T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ALTERNATİF KÜMELEME YÖNTEMLERİNİN KARŞILAŞTIRILMALI ANALİZİ VE BİR UYGULAMASI

MEHMET ŞAMİL GÜNEŞ

YÜKSEK LİSANS TEZİ
İSTATİSTİK ANABİLİM DALI
İSTATİSTİK PROGRAMI

DANIŞMAN
YAR. DOÇ. DR. DOĞAN YILDIZ

İSTANBUL, 2017

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ALTERNATİF KÜMELEME YÖNTEMLERİNİN KARŞILAŞTIRILMALI ANALİZİ VE BİR UYGULAMASI

Mehmet Şamil Güneş tarafından hazırlanan tez çalışması 21.06.2017 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Yard. Doç. Dr. Doğan Yıldız

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Yard. Doç. Dr. Doğan Yıldız

Yıldız Teknik Üniversitesi _____

Prof. Dr. Ali Hakan BÜYÜKLÜ

Yıldız Teknik Üniversitesi _____

Prof. Dr. Ali Karun NEMLİOĞLU

İstanbul Üniversitesi _____

Bu tez çalışması 1649B021506912 numaralı, TÜBİTAK 2210-C Öncelikli Alanlara Yönelik Yurt İçi Yüksek Lisans Burs Programı ile desteklenmiştir.

ÖNSÖZ

Tez çalışmamda planlanmasında, araştırılmasında, yürütülmesinde ve oluşumunda ilgi ve desteğini esirgemeyen, engin bilgi ve tecrübelerinden yararlandığım, yönlendirme ve bilgilendirmeleriyle çalışmamı bilimsel temeller ışığında şekillendiren sayın hocam Yrd. Doç. Dr. Doğan YILDIZ 'a sonsuz teşekkürlerimi sunarım. Ayrıca tez aşamasında maddi desteğini esirgemeyen, bilimin ve bilim insanının destekçisi olan TÜBİTAK 2211 Yurt İçi Lisansüstü Burs Programı 'na teşekkür ederim.

Haziran, 2017

Mehmet Şamil GÜNEŞ

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ viii
KISALTMA LİSTESİ ix
ŞEKİL LİSTESİ x
ÇİZELGE LİSTESİ xii
ÖZET xiii
ABSTRACT xv
BÖLÜM 1
GİRİŞ 1
1.1 Literatür Özeti 1
1.2 Tezin Amacı 6
1.3 Hipotez 7
1.4 Kümeleme Kavramına Giriş 7
1.5 Kümeleme Kullanımına Örnekler 9
1.5.1 Pazarlama Araştırmalarında Kümeleme 9
1.5.2 Astronomi Biliminde Kümeleme 10
1.5.3 Psikiyatri Biliminde Kümeleme 10
1.5.4 Hava Durumu Sınıflandırılmasında Kümeleme 11
1.5.5 Arkeoloji Biliminde Kümeleme 11
1.5.6 Bioinformatik ve Genetik Biliminde Kümeleme 11
1.5.7 Diğer Bilim Dallarında Kümeleme 12
BÖLÜM 2
KÜMELEME ANALİZİ 14
2.1 Kümeleme Analizinin Genel Amacı 14
2.2 Kümeleme Analizinin Aşamaları ve Varsayımları 15
2.3 Kümeleme Analizinde Veri Tipleri 18
2.3.1 Uzaklık (mesafe) Ölçüleri 18
2.3.2 Aralıklı Ölçekli Veriler İçin Uzaklık Ölçüleri 19
2.3.3 Aralık ve Oransal Ölçekli Değişkenler İçin Benzerlik Ölçüleri 19
2.3.3.1 Öklid Uzaklığı 19
2.3.3.2 Manhattan(City-Block) Uzaklığı 21
2.3.3.3 Minkowski Uzaklığı 21

2.3.3.4 Mahalanobis Uzaklığı 22
2.3.3.5 Chebychev Uzaklığı 22
2.3.3.6 Hotelling T^2 Uzaklığı 23
2.3.3.7 Korelasyon Benzerlik Ölçüsü (Correlation Similarity Measure) 23
2.3.4 Kategorik Değişkenler için Benzerlik Ölçütleri 24
2.3.4.1 İkili Veriler(Binary Data) için Ölçütler 24
2.3.4.2 İki Seviyeden Fazla Kategorik Değişkenler İçin Ölçütler 25
2.3.5 Sıralı Değişkenler için Mesafe Ölçütleri 25
2.3.6 Oran Ölçekli Değişkenler için Mesafe Ölçütleri 25
2.3.7 Kategorik ve Sürekli Değişkenler İçin Benzerlik Ölçütleri 26
2.3.8 Vektör Gözlemler 27
2.3.9 Değerlendirme Kriter Ölçütleri 28
2.3.9.1 İçsel Nitelik Kriteri 28
2.3.9.1.1 Hata Kareler Toplamı 28
2.3.9.1.2 Minimum Varyans Kriteri 28
2.3.9.1.3 Serpilme Kriteri 30
2.3.9.1.4 İz Matris Kriteri 30
2.3.9.1.5 Determinant Kriteri 31
2.3.9.1.6 Değişmezlik Kriteri 31
2.3.9.1.7 Condorcent Kriteri 31
2.3.9.1.8 C Kriteri 31
2.3.9.1.9 Kategori Fayda Metriği 32
2.3.9.1.10 Sınır Kesim Metriği 32
2.4. Standartlaştırma 32
2.4.1 Z Değerleri Standardizasyonu 33
2.4.2 0-1 ve -1,1 Standardizasyonu 33
2.4.3 $\mu=1$,$\delta=1$ Standardizasyonu 34
2.5 Benzerlik ve Benzemezlik Ölçütü Seçimi 34

BÖLÜM 3

KÜMELEME ANALİZİ YÖNTEMLERİ 36

3.1 Hiyerarşik Kümeleme Metotları 38
3.1.1 Birleştirici(Agglomerative) Aşamalı Kümeleme Yöntemleri 40
3.1.1.1 Tek Bağlantı (Single Linkage-SLINK) Yöntemi 41
3.1.1.2 Tam Bağlantı (Complete Linkage-CLINK) Yöntemi 42
3.1.1.3 Ortalama Bağlantı (Average Linkage) Yöntemi 43
3.1.1.4 Ward Yöntemi 44
3.1.1.5 Merkezileştirme(Centroid) Yöntemi 45
3.1.1.6 Medyan Yöntemi 46
3.1.2 Ayırıcı Aşamalı Kümeleme Yöntemleri 47
3.1.2.1 Monotetik Ayırıcı Aşamalı Kümeleme Yöntemi 47
3.1.2.2 Politetik Ayırıcı Aşamalı Kümeleme Yöntemi 48
3.2 Bölümleyici Kümeleme Teknikleri 55
3.2.1 K-Means Kümeleme 50
3.2.2 K-Medoid Kümeleme 54
3.3 Modele Dayalı Kümeleme Yöntemleri 55

3.4 Yoğunluğa Dayalı(Density Based) Kümeleme Teknikleri 58
3.4.1 DBSCAN Algoritması 58
3.4.2 OPTICS Algoritması 61
3.5 Izgara Tabanlı(Grid Based) Kümeleme Teknikleri 61
3.5.1 STING(İstatistiksel Bilgi Gridi) Algoritması 62
3.5.2 Wavecluster Kümeleme 63
3.5.3 CLIQUE Kümeleme Yöntemi 64
3.6 Küme Sonuçlarının Değerlendirilmesi 64
3.6.1 Silhouette Endeksi 65
3.6.2 Davies-Bouldin Endeksi 66
3.6.3 Dunn Endeksi 66
3.6.4 Calinski ve Harabasz Endeksi 67
3.6.5 Wilk's Lambda İstatistiği 67
3.7 Kümeleme Analizi Sonuçlarının Görselleştirilmesi 68
3.7.1 Dendogram 68
3.7.2 Ağaç Haritaları (Treemaps) 68
3.7.3 Rectangle Grafigi (Dikdörtgen plot) 69
3.7.4 ReClus Grafigi (ReClus Plots) 69
3.8 Çok Değişkenli Yöntemler Kullanılarak Kümelerin Grafiksel Olarak Tespiti 70
3.8.1 Temel Bileşenler Analizi 70
3.8.2 Çok Boyutlu Ölçekleme 71
3.8.3 Diskriminant Analizi 72

BÖLÜM 4

HİDROLOJİ ÜZERİNE KÜMELEME ANALİZİ UYGULAMALARI 73

BÖLÜM 5

SONUÇLAR ve YORUM 93

KAYNAKLAR 95

EK-A

HAVZA İSİMLERİ VE İSTASYON NUMARALARI 99

EK-B

AYLAR VE OLUŞTURDUĞU KÜMELER 101

EK-C

MDS İÇİN BOYUTLAR 102

EK-D

ANALİZLERDE KULLANILAN R KODLARI 103

ÖZGEÇMİŞ 106

SİMGE LİSTESİ

$a(i)$	*i*. birimin kendi kümesindeki tüm noktalara olan ortalama uzaklığı
$b(i)$	i. birimin diğer kümelerdeki tüm noktalara olan ortalama uzaklıkların minimumu
$D_{k(i,j)}$	k.kümenin , i. ve j. kümelerle olan uzaklığı
d_{ij}	i. birim ile j. birim arasındaki uzaklık
X_{ik}	i. birimin k. değisken değeri
X_i	i. küme vektörü
x_i	i. değisken vektörü
$\bar{X}$	küme merkezi
S	Koveryans matris

