme katsayısı orijinal değişkenlerle kümeleme yapıldığında elde edilen küme üyelikleri

ile boyut indirgeyerek yapılan kümeleme sonrası elde edilen küme üyelikleri arasındaki uyumun bir göstergesidir. Orijinal değişkenler yerine temel bileşenlerin kullanılmasının kümelemeye olan etkisi incelenirken küme sayısının farklı olması ve kümelerin kovaryans matrislerinin eşit olup olmaması göz önünde bulundurulmuştur. Çalışma sonucunda, doğru sınıflandırma yüzdesi bakımından boyut indirgemenin avantajlı, eşleştirme katsayısı bakımından ise avantajlı olmadığı tespit edilmiştir. Bu bölümde ayrıca standartlaştırmanın kümeleme olan etkisi incelenmiştir. Ham veri yerine standartlaştırılmış verinin kullanılmasının kümeleme performansını olumsuz etkilediği görülmüştür. Ancak, kümelerin kovaryans matrislerinin farklı olması durumunda standartlaştırmanın kümelemeye olan olumsuz etkisinin azaldığı tespit edilmiştir.

Kümeleme analizindeki en önemli problemlerden biride uygun küme sayısının belirlenmesidir. Kümeleme analizi öncesi küme yapısı hakkında herhangi bir ön bilginin olmaması, veri yapısındaki farklılıklar uygun küme sayısının tahmini için genel ve kabul görmüş bir formülün verilmesini zorlaştırmaktadır. Buna rağmen literatür incelendiğinde küme sayının belirlenmesi ile ilgili birçok farklı yaklaşımın olduğu görülmektedir. Uygun küme sayısının belirlenmesi için önerilen yöntemlerin çoğu deneme yanılmaya bağlı teknikler olduğu için seçilen kümeleme tekniği ve uzaklık fonksiyonundan etkilenmektedir. Dolayısıyla bu yöntemler ile aynı veri seti için farklı küme sayısı önerilebilmektedir.

Bölüm 6'da uygun küme sayısının tahmini için yeni bir yöntem önerilmiştir. Önerdiğimiz yöntemin temeli $n \times p$ boyutlu veri matrisinin $n \times 1$ boyutlu bir temsil vektörü ile temsil edilmesi esasına dayanmaktadır. Küçükten büyüğe doğru sıralanmış temsil değerleri arasındaki ardışık farkların belirlenen beklenen minimum farkla karşılaştırılması ile uygun küme sayısı tespit edilmeye çalışılmıştır. Uygun küme sayısının belirlenmesi için önerilen yöntem ardışık farklar yöntemi olarak isimlendirilmiştir. Ardışık farklar yöntemi uygun küme sayısının belirlenmesi için önerilen üst kuyruk yöntemi, Marriott indeksi, varyans açıklama oranı, Silhouette indeksi ve Gap istatistiği yöntemleri ile karşılaştırılmıştır. Karşılaştırmada kümeleme literatüründe iyi bilinen ve yaygın bir kullanıma sahip olan Diyabet (Diabetes), E.coli, Gayzer (Geyser), Cam (Glass) İris Çiçeği (Iris), Araç Görüntüsü (Vehicle

Silhouettes), Şarap (Wine) ve Maya (Yeast) veri setleri kullanılmıştır. Karşılaştırma sonucunda küme sayısı tahmini için önerdiğimiz ardışık farklar yöntemi ile mevcut yöntemlere göre daha iyi küme sayısı tahminleri elde edildiği görülmüştür. Ardışık farklar yöntemi ile incelenen tüm veri setlerinde, iç içe kümelemenin olmaması durumunda iki küme merkezi arasındaki beklenen minimum farkın hesaplanmasında $\gamma = 3$, iç içe kümelemede kümelerin kendi içerisinde bir kez parçalanması durumunda $\gamma = 2.5$ ve kümelerin kendi içerisinde birden fazla kez parçalanması durumunda $\gamma = 1.2$ alındığında küme sayısı tahminleri doğru olarak yapılmıştır. Buna karşın literatürdeki mevcut yöntemlerin küme sayısı tahminlerinin çok iyi olmadığı, bu yöntemlerle aynı veri seti için farklı kümeleme ve farklı uzaklık ölçülerinin kullanılması durumunda farklı küme sayısı tahminlerinin elde edildiği gösterilmiştir.

Bölüm 7’de, k - ortalamalar tekniğinde küme başlangıç merkezlerinin seçimi için yeni bir algoritma önerilmiştir. Önerilen algoritmanın temeli en büyük değişim katsayısına sahip değişken ve bu değişken ile mutlak değerce minimum korelasyona sahip değişken tarafından iki eksenin oluşturulmasıdır. Önerilen algoritma ile sınıflama problemlerinde yaygın bir kullanıma sahip olan gerçek veri setleri İris, Şarap, Harf ve Ruspini verilerinde rassal başlangıçlı k -ortalamalar algoritmasına göre daha başarılı kümelemeler elde edilmiştir. Ayrıca hatalı sınıflandırma yüzdesi bakımından Khan ve Ahmad (2004)’in önerdikleri CCIA algoritması ile Deelers ve Auwatanamongkol (2007) tarafından önerilen algoritmalarla göre de önerilen yeni algoritma daha başarılı bulunmuştur. Literatürdeki mevcut algoritmaların çoğuna göre daha basit ve etkin olan bu yeni algoritmanın yüksek boyutlu veri setlerinde kolaylıkla uygulanabilmesi algoritmanın önemli bir avantajıdır.

Bölüm 8’de uzaklık ölçülerinin aşamalı ve aşamalı olmayan kümeleme tekniklerinde kümelemeye olan etkisi incelenmiştir. Uzaklık ölçülerinin aşamalı ve aşamalı olmayan kümeleme tekniklerinde kümelemeye olan etkisi hem gerçek veri setlerinde hem de simülasyon çalışması ile incelenmiştir. Aşamalı kümeleme tekniklerinde uzaklık ölçülerinin etkisi incelenirken karşılaştırma kriteri olarak kofenetik korelasyon katsayısı ve Wilks’in lambda test istatistiği kullanılmıştır. Gerçek veri setlerinin farklı uzaklık ölçülerine göre kümelenmesi sonrası elde edilen

kofenetik korelasyon katsayı değerleri incelendiğinde ortalama bağlantı, ağırlıklı ortalama bağlantı ve merkez bağlantı kümeleme tekniklerinde Öklid uzaklık ölçüsü en başarılı uzaklık ölçüsü olmuştur. Wilks'in lambda test istatistiği değerlerinin ortalamasına göre tam ve ortalama bağlantı kümeleme tekniklerinde en başarılı uzaklık ölçüsü yeni tanımlanan geometrik uzaklık ölçüsü olmuştur.

Aşamalı kümeleme tekniklerinde uzaklık ölçülerinin etkisi ayrıca farklı örneklem hacimlerinde rassal olarak üretilen 21 değişkene ve 3 kümeye sahip olan Waveform veri setleri ile de incelenmiştir. Kofenetik korelasyon katsayısı ve Wilks'in lambda test istatistiği değerlerine göre örneklem hacminin düşük olduğu ($n = 75$) olduğu durumda tüm aşamalı kümeleme tekniklerinde Mahalanobis uzaklığı en başarılı uzaklık ölçüsü olmuştur. Örneklem hacminin artması durumunda en başarılı uzaklık ölçülerinin tekniklere göre farklılık gösterdiği gözlemlenmiştir.

k -ortalamlar tekniğinde incelemeye dahil edilen veri setlerinin farklı uzaklık fonksiyonları ile kümelenmesi sonrasında elde edilen Wilks'in lambda test istatistiği değerlerinin ortalamasına göre en başarılı uzaklık ölçüsü yeni önerilen karma uzaklık ölçüsü olmuştur.

k -ortalamlar tekniğinde uzaklık ölçülerinin kümelemeye etkisinin incelendiği simülasyon çalışmasında farklı iki değişken sayısı, farklı üç küme sayısı, farklı üç örneklem hacmi ve farklı dört kovaryans yapısında çok değişkenli normal dağılımdan veri setleri üretilmiştir. Simülasyon çalışmasında değişken sayısı p sırasıyla 4 ve 6, küme sayısı k sırasıyla 2, 3 ve 4, örneklem hacmi n 100, 200 ve 500 alınmıştır. Değişken sayısının 4 olduğu simülasyon çalışmasında üretilecek rassal örneklem için ortalama vektörü $\mu = [1 \ 3 \ 5 \ 7]$ iken değişken sayısının 6 olduğu durumda ortalama vektörü $\mu = [1 \ 10 \ 5 \ 1000 \ 7 \ 0.5]$ olarak alınmıştır. Rassal örneklemelerin çok değişkenli normal dağılımdan üretilmesi esnasında kovaryans matrisinin $\Sigma = \mathbf{I}$ olması Durum I, kovaryans matrisinin $\Sigma = \sigma_i^2 \mathbf{I}$ olması Durum II, kovaryans matrisinde farklı varyansa sahip değişkenler arasındaki korelasyonların aynı olması Durum III ve kovaryans matrisinde farklı varyansa sahip değişkenler arasındaki korelasyonların farklı olması Durum IV olarak ele alınmıştır. Simülasyon çalışmasında ilgilenilen 7 farklı uzaklık ölçüsü ile 72 farklı koşulda

1000'er tekrarla üretilen 72000 verinin k -ortalamalar tekniğiyle kümelenmesi ile 504000 kümeleme sonucu incelenmiştir. Simülasyon çalışmasında uzaklık ölçülerinin kümeleme olan etkisi Wilks'in lambda test istatistiğine göre değerlendirilmiştir.