KISALTMA LİSTESİ

AB	Avrupa Birliği
AGİ	Akım Gözlem İstasyonu
CBS	Coğrafi Bilgi Sistemi
CLARA	Clustering LARge Applications
CLIQUE	Clustering High-Dimensional Space
DB	Dünya Bankası
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DENCLUE	DENsity-based CLUstEring
DIANA	DIvisive ANAlysis
DSİ	Devlet Su İşleri
EEİ	Elektrik Etüt İdaresi
MDS	Çok boyutlu ölçekleme
MGM	Meteoroloji Genel Müdürlüğü
PCA	Temel bilesenler analizi
STING	STatistical INformation Grid
SYGM	Su Yönetimi Genel Müdürlüğü
TÜBİTAK	Türkiye Bilimsel ve Teknolojik Araştırma Kurumu
TÜİK	Türkiye İstatistik Kurumu
VM	Veri Madenciligi
YY	Yerel Yönetimler

ŞEKİL LİSTESİ

Sayfa

Şekil 1.1 Aynı veri setinin değişik yollarla kümelenmesi 2
Şekil 1.2 Küme içi homojenliği ve kümeler arası heterojenliği gösteren küme grupları 3
Şekil 1.3 Doğal, görünür küme yapısı içermeyen veri grubunun grafiksel gösterimi 4
Şekil 1.4 Gruplara ayrılmış(dissection) veri grubunun grafiksel gösterimi 5
Şekil 2.1 Kümeleme analizi aşamaları 16
Şekil 2.2 Öklid uzaklığının grafiksel gösterimi 20
Şekil 2.3 Manhattan ve Öklid uzaklığının grafiksel gösterimi 21
Şekil 2.4 Mahalanobis uzaklığının grafiksel gösterimi 22
Şekil 2.5 Standardlaştırma işleminin görselleştirilmesi 34
Şekil 3.1 Kümeleme Metotları 37
Şekil 3.2 Dendogram görseli 39
Şekil 3.3 Hayat ağacı(lifetree) görseli 39
Şekil 3.4 Hiyerarşik kümeleme yöntemleri 40
Şekil 3.5 Tek bağlantı tekniği 41
Şekil 3.6 Tam bağlantı tekniği 42
Şekil 3.7 Ortalama bağlantı tekniği 43
Şekil 3.8 Merkezi(Centroid) bağlantı tekniği 46
Şekil 3.9 Başlangıç noktası seçimi(1) 52
Şekil 3.10 Başlangıç noktası seçimi(2) 52
Şekil 3.11 Başlangıç noktası seçimi(3) 52
Şekil 3.12 Başlangıç noktası seçimi(4) 52
Şekil 3.13 Başlangıç noktası seçimi(5) 53
Şekil 3.14 Başlangıç noktası seçimi(6) 53
Şekil 3.15 K-medoid algoritmasının kümeleme adımları 54
Şekil 3.16 Model bazlı yaklaşımlar..... 56
Şekil 3.17 Cobweb sınıflandırma ağaç görseli 57
Şekil 3.18 Yoğunluk ulaşılabilirlik(a) ve yoğunluk bağlılık(b) 60
Şekil 3.19 STING kümeleme için hiyerarşik yapı 63
Şekil 3.20 Küme Sonuçlarının Değerlendirilmesi 65
Şekil 3.21 Dendogram görseli 68
Şekil 3.22 Treemaps görseli 69
Şekil 3.23 Rectangle plot görseli 69
Şekil 3.24 Reclus grafiği görseli 70

Şekil 3.25 TBA ve kümeleme arasındaki ilişki 71
Şekil 3.26 ÇBÖ ve kümeleme arasındaki ilişkiyi gösteren diagram 72
Şekil 4.1 Türkiye Nehir Havzaları Haritası 73
Şekil 4.2 Türkiye için kullanılan akım gözlem istasyonları 74
Şekil 4.3 Dicle-Fırat havzası için ortalama aylık yağış ve sıcaklık grafiği 76
Şekil 4.4 Türkiye ve havza için maksimum ve minimum sıcaklık değişimleri 77
Şekil 4.5 Aylık Ortalama Grafikleri 78
Şekil 4.6 Yıllık Ortalama Grafikleri 80
Şekil 4.7 Fırat Dicle Havzası için aylık betimleyici histogramlar.....80
Şekil 4.8 Fırat Dicle Havzası için istasyonların aylar üzerinde genel dağılımı 81
Şekil 4.9 Fırat Dicle Havzası için istasyonların aylar üzerinde scatter dot grafik görseli 81
Şekil 4.10 Fırat Dicle Havzası için istasyonların aylar üzerinde hiyerarşik kümeleme görselleri . 82
Şekil 4.11 Ward kümeleme yöntemi ile belirlenen bölgeler..... 84
Şekil 4.12 Ward kümeleme yöntemi ile belirlenen bölgeler..... 85
Şekil 4.13 Ward kümeleme yöntemi ile belirlenen bölgeler 86
Şekil 4.14 Yağış serilerinin sınıflandırılmasında K-Ortalamalar kümeleme yöntemi 87
Şekil 4.15 gCLUTO kümeleme çıktısı (1961-1970) 88
Şekil 4.16 gCLUTO kümeleme çıktısı (1971-1980) 89
Şekil 4.17 gCLUTO kümeleme çıktısı (1981-1990) 89
Şekil 4.18 gCLUTO kümeleme çıktısı (1991-2000) 89
Şekil 4.19 Dicle- Fırat Havzası İçin Silhouette indeksi çıktısı 90
Şekil 4.20 Kümeleme sonuçlarından oluşan kümelerin MDS ile gösterimi 91
Şekil 4.21 Türkiye bazında kümeleme analiziyle beraber oluşabilecek havza bölgeleri 92

ÇİZELGE LİSTESİ

Sayfa

Çizelge 2.1 İkili kategorik değişken için ikili sonuçların sayımı …. ……25
Çizelge 2.2 Değerlendirme kriter ölçütleri ……28
Çizelge 3.1 Ortalamalar arası farklılıklar tablosu ……49
Çizelge 3.2 Ortalamalar arası farklılıklar (2) ……49
Çizelge 3.3 Ortalamalar arası farklılıklar (3) ……50
Çizelge 4.1 Varyansların Homojenliği ve Anova ……77
Çizelge 4.2 Varyansların Homojenliği ve Anova ……79
Çizelge 4.3 Ward kümeleme yöntemi ile belirlenen bölgeler için uyumsuzluk ve bölgesel testi. 83
Çizelge 4.4 k-means çıktıları ……85
Çizelge 4.5 K-Ortalamalar kümeleme yöntemiyle belirlenen bölgeler için uyumsuzluk testi…..86
Çizelge 4.6 K-Ortalamalar kümeleme yöntemiyle belirlenen bölgeler için uyumsuzluk testi…..87
Çizelge 4.7 TBA için oluşan ana iki kümenin açıklanma oranı ……91

ÖZET

ALTERNATİF KÜMELEME YÖNTEMLERİNİN KARŞILAŞTIRILMALI ANALİZİ VE BİR UYGULAMASI

Mehmet Şamil GÜNEŞ

İstatistik Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Yar. Doç. Dr. Doğan YILDIZ

Kümeleme analizi metodları, data işleme ve sınıflandırma süreçlerinde çoğunlukla kullanılan çok değişkenli bir analiz tekniğidir. Kümeleme analizinde ana amaç ise işlenen nitelikler bakımından biribirine benzer değişkenleri belirleyerek bunların ana küme yapılarını belirlemektir. Sınıflandırılmış verileri benzerlik veya benzemezlik niteliklerine göre sınıflayarak genel yapıyı özetler. Kümeleme analizi teknikleri aynı akış karakteristiklerine sahip havzaları belirlemede kullanılan önemli bir yöntemdir. Bu tez çalışması havzaların bölgeselleştirilmesi üzerine konusunu oluşturmaktadır. Uygulamada farklı kümeleme yöntemleri kullanılarak havza bazında çalışmalar yapılmıştır. Türkiye bazında entegre havza yönetimi alanında daha iyi bilgiler sağlayacak analizler bu çalışmaya dahil edilmiştir. Kümeleme analizi sonuçlarını daha iyi göstermek amacıyla farklı görselleştirme teknikleri kullanılmıştır. Çalışmadan ve sonuçlardan hareketle yirmibeş su havzası bölgesinin çeşitli karakteristik özelliklerine göre entegre havza ve su yönetimine olumsuz bir şekilde etkilediği görülmüştür.

Kümeleme Analizi metotları kullanılarak oluşturulabilecek yeni havza bölgelerine rehber oluşturabilecek yeni haritalar verilmiştir. Bunun yanında Çoruh, Büyük Menderes, Konya Kapalı ve Sakarya havzalarında yüksek seviyede entegre havza yönetimine ihtiyaç duyduğu ve önem verilmesi gerektiği görülmüştür.

Anahtar Kelimeler: Kümeleme analizi, Entegre Havza yönetimi, Çok değişkenli istatistik yöntemler, Su havzaları, Su yönetimi

ABSTRACT

COMPARATIVE ANALYSIS OF ALTERNATIVE CLUSTER METHODS AND APPLICATION

Mehmet Samil GUNES

Department of Statistics

MSc. Thesis

Adviser: Assist. Prof. Dr. Dogan YILDIZ

Cluster analysis methods are multivariate analysis in data processing and classification with common usage. Main target of this analysis is to obtain variables similar or dissimilar to each other in terms of determine main structure of cluster. The purpose objective of cluster analysis are to support researchers about classified data and summarize the information's. Clustering techniques are useful to identify groups of watersheds which have similar flow characteristics. This thesis deals with regionalization of watersheds. In application for this thesis, we used the different cluster techniques on basin scale for Turkey. Included some cluster analysis that will provide us the more information about integrated basin management in Turkey. Different and useful visualization techniques used in this application for the seeing better cluster analysis results. From the results, it has been observed that, the twenty five water basins are adversely affecting integrated basin and water management according to various characteristics of the water-basins.

New river basin maps are provided that can guide new basin areas that can be created using Clustering Analysis methods. Besides, we can concluded that Coruh, Buyuk Menderes, Konya Kapali and Sakarya Basins needs significantly basin management policies for Turkey.

Keywords: Cluster Analysis, Integrated Basin Management, Multivariate statistics methods, River Basins, Water Management

BÖLÜM 1

GİRİŞ

1.1 Literatür Özeti

Kümeleme teknikleri son yıllarda birçok alanda kullanılan bir metot olmakla beraber önceki çalışmaların birinci bölümünde kümenin tanımlanması ve sınıflandırılması ile ilgili bilgiler yer alacaktır. İkinci bölümde ise kümeleme analizi teknikleri ile ilgili geçmişte yapılmış çalışmalardan örnekler yer alacaktır.

Günümüzde küme kavramının farklı şekillerde açıklanmasının önemli sebeplerinin başında kümeyi oluşturan birimlerin ortak özelliklerinin farklı olmasıdır [1].

Günümüze kadar küme(cluster) teriminin tanımları için genel anlamda araştırmacı ve bilim adamları ortak bir noktada buluşurken, daha sonradan çalışmalarda gelişen teknolojinin etkisiyle beraber küme (cluster) teriminin tanımlanmasında çeşitli farklılıklarda yaşanmıştır.

1964'te Bonner tarafından yapılan küme(Cluster) teriminin tanımlanması şu şekilde yapmıştır. "Küme(Cluster) grup ve sınıfların herhangi bir matematiksel biçime sokulmadan tamamen sezgisel biçimde araştırmacı tarafından oluşturulmasıdır. Kümenin matematiksel biçimde tanımlanmasının zor olmasının yanı sıra, bu tanımlama bile karışıklığa yol açabilir. Kümelemenin tanımı kullanıcının yargılama şekline göre değişir."

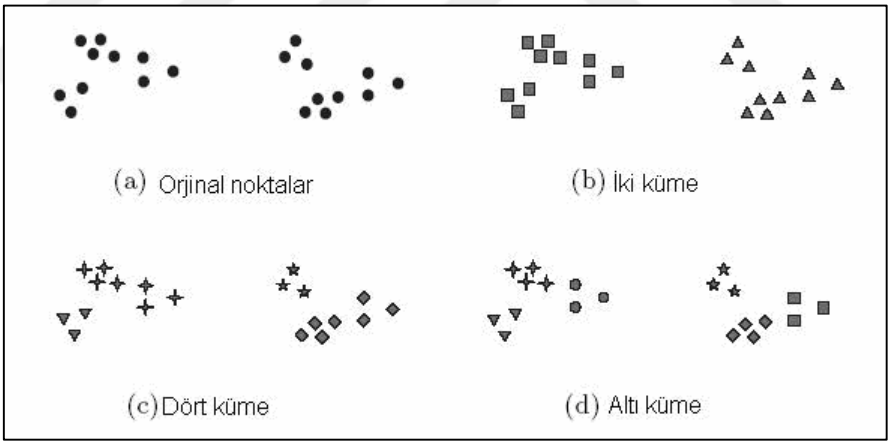
Kendall ve Buckland, kümelerin çeşitli özelliklerini dikkate alarak bir tanımlama yapmışlar ve çalışmalarında kümeyi "İstatistiksel evrenin yakın ögelerinden oluşan bir grup" olarak tanımlamışlardır. Bu tanımlamalarda, birim-birey, nesne-gözlem şeklinde

geçen iki önemli kavram bulunmaktadır. Bireylere ilişkin vasıflar ise karakter, değişken veya özellik olarak birbirleriyle eş anlamlı olarak kullanılmaktadır [2].

Everitt, farklı kaynaklarda, gruplandırma işleminin kullanım alanlarına göre değişik şekillerde isimlendirildiğini bildirmiş ve bunları; Tipleme (Typology), Yığma (Clumping), Q-Analizi(Q-Mode Analysis), Sınıflandırma (Classification), Sayısal sınıflandırma (Numerical Taxonomy), Örüntü Tanımlama (Pattern Recognition) ve Kümeleme Analizi (Cluster Analysis) şeklinde sıralamıştır [3].

Özdamar, popülasyonu meydana getiren bireylere ait birden fazla karakterin gözlemlenmesiyle elde edilen verilerin yorumlanmasının veya bireylerin yapısı hakkında tam olarak bilgi sahibi olmadan sınıflandırma yapılmasının oldukça güç olduğunu belirtmiştir [4].

Yapılan birçok uygulamada küme kavramı çeşitli şekillerde tanımlanmıştır. Küme tanımlamasının daha net anlaşılabilmesi açısından aşağıdaki şekilde farklı noktaların bulundukları yerlere göre veriyi kümelere ayırmanın farklı yolları gösterilmiştir.



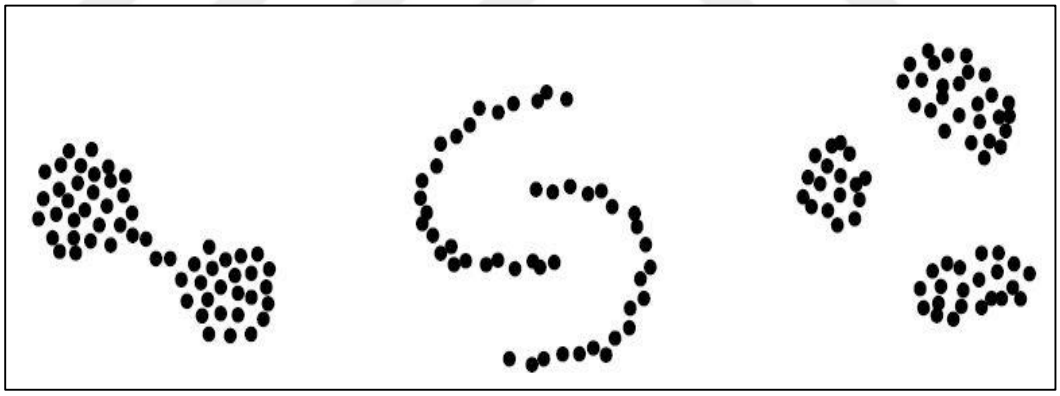
Şekil 1.1 Aynı veri setinin değişik yollarla kümelenmesi

Şekil 1.1(a) de noktaların orijinal noktalar halinde bulunduğu şekli verilmiştir. Şekil 1.1(b) ve Şekil 1.1(d) den hareketle veri setinin iki ve altılı kümelere ayrılış şekli görünmektedir. Şekil 1.1(b) de bulunan iki kümenin Şekil 1.1(d) de üç alt kümeye ayrılması araştırmacının kendi bakış açısına bağlı olmaktadır. Burada Şekil 1.1(c) de bu veri setinin dört gruba ayırabileceğini de görebilmekteyiz. Buradan hareketle küme(cluster) tanımının çeşitli

durumlara göre değişebileceğinin ve küme tanımının kesin ve net olmadığını görebilmekteyiz. En iyi kümelemeyi oluşturabilmenin verinin doğasına ve istenilen sonuçlara bağlı olduğunu söyleyebiliriz.

Öner, bir populasyondaki çok değişkenli gözlem değerlerini tanımlamanın oldukça güç olduğunu ve istenilen bilginin çıkarılıp özetlenmesi gerektiğini belirtmiştir. Bu bakımdan araştırmacı bireyleri kontrol edilebilir gruplara ayırmadıkça, istenilen bilginin kolaylıkla elde edilemeyeceğini belirtmiştir. Böylece elde edilen gruplar arası homojenliği, gruplar arasında da heterojenliği sağlayarak, çok sayıda gözlem değeri yerine birkaç grup ile çalışma kolaylığı sağlanmış olmaktadır.

1964'te Bonner tarafından yapılan açıklama birçok araştırmacı tarafından ikna edici olarak tanımlanmamıştır. Cormack(1971) ve Gordon(1999) 'un yaptığı tanımlamalar buna örnek olarak gösterilebilir. Küme(cluster) kavramının tanımlanmasında Cormack ve Gordon küme içinde iç uyumun(internal cohesion), küme içi homojenliğin(homogenity) ve kümeler arası heterojenliğin önemini vurgulamışlardır. Bu küme(cluster) tanımına aşağıdaki şekillerle daha iyi kavrayabiliriz.