Durum I'de göz önünde tutulan 18 farklı koşuldan 12'sinde Mahalanobis uzaklığı, 3'ünde karma uzaklık, 2'sinde standartlaştırılmış Öklid uzaklığı ve 1'inde Öklid uzaklığı en başarılı uzaklık ölçüsü olmuştur. Yeni tanımlanan karma uzaklık ölçüsü Durum I'de değişken sayısının 4, küme sayısının 3 olduğu 500 birimlik örneklem ve değişken sayısının 4, küme sayısının 4 olduğu 200 ve 500 birimlik örneklemelerin kümelendirilmesinde Wilks'in lambda test istatistiğine göre en başarılı uzaklık ölçüsü olarak belirlenmiştir.

Durum II'de göz önünde tutulan 18 farklı koşuldan 15'inde Mahalanobis uzaklığı ve 3'ünde Öklid uzaklığı en başarılı uzaklık ölçüsü olmuştur. Örneklem hacminin 100 olduğu durumda çok değişkenli normal dağılıma uyan ve kovaryans matrisi $\Sigma = \sigma_i^2 \mathbf{I}$ şeklinde olan veri setlerinin k -ortalamalar tekniği ile kümelenmesinde Mahalanobis ve standartlaştırılmış Öklid uzaklıklarının diğer uzaklık fonksiyonlarına göre daha başarılı olduğu görülmüştür.

Durum III'de göz önünde tutulan 18 farklı koşuldan 13'ünde City-Blok uzaklığı ve 5'inde Mahalanobis uzaklığı en başarılı uzaklık ölçüsü olmuştur. Simülasyon çalışması sonucuna göre çok değişkenli normal dağılıma uyan ve kovaryans matrisinde farklı varyansa sahip değişkenler arasındaki ilişkilerin aynı olması durumunda veri setlerinin k -ortalamalar tekniği ile kümelenmesinde en başarılı uzaklık ölçüsü City-Blok uzaklığı olmuştur.

Durum IV'de göz önünde tutulan 18 farklı koşuldan 13'ünde Mahalanobis uzaklığı ve 2'sinde Öklid uzaklığı, 2'sinde standartlaştırılmış Öklid uzaklığı ve 1'inde karma uzaklık en başarılı uzaklık ölçüsü olmuştur. Yeni tanımlanan karma uzaklık ölçüsü Durum IV'de değişken sayısının 6, küme sayısının 4 olduğu 500 birimlik örneklemelerin kümelendirilmesinde en başarılı uzaklık ölçüsü olmuştur.

Simülasyon çalışması sonucunun genel olarak değerlendirilmesi sonrasında çok değişkenli normal dağılımdan üretilen verilerin kümelendirilmesinde en başarılı uzaklık ölçüsü Mahalanobis uzaklık ölçüsü olmuştur.

KAYNAKLAR

- ABDI, H., 2007. Distance, In Salkind, N.J. (eds.) Encyclopedia of Measurement and Statistics. 280–284. Sage Thousand Oaks. 1136p.
- AKIN, K. Y., 2008. Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi. Marmara Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı, Yayınlanmamış Doktora Tezi, İstanbul. 174s.
- ALDENFERDER, M. S. and BLASHFIELD, R. K., 1984. Cluster Analysis. Sage University Paper, London. 88p.
- ANDERBERG, M. R., 1973. Cluster Analysis for Applications. Academic Press Inc., London. 359p.
- BABU, G. P. and MURTY, M. N., 1993. A near-optimal initial seed value selection in K-means algorithm using a genetic algorithm. Pattern Recognition Letters, 14 (10): 763-769.
- BAKER, F. B., and HUBERT, L. J., 1975. Measuring the power of hierarchical cluster analysis. Journal of the American Statistical Association, 31–38.
- BALL, G., and HALL, D., 1965, ISODATA, A novel method of data analysis and pattern classification. Technical report NTIS AD 699616. Stanford Research Institute, Stanford, CA.
- BANFIELD, J. D. and RAFTERY, A. E., 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics, 49: 803-822.
- BEZDEK, J. C., 1974a. Numerical taxonomy with fuzzy sets. Journal of Mathematical Biology, 57-71.
- BEZDEK, J. C., 1974b. Cluster Validity with fuzzy sets. Journal of Cybernetics, 3: 58-78.
- BINDER, D.A., 1978. Bayesian Cluster Analysis, Biometrika, 65:31-38.
- BLASHFIELD, R. K., 1976. Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. Psychological Bulletin, 83: 377-388.
- BOUYEYRON, C., GIRARD, S. and SCHMID, C., 2007. High-Dimensional Data Clustering. Computational Statistics & Data Analysis, 52: 502–519.

- BRADLEY, P. S., and FAYYAD, U.M., 1998. Refinining initial points for k-means algorithm. Proceeding of The 15. International Conference on Machine Learning.
- BREIMAN, L., FRIEDMAN, J., OLSHEN, R. and STONE, C., 1984. Classification and Regression Trees. Wadsworth Belmont, CA Press. 358p.
- BOUGUESSA, M., WANG, S., SUN H., 2006. An objective approach to cluster validation. Pattern Recognition Letters, 27 (13): 1419-1430.
- CALINSKI, R.B. and HARABASZ, J., 1974. A dendrite method for cluster analysis. Communications in Statistics, 3: 1-27.
- CELEUX, G. and GOVAERT, G., 1992. A Classification EM Algorithm for Clustering and Two Stochastic versions. Computational Statistics and Data Analysis, 14: 315-332.
- CELEUX, G. and GOVAERT, G., 1993. Comparison of the Mixture and the Classification Maximum likelihood in Cluster Analysis. Journal of Statistical Computation and Simulation, 47: 127-146.
- CHANG, W. C., 1983. On using principal components before separating a mixture of two multivariate normal distributions. Applied Statistics, 32: 267–275.
- CLEMENTS, F. E., 1954, Use of Cluster Analysis with Anthropological Data. American Anthropologist, 56 (2): 180-199.
- COLE-RODGERS, P., SMITH, D. W. and BOSLAND, P. W., 1997. A novel statistical approach to analyze genetic resource evaluations using Capiscim as an example. Crop Science, 3: 1000-1002.
- COX, D. R., 1957. Note on grouping. Journal of the American Statistical Association, 52 (280): 543-547.
- CRONBACH, L. J. and GLESER, G. C., 1953. Assessing Similarity Between Profiles. Psychology Bulletin, 50 (6): 456-473.
- DAVIES, D. and BOULDIN, D., 1979. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1: 224–227.
- DAY, N.E., 1969. Estimating the components of a mixture of normal distributions. Biometrika, 56: 463-474.

- DEELERS, S. and AUWATANAMONGKOL, S., 2007. Enhancing k-means algorithm with initial cluster centers derived from data partitioning along the data axis with the highest variance. *International Journal of Computer Science*, 2 (4): 247-252.
- DEZA, M. M. and DEZA, E., 2006. *Dictionary of Distances*. Elsevier. 391p.
- DİNÇER, K. S., 1992. Kümeleme Çözümlemesinde Uygun Kümeleme Ölçütlerinin Karşılaştırılması. *Yayınlanmamış Doktora Tezi*, Anadolu Üniversitesi Sağlık Bilimleri Enstitüsü, Eskişehir. 128s.
- DUDA, R. O., HART, P.E. and STORK, D. G., 2001. *Pattern Classification*. John Wiley & Sons, INC. New York. 654p.
- DUNN, J., 1974. Well separated clusters and optimal fuzzy partitions . *Journal of Cybernetics*, 4: 95–104.
- DURAN, B.S. and ODELL, P.L., 1974. *Cluster Analysis*. New York: Springer-Verlag. 137p.
- ERİŞOĞLU, M., PEKGÖR, A., KARAKOCA, A. and GENÇ, A., 2006. A Comparison of Nonhierarchical Clustering Methods. *MS2006 Proceedings of the International Conference on Modelling and Simulation*, 2: 741-744.
- EVERITT, B. S., 1993. *Cluster analysis*. 3rd ed. London: Edward Arnold. 170p.
- EVERITT, B.S., LANDAU, S. and LEESE, M., 2001. *Cluster Analysis*. Oxford University Press Inc., New York, NY. 237p.
- FERREIRA, L. A., 2007. *Clustering Functional Data: A Comparison of Hierarchical Clustering Methods*. Science in Statistics College of Arts and Sciences University of South Carolina, Unpublished Masters Thesis , South Carolina. 81p.
- FINCH, H. and HUYNH H., 2000. Comparison of Similarity Measures in Cluster Analysis with Binary Data. *Annual Meeting of the American Educational Research Association in New Orleans*.
- FISHER, W. D., 1958. On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53 (284): 789-798.