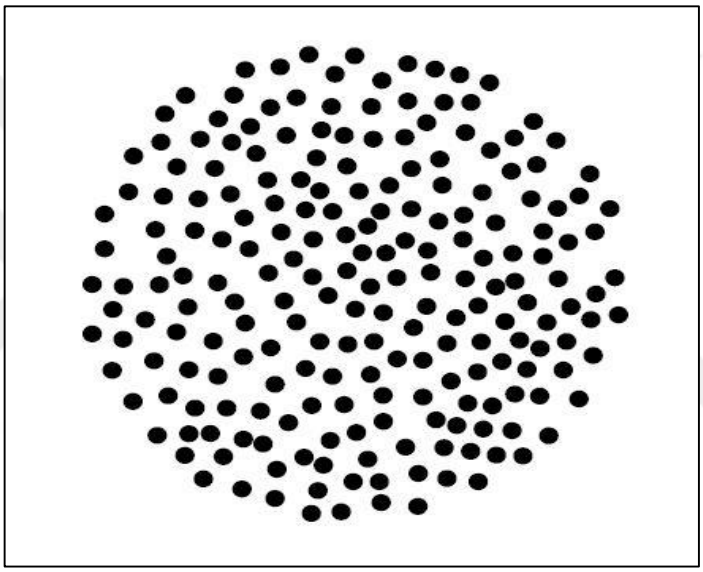


Şekil 1.2 Küme içi homojenliği ve kümeler arası heterojenliği gösteren küme grupları, CRC Press from Gordon, 1980

Matematiksel işlemlere girmeden Şekil 1.2'ye bakarak küme(cluster) kavramının nasıl oluştuğunu anlayabiliriz. Buradan ve önceki araştırmacıların yaklaşımlarından hareketle küme kavramının açıklanmasında neden çeşitlilik olduğunu görebiliriz.

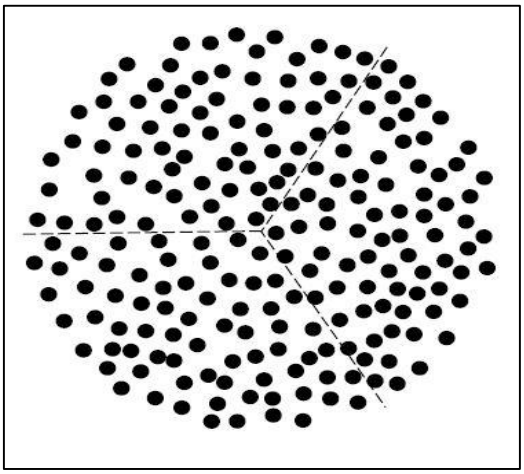
Düzlemsel alanda gösterilen verilerin nasıl küme oluşturduğunu anlamak veya fark etmek tam anlamıyla açıklanabilmiş bir kavram değildir. Noktalar arası göreli mesafe özelliği(relative distance between points) ile kümelerin düzlemde daha belirgin bir

şekilde fark edilmesi mümkündür. Küme kavramına açıklık getirmek isteyen bir diğer araştırmacı ise Hair'dir. Hair'e göre "Kümeleme analizi esas amacı birimlerde bulunan karakteristik özellikleri dikkate alarak gruplamak olan çok değişkenli istatistiksel teknikler grubudur. Kümeleme analizi birimler, küme içerisinde çok benzer biçimde, kümeler arasında farklı olacak şekilde kümeler. Kümeleme işlemi başarılı olursa bir geometrik çizim yapıldığında birimler küme içerisinde çok yakın, kümeler ise birbirinden uzak olacaktır. Hair'in küme tanımı ile Cormack ve Gordon'un yaptığı küme(cluster) kavramı tanımlaması birbirine oldukça oldukça benzer olup, kümeler arası heterojenliği ve küme içi homojenliğin önemini vurgulamışlardır.



Şekil 1.3 Doğal, görünür küme yapısı içermeyen veri grubunun grafiksel gösterimi, CRC Press from Gordon, 1980

Yukarıdaki iki boyutlu verinin oluşturduğu şekle bakan çoğu gözlemciler, burada bir araya toplanmış noktalardan oluşan homojen bir yapının varlığından, yani doğal bir kümeleme yapısının olmadığını söyleyebilirler. Kümeleme yöntemlerinin birçoğu bu homojen yapıyı gruplara bölecektir. Bu ve benzeri homojen yapıları gruplara ayırmanın bazı durumlarda önemli avantajları vardır. Örnek olarak Şekil 1.3'ün bir şehirdeki evlerin coğrafik yerleri olduğunu varsayalım. Ev sayısına göre karşılaştırma yapabilmemizi sağlayacak, evleri bölgelere ayıran bir küme yapısı oluşturulup yorumlanabilir. Bu ayırma aşağıda Şekil 1.4 'te görebiliriz.



Şekil 1.4 Gruplara ayrılmış(dissection) veri grubunun grafiksel gösterimi,
CRC Press from Gordon, 1980

1963 yılında Robert Sokal ve Peter Sneath adlı biyologlar tarafından yayınlanan "Principles of Numerical Taxonomy" adlı kitap Kümeleme analizinin gelişiminde önemli bir kaynak olmuştur. Burada Sokal ve Sneath biyolojik sınıflandırma için kullanılabilecek etkili işlemlerin neler olabileceği konusunda çeşitli çalışmalar yapmışlardır. Bu çalışmalar sonucunda benzerlik derecesi belirlenip, benzer organizmalar bir küme oluşturacak şekilde oransal olarak gruplara yerleştirilmişlerdir. Kümeleme analizi çalışmaları bu araştırmalar ve kitaptan sonra hızlı bir şekilde artmaya başlamıştır.

Edwards ve Cavalli-Sforza, genel kümeleme tekniği dışında farklı bir yaklaşımla, çok değişkenli varyans tekniğini kullanmışlardır. Bu teknikte çok boyutlu uzayda bulunan noktalar arası ilişkilerde, noktaları en yoğun iki kümeye ayırıp, bu işleme sırasıyla devam etmişlerdir. Birimleri ağaç grafiği veya dendogram adı verilen şekillerle birbirinden ayırmışlardır [5].

Kümeleme analizi için Edwards ve Cavalli-Sforza tarafından kullanılan varyans analizi tekniğinin bazen faydalı olabileceğini belirten Scott ve Knott, en doğrusunun kümeleme yöntemlerini kullanmak olduğunu belirtmişlerdir. Bu sisteme dayanarak "olabilirlik oran testi" ni geliştirmişlerdir [6].

Jardine ve Sibson ise veri matrisinden oluşturulan uzaklık ölçüleri (distance measures) yardımıyla sınıflandırma işleminin yapılması gerektiğini söylemişlerdir. Teorik olarak

araştırmacılar kümeleme metodunu sağlayacak şartları ortaya koymuşlardır. Bu şartlar aşağıda maddeler halinde verilmiştir.

1. $d(a,b) \neq 0 \qquad |P|>1$
2. $d(a,b) \geq 0 \qquad a,b \in P$
3. $d(a,b) = 0 \qquad a=b$
4. $d(a,b) = d(b,a) \qquad a,b \in P$
5. $d(a,c) \leq \max \{ d(a,b) , d(b,c) \} \qquad a,b,c \in P$
6. $d(a,c) \leq d(a,b) + d(b,c) \qquad a,b,c \in P$

Bu şekildeki uygulamalarla beraber tek bağlantı yönteminin en iyi yöntem olduğunu savunmuşlardır.

Biyoloji ve genetik alanında çalışan Franco (1997), bitki genetik kaynakları adlı çalışmalarında genel olarak kümeleme analizinin genetik farklılıkların belirlenmesinde kullanıldığını belirtmişlerdir. Burada kümeleme analizi için yapılan çalışmada; belirlenen kritere göre farklı kümeleme yöntemlerinin performansını ortaya koymak amacıyla, istatistiksel çeşitli özellikleri kullanılarak germplasm tipinde çoğalmalar için en iyi şekilde kullanılabilecek sınıflandırma metodu tercih edilmiştir. Bu çalışmada Ward yöntemi kullanılmıştır.

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

1.2 Tezin Amacı

Bu tez çalışmasının ana amaçları, ülkemiz su havza kaynaklarının sürdürülebilir olarak yönetimi ve kullanımı, su havzalarının geliştirilmesi ve sürdürülebilir kullanımı ile ilgili orta ve uzun vadeli kararlara ve yatırım programlarına rehberlik sağlamak, toplumumuzun havzaların ekolojik, ekonomik, sosyal ve kültürel fayda ve hizmetleri ile ilgili ihtiyaç ve beklentilerinin yeterli düzeyde ve sürdürülebilir olarak karşılanması, havza yönetimine iklim değişikliğinin muhtemel etkilerinin ve bu etkilere uyumun dahil

edilmesi, uyum ve mücadele mekanizmalarının geliştirilmesi için yapılacak çalışmalara ortak bir yol göstermektir.

1.3 Hipotez

Bu tez çalışmasında havzalarımızın yönetiminin geliştirilmesine ana hipotez konuları arasında: (i) Su kaynaklarının korunması, geliştirilmesi ve sürdürülebilir kullanımı; (ii) farklı kurum ve kuruluşlarca su havzalarında yürütülen havza yönetim planlaması ve uygulamaları için mevcut 25 nehir havzası ile bunların alt havzaları ile mikro havzalarından oluşan hidrolojik temelli havza tanımlamasının sağlanması; (iii) havza alanları ve yatırımlarının uygun bilimsel kriter ve yöntemlere dayalı olarak önceliklendirilmesi; (iv) Su kullanım verimliliğinin ve tasarrufunun arttırılması; (v) Tarımsal sulama alanlarının havzanın iklim, toprak ve su koşullarına uygun şekilde genişletilmesi, sulama sistemleri ve sulama verimliğinin geliştirilmesi yer almaktadır.