- FLOREK, K., LUKASZEWICZ, J., STEINHAUS, H., and ZUBRZYCKI, S., 1951. Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum*, 2: 282–285.
- FORGY, E. W., 1965. Cluster analysis of multivariate data: Efficiency vs interpretability of classifications. *Biometrics*, 21: 768–769.
- FOWLKES, E. B. and MALLOWS, C. L., 1983. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78: 553-584.
- FRIEDMAN, H. P. and RUBIN, J., 1967. On some invariant criteria for grouping data. *Journal of American Statistical Association*, 62: 1159-1178.
- GAN, G., MA, C. and WU J., 2007. *Data Clustering: Theory, Algorithms, and Applications*. Society For Industrial & Applied Mathematics, U.S. 466p.
- GILBERT, E., W., 1958. Pioneer maps of health and disease in England. *Geographical Journal*, 124:172-183.
- GORDON, A. D., 1998. Cluster validation. In: C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.H. Bock and Y. Baba, Editors, *Studies in Classification, Data Analysis, and Knowledge Organization: Data Science, Classification, and Related Methods*, Springer, Tokyo, 22–39.
- GORDON, A. D., 1999. *Classification*. 2nd edn. Chapman & Hall, New York, NY. 256p.
- GOWER, J. C., 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53 (3/4): 325-338.
- GOWER, J. C., 1967. A comparison of some methods of cluster analysis. *Biometrics*, 23(4): 623–637.
- GOWER, J. C., 1971. A general coefficient of similarity and some of its properties. *Biometrics*, 27: 857-872.
- GREEN, P. E. and RAO V. R., 1969. A Note on Proximity Measures and Cluster Analysis. *Journal of Marketing Research*, VI: 359-364.
- GUNJACA J., SATAVIĆ Z. and KOLAK I., 2000. Combining qualitative and quantitative TR4IT data in classification of gene bank accessions. 22 nd Int. Conf. information Technology Interfaces ITI-2000: 311-315.

- HAIR, J. F. and BLACK, W.C., 2000. Cluster Analysis. In Laurence G.Grimm & Paul R. Yarnold (Eds.). Reading and Understanding More Multivariate Statistics. Washington, D.C.: American Psychological Association.
- HALKIDI, M., BATISTAKIS, Y., and VAZIRGIANNIS, M., 2002a. Cluster Validity Methods: Part I. SIGMOD Record 31 (2): 40-45.
- HALKIDI, M., BATISTAKIS, Y., and VAZIRGIANNIS, M., 2002b. Clustering Validity Checking Methods: Part II. SIGMOD Record 31 (3): 19-27.
- HANDS, S. and EVERITT, B., 1987. A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques. Multivariate Behavioral Research 22, 235-243.
- HANSEN, P., HERTZ, A. and QUINODOZ, N., 1997. Splitting trees. Discrete Mathematics 165-166: 403-419.
- HARTIGAN, J., 1975. Clustering Algorithms. Toronto, JohnWiley & Sons. 351p.
- HARTIGAN, J. and WONG, M., 1979. Algorithm AS136: A k-means clustering algorithm. Applied Statistics, 28 (1): 100–108.
- HAWKINS, D. M., MÜLER, M. W. and KROODEN, J. A. (1982). Topics in applied multivariate analysis. Cambridge University Pres, USA. 362p.
- HE, J., LAN, M., TAN, C. L., SUNG, S. Y., LOW, H. B., 2004. Initialization of cluster renement algorithms: A review and comparative study. In: Proceedings of International Joint Conference on Neural Networks (IJCNN). Budapest, Hungary.
- <http://apps.isiknowledge.com> (Erişim Tarihi: 03 Temmuz 2010)
- HUBERT, L. and ARABIE, P., 1985. Comparing partitions. Journal of Classification, 193–218.
- HUBERT, L. SCHULTZ, J., 1976. Quadratic assignment as a general data-analysis strategy. British Journal of Mathematical and Statistical Psychologie, 29, 190-241.
- İŞİK, M. VE ÇAMURCU A. Y., 2007. K-Means, K-Medoids ve Bulanık C-Means Algoritmalarının Uygulamalı Olarak Performanslarının Tespiti. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 11 (1): 31-45.

- JACCARD, P., 1908. Nouvelles recherches sur la distribution florale. Bull. Soc. Vaud. Sci. Nat. 44 (1908): 223–270.
- JAIN, A., MURTY, M., and FLYNN, P., 1999. Data clustering: A review. ACM Computing Surveys, 31 (3): 264–323.
- JAIN, A.K. and DUBES, R.C. 1988. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ. 320p.
- JAIN, R. and KORONIOS, A., 2008. Innovation in the cluster validating techniques. Fuzzy Optim Decis Making, 7:233–241
- JENSEN, R. E., 1969. A Dynamic Programming Algorithm for Cluster Analysis. Operations Research, 17 (6): 1034-1057.
- JIE, Y., JAUME, A., NICU, S. and QI, T., 2006. A New Study On Distance Metrics As Similarity Measurement. IEEE International Conference on Multimedia and Expo(ICME), 533-536.
- JOHNSON, R. A. and WICHERN, D. W., 2007. Applied Multivariate Statistical Analysis. Sixth Edition, Pearson Prentice Hall, New Jersey.
- JOLLIFFE, I. T., JONES, B. and MORGAN, B. J. T., 1980. Cluster analysis of the elderly at home: a case study. Data Analysis Information, 745–757.
- KANTARDZIC, M., 2003. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons Inc., Hoboken, New Jersey. 345p.
- KATSAVOUNIDIS, I., JAY KUO, C., ZHANG, Z., 1994. A new initialization technique for generalized Lloyd iteration. Signal Processing Letters, 1: 144-146.
- KAUFMAN, L., and ROUSSEEUW, P. J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons, Inc. 342p.
- KHAN S. S. and AHMAD A., 2004. Cluster center initialization algorithm for k-means algorithm. Pattern Recognition Letters, 25: 1293-1302.
- KOTHARI, R. and PITTS, D., 1999. On Finding The Number of Clusters. Pattern Recognition Letters, 20 (4): 405-416.

- KROLAK-SCHWERDT, S. AND ECKES T., 1992. A graph Theoretic Criterion for Determining Number of Clusters in a Data. *Multivariate Behavioral Research*, 27 (4): 541-565.
- KRZANOWSKI, W. J. and LAI, Y. T., 1988. A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics*, 44: 23-34.
- KUIPER, F. K. and FISHER, L., 1975. A Monte Carlo Comparison of Six Clustering Procedures. *Biometrics*, 31: 777-783.
- LANCE P., EHTESHAM H., HUAN L., 2004. Subspace Clustering for High Dimensional Data: A Review. *Sigkdd Explorations*, 6 (1): 90-105.
- LANCE, G.N. and WILLIAMS, W.T., 1967. A general theory of classificatory sorting strategies. I. Hierarchical systems. *Comput. J.* 9: 60-64.
- LEGENDRE, L. and LEGENDRE, P., 1983. *Numerical Ecology*. New York, Elsevier Scientific. 853p.
- LIAO, H., 2003. Comparison of Clustering Algorithms for Gene Expression Microarray Data. Unpublished PhD Thesis, Actuarial Science, Boston University. 156p.
- LIKAS A., VLASSIS N., JAKOB J. V., 2003. The global k-means algorithm. *Pattern Recognition*, 36: 451-461.
- LLOYD, S. P., 1957. Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: LLOYD, S. P. (1982). "Least squares quantization in PCM". *IEEE Transactions on Information Theory* 28 (2): 129-137.
- MAC QUEEN, J., 1967. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, eds. L. M. L. Cam and J. Neyman, Berkeley, CA: University of California Press, 281-297.
- MAO, J. and JAIN, A. K., 1996. A self-organizing network for hyperellipsoidal clustering (HEC). *IEEE Transactions on Neural Networks*, 7: 16-29.

- MARRIOTT, F. H. C., 1971. Practical Problems in a Method of Cluster Analysis. *Biometrics*, 27: 501–514.
- MARTINEZ, W.L. and MARTINEZ, A.R., 2005. Explanatory Data Analysis with Matlab. Computer Science and Data Analysis Series, Chapman & Hall/CRC. NewYork. 405p.
- MAULIK, U. and BANDYOPADHYAY, S., 2002. Performance evaluation of some clustering algorithms and validity indices. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24: 1650–1654.
- MCLACHLAN, G.J. and PEEL, D., 1996. An algorithm for unsupervised learning via normal mixture models. In *ISIS: Information, Statistics and Induction in Science*, D.L. Dowe, K.B. Korb, and J.J. Oliver (Eds.). Singapore: World Scientific Publishing, 354-363.
- MCQUITTY, L., 1957. Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies. *Educational and Psychological Measurement*, 17: 207–222.
- MILLIGAN, G. W., 1981. A Monte Carlo Study Of Thirty Internal Criterion Measures For Cluster Analysis. *Psychometrika*, 46 (2): 187-199.
- MILLIGAN, G. W., COOPER M. C., 1985. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*, 50: 159–179.
- MIMMACK, G. M., MASON, S. J. and GALPIN, J. S., 2001. Choice of Distance Matrices in Cluster Analysis: Defining Regions. *Journal of Climate*, 14: 2790-2797.
- MIRKIN, B., 2005. Clustering for Data Mining: A Data Recovery Approach. Chapman & Hall/CRC. 266p.
- MOJENA, R., 1977. Hierarchical Grouping Methods and Stopping Rules an Evaluation. *The Computer Journal*, 20 : 359–363.
- MORRISON, D., 1967. Measurement problems in cluster analysis. *Management Science (Series B, Managerial)*, 13(12): B775–B780.