1.4 Kümeleme Kavramına Giriş

Tüm dünyada araştırmacılar, gözlemciler ve bilim insanları için canlıların veya nesnelerin belirli özelliklerine göre sınıflandırılması çok eskilerden beri ilgilenilen bir alan olmuştur. Sınıflandırma (taxonomy); insanların yasadıkları çevreyi, hayatı veya dünyayı daha iyi kavrayabilmesine olanak sağlayan önemli bir araçtır. İlgilenilen birimlerin sayısı arttıkça birimleri sınıflandırmak gün geçtikçe daha da sıkıntılı hale gelmiş ve yeni yöntemler bulma ihtiyacını doğurmuştur. Bu ihtiyaç sonucu kümeleme analizi kavramı ortaya çıkmıstır. Kümeleme Analizi, Linnaeus'un 1753 yılında hayvanların ve burda bitkilerin sınıflandırılması üzerine yaptığı çalışmaya dayanan bir araştırma metodudur. Günümüzde bu analiz tıptan ekonomiye varan bir yelpazede uygulanmaktadır. Daha sonra günümüzde gruplandırmada en sıklıkla kullanılan yöntemlerden olan kümeleme analizi 1939 yılında ilk kez Tryon tarafından kullanılmıştır. Bu kavram olarak 1960 yıllarından sonra hızlı bir şekilde gelişmiş ve geniş uygulama alanlarına sahip olmuştur. Bu alanda kayda değer önemli bir katkıyı, 1963 yılında yayınladığı "Sayısal Sınıflandırma Biliminin Temelleri" adlı kitapla Robert Sokal ve Peter sağlamıştır. Günümüzde bu analiz tıptan ekonomiye kadar varan bir yelpazede kullanılmakta ve uygulanmaktadır. Kümeleme analizinin tıp, biyoloji, kimya, ekonomi, arkeoloji, biyoinformatik ve genetik,

hava durumları, astronomi, veri madenciliği, sosyoloji, eğitim vb. gibi çok uygulama alanları mevcuttur.

Kümeleme analizinde temel amaç, birimler içinde benzer özellikleri gösterenleri aynı kümede(grupta) toplayarak, bu birimler hakkında özet bilgi elde etmektir. Hartigan'ın 1975'te kümeleme analizi ile ilgili yaptığı tanımda şunları demiştir. "Kümeleme analizi doğal sınıflamaları hakkında kesin bir bilgi bulunmayan durumlarda, yığına ilişkin tahminlerin yapılmasında kullanılan teknikler topluluğudur." Bu kısımda hakkında prior bilgi(ön bilgi) bulunmayan grupları sınıflandırırken, birimleri kaç kümeye ayırmak gerektiği kümeleme analizinde üzerinde düşünülmesi gereken bir konudur.

Kümeleme ilgili yapılan tez çalışmasının birinci bölümünde küme, kümeleme kavramının geçmişten günümüze nasıl değiştiği ve bu değişime sebep olan çalışmalara değinilecektir. Küme kavramının anlaşılması için verilen çeşitli örnek ve şekillerle küme kavramı üzerinde durulacak olan birinci bölümde, bunun yanı sıra kümeleme analizi tekniklerinin de geçmişten bugüne nasıl, ne amaçla kullanıldıkları açıklanacaktır. Çeşitli bilim alanlarında kümelemenin hangi amaçla ve nasıl kullanıldığı kümeleme kullanımına örnekler kısmında belirtilecektir.

İkinci bölümde kümeleme analizinin amaçlarından bahsedilerek, kümeleme analizi kavramı bütün yönleriyle ele alınacaktır. Kümeleme analizi için gerekli değişkenlerin nasıl seçileceği, kümeleme analizi için çok önemli bir kavram olan uzaklık ölçüleri kavramı bu bölümde detaylı bir şekilde ele alınacaktır. Değişken türlerine göre incelenecek olan mesafe ölçütleriyle beraber kümeleme analizinde önem arz eden ağırlıklandırma ve standartlaştırma kavramının incelenmesiyle devam edilecektir.

Kümeleme analizi yöntemlerinin beş gruba ayrılarak detaylı bir şekilde inceleneceği bu bölümde, kümeleme analiz yöntemleri incelendikten sonra küme sonuçlarının değerlendirilmesi için kullanılan yöntemler ele alınacaktır. Analiz sonuçlarının kümelemede nasıl görselleştiği ve görsel ögelerin dayandığı temellere değinilecektir.

Çok değişkenli diğer analiz tekniklerinden, özellikle Temel Bileşen Analizi (Prinicipal Component Analysis, TBA), Diskriminant analizi ve Çok Boyutlu Ölçekleme(Multi-dimensional scaling) analizi ile kümeleme analizi arasındaki ilişkiler incelecektir. Bu bölümün sonunda küme sonuçları değerlendirilip diğer bölüme geçilecektir.

Uygulama bölümünde kümeleme analizinin istatistiksel hidroloji alanında uygulamalarına yer verilecektir. Türkiye'de havzaların akım gözlem istasyonlarınca nasıl kümelendiği görmek amaçlanmaktadır. Oluşturulacak kümelerin sayısının önceden bilinmemesi ve daha özet bilgi vermesi nedeniyle illerin kümelenmesi için, yöntem olarak kümeleme analizi tercih edilmiştir.

1.5 Kümeleme Kullanımına Örnekler

Kümeleme analizinin özellikle 1970 yılı ve sonrasında teknolojinin gelişmesiyle beraber kullanımı da önemli bir şekilde artmıştır. Özellikle sosyal bilimlerde, fen bilimlerde, doğa bilimleri, tıp, endüstri ve diğer birçok alanda kullanıldığını görebiliriz. Bu bölümde kümeleme analizinin kullanıldığı çeşitli alanlar ve bu alanlarda kümeleme analizinin kullanım şeklinin nasıl olduğu anlatılacaktır.

1.5.1 Pazarlama Araştırmalarında Kümeleme

Pazarlama sektörünün ana hedeflerinden biri müşterileri homojen gruplara ayırmaktır. Bu sektörde araştırmacı, örnek olarak, belli bir üründen benzer avantajlar(fiyat, ebat vb. faydalar) sağlayan veya benzer faydalar arayan müşterileri nasıl gruplaması gerektiğini sorununu kümeleme analizi sayesinde çözebilir. Araştırmacı aynı zamanda şirketlerin finansal karakterlerini gruplayıp,bu kriterlerin borsa üzerindeki performansıyla inceleme yapmak için kümelemeden yararlanabilir.

Pazarlama araştırmalarında kümeleme yöntemlerinin kullanımına belirli bir örnek 1967 yılında Green tarafından verilmiştir. Bu verilen örnekte pazar araştırması için birçok şehir varken, ekonomik faktörlerden dolayı, bu araştırma şehirlerin hepsine değil de küçük bir bölümüne uygulanmak zorunda kalınmıştır. Green bu araştırmada Kümeleme analizini kullanarak şehirleri, şehir büyüklüğü, şehirdeki gazete satışı ve kişi başına düşen milli gelir gibi 14 değişkenin içinde bulunduğu daha küçük sayılarla gruplara ayırmıştır [7].

Chakrapani tarafından 2004 yılında kümeleme analizi kullanılarak yapılan araştırma kümelemenin pazarlama araştırmalarında kullanımına bir diğer örnek olarak gösterebilir. Burada spor araba üreticileri müşteriler tarafından spor araba almanın sadece müşterinin yaşı ve araç ihtiyacından değil, daha çok kişinin yaşam tarzıyla ilgili olduğunu düşünmektedirler. Üreticiler için spor araba alan müşterilerin yaşam tarzı

modelleri, spor araba sahibi olmayanlara göre farklıdır düşüncesi oluşmaktadır. Sonuç olarak üreticiler kümeleme analizini kullanarak yaşam tarzlarına göre spor araba alacak müşterileri belirlemeye odaklanarak pazarlama stratejisi oluşturmaktadırlar [7].

1.5.2 Astronomi Biliminde Kümeleme

Geniş boyuttaki çok değişkenli astronomik veri setleri içinde birbirinden ayırt edilmesi gereken, nispeten birbirinden farklı gruplara ait olduklarından şüphelenilen nesneler vardır.

Astronomlar, örnek olarak yıldızları hangi istatistiksel kritere göre, nasıl farklı gruplara ayrılacağını düşünmüşlerdir. Burada önemli sorular şu şekildedir. 'İstatistiksel olarak nesneler kaç farklı gruba ayrılmalıdır? Hangi nesne hangi gruba dâhil olmalıdır? Daha önceden sınıfı bilinmeyen nesne var mıdır?' Burada kümeleme analizi kullanılarak nesneler sınıflandırılabilir ve anormal nesneler bu şekilde belirlenebilir. Bu keşiflerden örnek olarak yüksek kırmızı rotasyonlu kuazarlar, tip 2 kuazarlar ve kahverengi cüceler(Brown dwarfs) nesneleri gösterilebilir.

Astronomi alanında bir diğer önemli kümeleme analizi çalışması ise 1996 yılında Faundez-Abans tarafından bulutsu 192 gezegenin kimyasal bileşimini incelemek için yapılmıştır. Bu araştırma sonucunda birbirine benzeyen 6 grup bulmuştur. Aynı zamanda sonuçlar ilginç farklılıklar göstermiştir.

1.5.3 Psikiyatri Biliminde Kümeleme

Psikiyatri de akıl hastalıkları fiziksel vücut hastalıklarından çok daha yanıltıcı olabilir. Kümeleme analizi kullanılarak tanı kategorileri yeniden şekillenebilir ve tanımlanabilir. Buna örnek olarak Wallace ve Boulton'un 1968 yılında 200 hastaya depresyon anketi uygulamıştır. Mental durumu, cinsiyet, yaş ve hastalık süresi değişkenleri ele alınarak yapılan kümeleme analizi sonucunda bir kümeyi içsel depresyon grubunun oluşturduğu gözlenmiştir.

Kümeleme analizi aynı zamanda intihara eğilimli bireyleri bulmak için kullanılmıştır. Paykel ve Rassaby 1978 yılında 236 intihar girişimini çalışma grubu olarak ele almışlardır. Ward yöntemi kullanılarak uygulama sonucunda ortaya 3 grup çıkmıştır ve intihara eğilimli bireylerin oluşturduğu grup kümeleme analiziyle tespit edilmiştir. Buradan

kümeleme yöntemlerinin psikiyatri bilim dalı için vazgeçilmez bir yöntem olduğu anlaşılmaktadır.

1.5.4 Hava Durumu Sınıflandırılmasında Kümeleme

Dünya çapında çok büyük miktarda hava ile ilgili veri bulunmaktadır. Bu tarz verileri kümeleme yöntemleriyle araştırmak klimatoloji ve ilgili birçok bilim dalında yeni bakış açıları getirebilir. Örnek olarak Akdeniz'de hava durumu için günlük tekrarlar kümeleme analizi ile incelenmiş ve Akdeniz'de yağış değişkenliğini açıklayan 20 grup bulunmuştur [8]. Hava durumu araştırmalarında kümeleme yöntemi sıklıkla kullanılmaktadır.

1.5.5 Arkeoloji Biliminde Kümeleme

Arkeolojide, sayısal sınıflandırma bir eserin kullanıldığı zamana ve topluma bağlı olarak farklı kullanım şekillerini belirleyebilir. Buna benzer olarak kümeleme kullanılarak fosilleşmiş maddeler sayesinde eski dönemlerde toplulukların nasıl yaşadığına dair ipuçları bulunabilir [9].

Kümeleme analizinin arkeolojik eserlerde kullanımına örnek olarak 1966'da Hudson tarafından yapılan çalışmayı gösterebiliriz. Burada Hudson tek bağlantı (single linkage-SLINK) ve ortalama bağlantı(average linkage) yöntemlerini kullanarak, demir çağdan(iron age) kalma broşları incelemiştir. İnceleme sonunda kanıtlanabilir arkeolojik güvenilirlikte gruplar bulmuştur [10].

1.5.6 Bioinformatik ve Genetik Biliminde Kümeleme

Bioinformatik dalının son zamanlarda moleküler biyoloji, bilgisayar bilimleri, matematik ve istatistik ile gelişmesiyle beraber büyük bir ilerleme göstermiştir. Genişleyen genomik ve proteomik veri altyapısıyla kümeleme yöntemleriyle ilişkilendirilmiştir. İstatistikçiler gen ekspresyonu(gene-expression) çalışmalarında büyük ve çok önemli rol oynamışlardır. Bu tarz çalışmalarda oluşan veriler ortak karakteristik özellikler taşır ve değişkenler gözlem saylarını aşabilir. Bu tür veriler çok boyutlu (high dimensional) veriler olarak adlandırılır. Birçok istatistiksel yöntem bu tür verileri önemli modifikasyonlar yapmadan kullanamaz. Ama kümeleme analizi bu tür verileri kullanabilir ve ortak özellik gösteren genleri bulabilir. Burada gen aktarımının hangi hastalıklardan etkilendiğini,

hangi genin kalıtsal hastalıkların aktarımından sorumlu olduğunu kümeleme analizi yardımıyla bulabiliriz.

Bioinformatik ve genetikte kümeleme analizinin kullanılmasında bir diğer örnek, Selinski ve Ickstadt adlı bilim insanlarının 2008 yılında tek nükleotit polimorfizleri üzerine yaptığı çalışmadır. Burada araştırmacılar vaka-kontrol çalışmalarındaki kontrol bireyleri ve hastalıklı olanlar arasındaki farkı anlamak için kümeleme analizini kullanmışlardır.

1998 yılında Eisen kümeleme analizini gen ekspresyon verisi üzerinde kullanarak kanser tiplerinin hayatta kalmayla ile olan ilişkisini araştırmıştır.

1.5.7 Diğer Bilim Dallarında Kümeleme

Kümeleme yöntemleri gelişen teknolojiyle beraber araştırmacı ve bilim adamlarına çalışmalarında önemli sonuçlar çıkarmasında yardımcı olmuştur. Bir önceki başlıklarda bahsi geçmeyen çeşitli bilim dallarında kullanımına kısa örnekler verilecektir.

- Sosyal Bilimlerde Kümeleme
  - Demografi: Belirli bölgelerde bulunan idari kayıtlardan yararlanarak çeşitli göç kalıplarının sosyal ve genetik sonuçlarını anlamada kullanılır.
  - Dil Bilim: Dünya üzerinde bulunan dillerin ve dillere ait konuşma çeşitliliğinden yararlanarak, karşılıklı kelimeleri ve detayları anlamada kullanılır.
  - Ekonomi: Bir ülke ekonomisinin sistemsel olarak sektörlerini ortaya çıkarmak, ekonomik yapıyı anlamada kullanılır.
  - Hukuk: Çeşitli kanunların belirlenen Avrupa Konseyi Antlaşmaları hükümlerinden yararlanılarak yönetimsel sınıfları ortaya çıkarmada kümeleme kullanılır.
  - Kütüphanecilik: Bilgi indeksleme, özetlemenin gruplandırılmasında kümeleme analizi yöntemlerinden yararlanılır.
  - Politika: Özellikle ülkede yapılan seçim sonuçlarından yararlanılarak, oylama üzerindeki çelişkileri değerlendirmede kullanılır. Bunun yanı sıra gelecekteki oy verme eğilimlerini belirlemede önemli seviyede kümeleme analizinden yararlanılır.

- Sosyoloji: Toplumsal sorunların çözümlerinde kümelemeden sıklıkla yararlanılır. Toplumlara ait sosyal özelliklerden yararlanılarak, sosyal sınıfların tipolojik şeklini görmede ve tanımlama kümeleme analizinden yararlanılır.

❖ Fen ve Mühendislik Alanında Kümeleme

- Jeoloji: Mineral çıkarma ve benzeri çalışmalarda kümeleme yöntemlerinden jeoloji alanında çalışan bilim insanları sıklıkla yararlanmaktadırlar.
- Kimya: Kimyasal bileşiklerin, yapıların çeşitli yönlerden(fiziksel, kimyasal vb.) incelenmesi ve özelliklerini belirlemede kullanılan yöntemlerden biri kümeleme analizidir.
- Mühendislik: Desenlenmiş makine modellerini kurarak, parmak izi, yüz tanıma, el yazısı karakterlerini tanıma gibi işlemlerde kümeleme analizinden yararlanılır.

❖ Doğa Bilimleri ve Tıp Alanlarında Kümeleme

- Antropoloji: Bu alandan elde edilen veriler kullanılarak, geçmiş dönemlere ait insan ırkları arasındaki genetik farklılıkları belirlemede ve evrimsel tarihleri yeniden düzenlemede kümeleme analizinden yararlanılır.
- Biyokimya: Genler ve proteinlerin çeşitli özellikleri kullanılarak, yeni oluşacak türlerin kökenine ilgilendiren değişimlerin etkilerini bulmada kümeleme analizinden yararlanılır.
- Tıp: Hastalıkların belirtileri kullanılarak, beyincik işlevleri, hastaların mental durumlar, salgıların yayılma durumları ve gelişme modellerinin belirlenmesinde kümeleme analizinden sıklıkla yararlanılır.

❖ Endüstri ve Yönetim Alanında Kümeleme

- Endüstri Ekonomisi: Fabrikaların ürettikleri malları, görevleri ve şekillerinden yararlanılarak, çeşitli tanımlamalar yapmada, gelecek modellemesi yapmada kümeleme yöntemlerinden yararlanılır.
- Yöneylem: Bir organizasyonun dağıtım, işletim, kontrol mekanizmalarından yararlanılarak çeşitli optimizasyon işlemleri yapmada önemli kolaylık sağlar.

BÖLÜM 2

KÜMELEME ANALİZİ

2.1 Kümeleme Analizinin Genel Amacı

Kümeleme analizinin genel amacı, toplanan çok sayıdaki gruplanmamış gözlemlerden oluşan veriyi birimlerin benzerliklerine göre anlamlı gruplardan oluşan özel alt kümelere bölerek veriyi indirgemektir. Böylece araştırmacı en az bilgi kaybıyla, daha net ve anlaşılabilir tanımlı gözlemlere sahip olacaktır. Kümeleme analizi ilk kez 1939 yılında Tryon tarafından kullanılmıştır. 1960'lı yıllardan sonra kullanımı yaygınlaşmıştır. 1963 yılında Robert Sokal ve Peter Sneath'ın yazdığı "Sayısal Sınıflandırma İlminin Temelleri" adlı kitap bu alanda önemli bir adım olmuştur [11].