- MUFTI, G.B., BERTRAND, P. and MOUBARKI L., 2005. Determining the number of groups from measures of cluster stability. In: Proceedings of ASMDA2005, 404–414.
- ORLOCI, L., 1967. An agglomerative method for classification of plant communities. *Journal of Ecology*, 55: 193–205.
- OVERALL, J. E. and KLETT, J. C., 1972. *Applied Multivariate Analysis*. McGraw-Hill Book Company, New York. 362p.
- ÖZTÜRK, İ., 1999. n<p Boyutlu biyolojik verilerde farklı Kümeleme Yöntemlerinin Karşılaştırmalı Olarak İncelenmesi. Yayınlanmamış Doktora Tezi, Harran Üniversitesi Fen Bilimleri Enstitüsü, Şanlıurfa. 142s.
- PEARSON, K., 1894. Contributions to the mathematical theory of evolution. *Phil. Trans. Roy. Soc. London A* 185 : 71-110.
- PEETERS, J. P. and MARTINELLI, J. A., 1989. Hierarchical Cluster Analysis As A Tool To Manage Variation in Germplasm Collections. *Theory Appl. Genet.*, 78: 42-48.
- PENA, J. M., LOZANO, J. A. and LARRANAGA P., 1999. An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recognition Letters*, 20 (10) : 1027-1040.
- PUN, WK D. and ALI ABM S., 2007. Unique Distance Measure Approach for K-means (UDMA-Km) Clustering Algorithm. *TENCON 2007, 2007 IEEE Region 10 Conference*, 1-4.
- DINGXI Q., AJIT C. T., 2007. A comparative study of the *K*-means algorithm and the normal mixture model for clustering: Univariate case. *Journal of Statistical Planning and Inference*, 137, 3722 – 3740.
- RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66: 846–850.
- RENCER, A. C., 2002. *Methods of Multivariate Analysis*. Second Edition, John Wiley&Sons Inc. Publication. 708p.
- ROMESBURG, H. C., 1984. *Cluster Analysis for Researchers*. Lifetime Learning Publications, Belmont, California. 334p.

- ROUSSEEUW, P., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65.
- ROUSSOS, L. A., WILLIAM, E. S. And JOHN I. M., 1998. Using New Proximity Measures With Hierarchical Cluster Analysis to Detect Multidimensionality. *Journal of Educational Measurement*, 35 (1): 1-30.
- SAMPATH D., HENRIK B., 2006. Reducing High-Dimensional Data by Principal Component Analysis vs. Random for Nearest Neighbor Classification. *Proceedings of the 5th International Conference on Machine Learning and Applications (ICMLA'06)*.
- SEBER, C. A. F., 1984. *Multivariate Observations*. John Wiley & Sons, New York. 686p.
- SERVİ, T., 2009. Çok Değişkenli Karma Dağılım Modeline Dayalı Kümeleme Analizi. Yayınlanmamış Doktora Tezi. Çukurova Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, Adana. 239s.
- SERVİ, T. and EROL, H., 2007. On Total Number Of Candidate Component Cluster Centers and Total Number of Candidate Mixture Models In Model Based Clustering. *Selçuk Journal of Applied Mathematics*, 8 (2): 57–69.
- SHARMA, S., 1996. *Applied Multivariate Techniques*. New York, JohnWiley & Sons. 493p.
- SNEATH, P., 1957. The applications of computers to taxonomy. *Journal of General Microbiology*, 17: 201–226.
- SNEATH, P. and SOKAL, R., 1973. *Numerical Taxonomy*. San Francisco, CA: Freeman. 573p.
- SOKAL, R. and MICHENER, C., 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38: 1409–1438
- SOKAL, R. and ROHLF, F., 1962. The comparison of dendrograms by objective methods. *Taxonomy*, 11: 33–40.
- SOKAL, R. and SNEATH, P., 1963. *Principles of Numerical Taxonomy*. Freeman, San Francisco. 359p.

- SORENSEN, T., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, 5: 1–34.
- STEINHAUS, H., 1955. Sur la division des corp materiels en parties. *Bulletin of acad. polon. sci.*, IV(C1. III): 801–804.
- STEINLEY, D., and BRUSCO, M. J., 2008. Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika*, 73: 125-144.
- SUGAR, C. A. and JAMES, G. M., 2003. Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Association* 98: 750–763.
- SUN, H., WANG, S. and JIANG, Q., 2001. A New Validation Index for Determining The Number of Clusters in A Data Set. *Neural Networks*, 2001. Proceedings. IJCNN '01. International Joint Conference, 3:1852–1857.
- TANTUĞ, A. C., 2002. Veri madenciliği ve Demetleme. Yayınlanmamış Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü. İstanbul. 73s.
- THEODORIDIS, S. And KOUTROUMBAS, K., 2006. *Pattern Recognition*. Third ed. Elsevier Academic Pres. 837p.
- TIBSHIRANI, R., WALTHER, G. and HASTIE, T., 2001. Estimating the number of clusters via the gap statistic, *J Royal Statist Soc B*, (63): 411-423.
- TOMBAK, H. F., 1996. Çok Değişkenli Analizlerde Kullanılan Uzaklık ve Benzerlik Ölçülerinin Kullanım Yerleri Üzerine Uygulamalı Bir Çalışma. Yayınlanmamış Doktora Tezi, Hacettepe Üniversitesi, Sağlık Bilimleri Enstitüsü, Ankara. 80s.
- TRYON, R. C., 1939. *Cluster Analysis*. Edwards Brothers, Ann Arbor, MI. 347p.
- ÜNLER, A., 2006. Sezgisel Yöntemlerle K-Harmonik Ortalama Veri Kümeleme Eniyilemesi. Yayınlanmamış Doktora Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara. 147s.

- VALARMATHIE, P., SRINATH M. and DINAKARAN K., 2009. An increased performance of clustering high dimensional data through dimensionality reduction technique. *Journal of Theoretical and Applied Information Technology*, 13: 271-273.
- VASSILIOU, A., TAMBOURATZIS, D. G., KOUTRAS, M. V. and BERSIMIS, S., 2004. A New Similarity Measure and Its Use in Determining the Number of Clusters in a Multivariate Data Set. *Communications in Statistics*, 33 (7): 1643–1666.
- VIMAL, A., VALLURI, S. R. and KARLAPALEM, K., 2008. An Experiment with Distance Measures for Clustering. Technical Report: IIIT/TR/2008/132, Center for Data Engineering International Institute of Information Technology Gachibowli, Hyderabad, India.
- WARD, J. H., 1963. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*. 58: 234-244.
- WARD, JR., J. and HOOK, M., 1963. Application of a hierarchical grouping procedure to a problem of grouping profiles. *Educational and Psychological Measurement*, 23 (1):69–81.
- WEBB, A. R., 2002. *Statistical Pattern Recognition*. Second Edition. John &Wiley Sons. Chichester, England, 496p.
- WHITTAKER, R., 1952. A study of summer foliage insect communities in the Great Smoky Mountains. *Ecol. Monogr.*, 22:1–44.
- WILLETT, P., 1988. Recent trends in hierarchical document clustering: a critical review. *Information Processing & Management*, 24 (5): 577-597.
- WOLFE, J. H., 1965. A computer Program for the Maximum-Likelihood Analysis of Types, USNPRA Technical Bulletin 65-15, U.S. Naval Personal Research Activity, San Diego.
- WOLFE, J. H., 1967. NORMIX: Computational Methods for Estimating the Parameters of Multivariate Normal Mixture Distributions, Technical Bulletin USNPRA SRM 68-2, U.S. Naval Personnel Research Activity, San Diego.

- XIE, X. and BENI, G., 1991. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13 (8): 841–847.
- XU, R. and WUNSCH, D. C. II., 2009. *Clustering*, John Wiley & Sons, Inc., Hoboken, New Jersey. 358p.
- YAN, M. and YE, K., 2007. Determining the Number of Clusters Using the Weighted Gap Statistic. *Biometrics*, 63: 1031–1037.
- YE, N., 2003. *The Handbook of Data Mining*. Lawrence Erlbaum Associates Publishers, Mahwah, New Jersey, London. 689p.
- YEUNG, K. Y., RUZZO, W. L., 2001. Principal Component Analysis for Clustering Gene Expression Data. *Bioinformatics*, 17 (9): 763-774.
- ZALIK, K. R. and ZALIK B., 2011. Validity index for clusters of different sizes and densities. *Pattern Recognition Letters*, 32 (2): 221-234.
- ZIVIANI, A., FDIDA, S., DE REZENDE, J. F., and DUARTE, O.C.M.B., 2004. Toward a Measurement-based Geographic Location Service. In: *Proc. of the Passive and Active Measurement Workshop-PAM'2004*. Lecture Notes in Computer Science (LNCS), Antibes Juanles-Pins, France.

ÖZGEÇMİŞ

1977 yılında Karabük’te doğdu. İlk ve orta öğrenimini Karabük’te tamamladı. 1996 yılında Eskişehir Osmangazi Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümünü kazandı. Lisans eğitimini 2000 yılında tamamlayarak aynı yıl Eskişehir Osmangazi Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalında başladığı yüksek lisans eğitimini 2003 yılında bitirdi. 2002 yılının sonunda Selçuk Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümünde araştırma görevlisi olarak çalışmaya başladı. 2007 yılında doktora eğitimi için 2547 Sayılı Yükseköğretim Kanunu’nun 35. Maddesi gereği Çukurova Üniversitesi Fen Bilimleri Enstitüsü’ne araştırma görevlisi olarak görevlendirilerek İstatistik Anabilim Dalı’nda doktora eğitimine başladı.