Sokal ve Peter bu yöntem için yaptığı çalışmalarla beraber kümeleme analizinde kullanılan tekniklerin işlevlerini kısaca aşağıdaki maddelerle açıklayabiliriz.

- İncelenen çalışmadaki n sayıdaki birimi, p tane değişken için belirlenen özelliklerinin gruplar içinde benzer, gruplar arasında benzer olmayan farklı alt kümeler oluşturmasını sağlamaktır.
- İlgilenilen veri seti, çalışma için uygun uzaklık kriteri belirlenerek birimler arası uzaklıkların belirlenmesi önemli bir adımdır.
- Birimleri alt gruplara ayırabilmek bu analiz türünün ana amaçlarından biridir.

Genel olarak çalışma alanlarında araştırmacıların yüz yüze kaldığı en genel sorun gözlenen değerlerin en anlamlı şekilde nasıl sınıflandırılacağıdır. Kümeleme analizinin ihtiyacının temelinde yatan ana sorunun bu olduğunu rahatlıkla söyleyebiliriz. Kümeleme analizde temel amaç grup içi veriler arasında maksimum benzerliği sağlamak ve gruplar arasında ise minimum benzerlikte olmasını sağlamaktır.

Everitt, yaptığı çalışmada, Fisher ve Van Ness ile Jardine ve Sibson'ın çalışmalarını inceleyerek kümeleme analizi için ön bilgi oluşturacak genel bilgilerinin yanında araştırmacılar için çeşitli özel amaçlardan da söz etmiştir.

Genel amacın yanı sıra Everitt tarafından belirlenen kümeleme analizi için aşağıdaki özel amaçlardan da bahsedilebilir [7].

Bu özel amaçları aşağıdaki gibi sıralayabiliriz.

i. Gerçek tipleri belirlemek
ii. Gruplar için ön tahmin yapmak
iii. Hipotezlerin testi
iv. Veri yapısını netleştirmek
v. Veri indirgemek (Veriler yerine kümelerin değerlendirilmesi)
vi. Aykırı değerlerin bulunması

2.2 Kümeleme Analizinin Aşamaları ve Varsayımları

Kümeleme analizi; birkaç adımdan oluşan çalışma yöntemidir. Çalışmayı oluşturacak veri setinin uygun teknolojik platforma konulması kümeleme analizinin ilk aşamasıdır. Öncelikle doğal sınıflamaları hakkında kesin bilgilerin bulunmadığı durumlarda ana kütlelerden çekilen n sayıda birimin incelenen p sayıda değişkene ilişkin gözlem sonucu değerleri elde edilir. Böylece veri matrisi oluşturulmuş olur. Daha sonra verinin ölçüm tipine uygun bir benzerlik ölçüsü ile nesnelerin uzaklıklar matrisi elde edilir. Uygun kümeleme tekniği seçilir ve uygulanır.

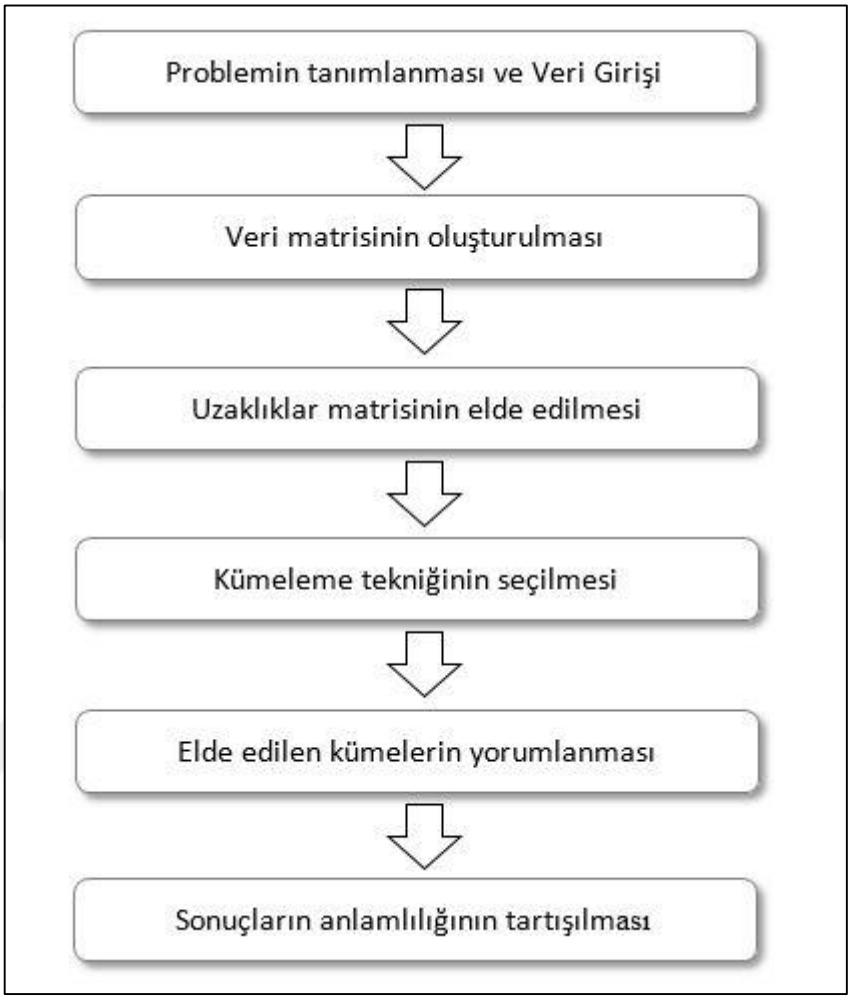
Tekniğin uygulanması sonucu nesneler kümelere ayrılmış olur. Kümeleme sonuçlarının anlamlılığının yorumlandığı aşama ise analizin son aşamasıdır [12].

Kümeleme analizi aşamalarının uygulanması neticesinde küme içi birimler, kümeler arasındaki birimlere göre karakteristik özellikleri bakımından daha çok benzeşirler. Bu uygulamanın sonucu olarak, küme içi homojenlik, kümeler arası heterojenlik sağlanmış olur.

Kümeleme analizi yöntemlerini uygulamak isteyen araştırmacı için analizde en önemli amaç, küme içinde benzerliğin maksimum düzeyde olması, yani küme içi homojenliğin

olması ve kümeler arası minimum benzerliğin yani küme grupları arasında heterojenliğin oluşmasıdır.

Kümeleme analizinin aşamalarının işlem sırasını şekilde gösterilmektedir.



Şekil 2.1 Kümeleme analizi aşamaları

Kümeleme analizinin aşamaları hakkındaki bilgileri aşağıda bulunan maddelerle açıklayabiliriz.

1. Problemin Tanımlanması ve Veri Girişi: Araştırmacının kendisi tarafından çalışmaya problemin tanımlanması belirlenerek başlanır. Çalışmamızda ilgilenilen değişkenler ve birimlerle beraber analizi uygulamak için veri girişi yapılır.
2. Veri Matrisinin Oluşturulması: Veri matrisinde yer alacak kümeleme değişkenlerinin, konumun kuramsal ve uygulamaya yönelik yanları dikkate

alınarak seçilmesi önerilir. Dolayısıyla bu seçimin ilgili konu üzerinde yapılmış araştırmalara, temel bilgiye ve belirli varsayımlara dayalı olarak yapılmasının sağlanması çok önemlidir; çünkü diğer çok değişkenli çözümlemeler gibi kümeleme analizi de değişken seçimi gibi konularda yapılacak yanlış seçimleri ortadan kaldırma becerisine sahip değildir. Bu gibi nedenlerden ötürü, araştırıcıların sonuçları denetleyerek ayırt edici özelliğe sahip olmayan(kümeler arasında istatistiksel olarak fark bulunmayan) değişkenleri çalışmasından çıkartmasında yarar vardır. Diğer yandan, verinin genel eğiliminin dışına çıkan ve de veri yapısında bulunmaması gereken yapıların dikkatle seçilip gerekli işlemlerin yapılması önemli bir basamaktır. Kümeleme analizi sonucunda bir ya da iki değişkenden oluşan bir kümenin oluşması, ilgili gözlem ya da gözlemlerin aşırı gözlem olacağı şüphesini artırır. Verilerin sürekli ya da kesikli sayısal veri türünde olduğu durumlarda gözleme ilişkin Mahalanobis uzaklıklarının hesaplanması ile bu konuda ön bilgi elde edilmektedir; ancak aşırı olarak nitelenen gözlemler veriden hemen çıkartılmamalı, bu gözlemlerin veride olduğu ve olmadığı durumlarda elde edilecek kümelenmeler incelenerek gerekli karara varılmalıdır [13].

3. Uzaklıklar Matrisinin Elde Edilmesi: Kümeleme analizinin bu aşamasında uygulayacağımız yöntemin temelini oluşturacak uzaklık ölçüsünün seçimi önemlidir. Burada dikkat edilmesi gereken önemli bir ayrıntı şudur; bir veriden farklı benzerlik ya da uzaklık matrisleri ile elde edilen kümeler arasında küçük farklılıklar olabileceği gibi bu farklılıkların seçilen kümeleme yöntemine göre de değişebileceği unutulmamalıdır [14].
4. Kümeleme Yönteminin Seçilmesi: Geldiğimiz bu aşamada çalışma için kümeleme yöntemlerinden hangisinin kullanılacağına kara verilir.
5. Elde Edilen Kümelerin Yorumlanması: Araştırmacı çalışma için en uygun kümeleme analiz yöntemini seçtikten sonra analiz sonucu oluşan kümeleri, küme içi homojenlik, kümeler arası heterojenlik vb. durumlara göre değerlendirir.
6. Sonuçların Anlamlılığın Tartışılması: Doğru küme sayısını belirlemede nesnel ve belirgin bir yaklaşım olmadığı için en uygun kümesel yapılanma konusunda yoruma giderken, bir yada iki gözlemli kümelere kuşku ile bakılmalı, kümeleme

değişkenleri açısından kümeler arasında anlamlı farkların olmasına dikkat edilmeli ve sonuçların kuramsal açıdan geçerli olmasına özen gösterilmelidir. Aşamalı kümeleme yöntemlerinde olası küme sayısı çözümleme sonrasında belirlenirken, aşamalı olmayan kümeleme yöntemlerinde önceden belirlenir [14].

- Kümeleme analizinin varsayımları genel olarak aşağıdaki gibi açıklayabiliriz.

Genel olarak birçok istatistiksel yöntemde önemli sayılabilecek varsayımların(normallik, sabit varyanslılık, doğrusallık) kümeleme analizinde önemi azdır. Burada araştırmacı için önemli olan kısım örneklem seçiminin doğru yapılıp yapılmadığı ve değişkenler arasında çoklu bağlantı probleminin olup olmamasıdır. Kümelemede uygulama konusu olan birimlerin evreni temsil etme yeteneği olması önemli bir konudur. Yine aynı şekilde çoklu bağlantı sorunun olduğu kümeler araştırma sonucunda ciddi sorunlar ortaya çıkaracaktır. Bu tür sorunların olduğu durumlarda kümeleme için oluşturulan boyutlarda eşit sayıda değişken alınması uygun bir yol olacaktır. Kümeler arasında benzer değişken sayılarının olması iyi bir yaklaşım olabilir.

Birçok istatistiksel teknikte olduğu gibi kümeleme analizi de aşırı değerlere karşı duyarlıdır. Çalışma verisinin çalışmaya başlamadan önce gözden geçirilmesi önem arz etmektedir.

2.3 Kümeleme Analizinde Veri Tipleri

Kümeleme analizi için araştırmacı tarafından çalışmaya dâhil edilecek verilerden hangilerinin uygulamada kümeleri oluşturmada etkili olup, hangilerinin etkili olmadığı belirlenir. Bu bölümde kümelemede kullanılan veri tipleri ve bunlara göre oluşturulan uzaklık ölçülerinin neler olacağını göreceğiz.

2.3.1 Uzaklık (mesafe) Ölçüleri

Kümeleme analizinde genel amaç küme içi homojenliği, kümeler arası heterojenliği sağlamaktır. Bu da, karakteristik özellikleri bakımında benzer bireylerin aynı kümede toplanmasıyla temin edilmiş olur. Kümeleme birbirine benzer, yakın olan nesnelerin gruplaması olduğu için, iki nesnenin birbirine benzer ya da farklı olduğuna dair bir ölçünün olması gerekmektedir. Bu ilişkiyi ortaya koyabilmek iki tip ölçüm

gerçekleşmektedir: Benzerli ve Başkalık ölçütleri. Benzerlik ve başkalık ölçütleri kümeleme analizinin temel yapısını oluşturmaktadırlar. Veri kümesinde yer alan gözlemleri kıyaslayabilmek için, onların arasında bir mesafe ölçütü veya bir yakınlık bulmaya gereksinim vardır. Gözlem çiftleri arasındaki yakınlık indeksleri, homojen grupları belirleyebilmek için gerekli başlangıç bilgileri sağlar.$x_i$vex_j gözlemleri arasındaki mesafe $d(x_i, x_j)$ ile gösterilmektedir. Farklı değişken niteliklerine göre farklı başkalık ölçütleri kullanılabilmektedir.

2.3.2 Aralıklı Ölçekli Veriler İçin Uzaklık Ölçüleri

Bu tip ölçekle ölçülmüş değişkenlere örnekler verilecek olursa bunlar; uzunluk, sıcaklık, ağırlık gibi değişkenler örnek gösterilebilir. Çalışmada ilgilenilen veri ölçek tipinin kümeleme analiziyle önemli bir ilişkisi vardır.

Araştırmada daha küçük ölçüm birimleri değişkenlerin daha geniş aralıklarda değerlere yönlendirilir ve bu kümeleme yapısında daha büyük etkilerle neden olmaktadır. Etkinin büyüklük derecesi direk olarak kümeleme analizinin etkileme ihtimali olduğundan standardizasyon işleminin kullanılması iyi bir yaklaşım olabilir.

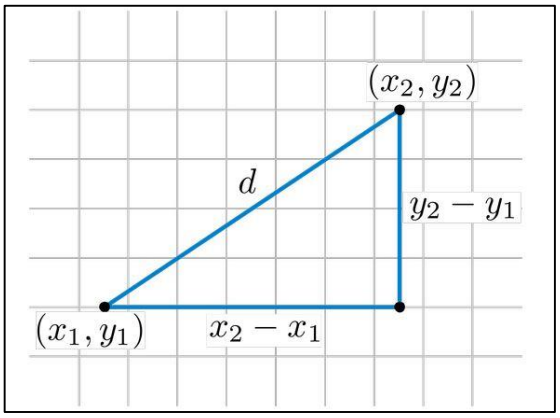
Standardizasyon işleminde dikkat edilmesi gereken değişkenlere verilen ağırlıklardır.

2.3.3 Aralık ve Oransal Ölçekli Değişkenler İçin Benzerlik Ölçüleri

Benzerlik kavramı, araştırmada birimleri gruplandırmak ve ya birbirinden ayırmak için kullanılan bir kümeleme kavramıdır. Diğer adıyla uzaklık ölçüsü de denilen benzerlikler, tek boyutlu veya çok boyutlu olarak yapılabilirler. Bu kavrama örnek vermek istersek fastfood sektöründe kullanılan yiyecekleri gruplandırmak gerektiğini düşünelim. Burada bu gruplandırmayı yapabilmemiz için kalori miktarı, fiyatları, tatları gibi değişkenleri dikkate almalıyız. Araştırmacı iki veya üç boyutlu uzayda çalışıyorsa bu ölçüm, uzayda iki nesnenin arasındaki geometrik uzaklıktır.

2.3.3.1 Öklid Uzaklığı

Öklid(Euclid) uzaklığı, kümeleme analizinde veri nesneleri arasındaki uzaklığı ölçmede kullanılan en yaygın tekniktir.



Şekil 2.2 Öklid uzaklığının grafiksel gösterimi

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ip} - x_{jp})^2} \quad (2.1a)$$

Kümeleme analizi için değişkenler belirli bir önem derecesine göre ağırlıklandırılmışsa;

$$d_{ij} = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + ... + w_p(x_{ip} - x_{jp})^2} \quad (2.1b)$$

p değişkenli bir yapıda i. ile j. gözlem arasındaki genelleştirilmiş Öklid uzaklığı eşitlik (2.2) ile verilir.

$$\mathrm{d_{ij}} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \quad (2.2)$$

Burada,

$x_{ik} = i.$ gözlemin $k.$ değişken değeri, $x = (x_1, x_2, ..., x_i)$ için;

$x_{jk} = j.$ gözlemin $k.$ değişken değeri, $x = (x_1, x_2, ..., x_k)$ için;

$p =$ değişken sayısıdır.

Eşitlik (2.2) den hareketle değişkenlerin ölçüm birimlerinin birbirinden farklı olması Öklit uzaklığı üzerinde etkilidir. Sonuç olarak Öklit uzaklık ölçüsü, değişkenler uzayının ortogonal olduğunu yani değişkenlerin birbirinden bağımsız olduğunu varsayar. Ayrıca teorik olarak Öklit uzaklık ölçüsünün uygulanabilmesi için değişkenlerin en azından aralıklı ölçekte ölçülmüş olması koşulunun sağlanması gerekir [15].

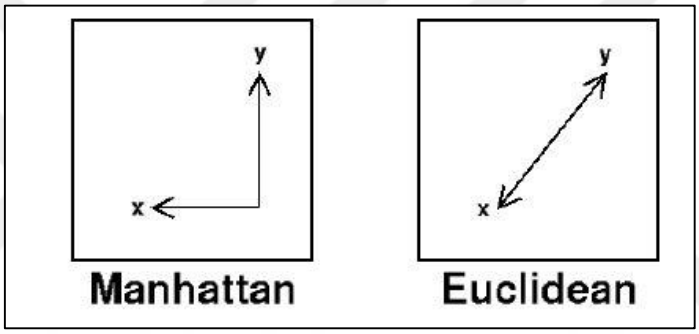
- **Kareli Öklid Uzaklığı**

Kareli öklid uzaklığı normal öklid uzaklığının karesine eşittir.

$$d_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \tag{2.3}$$

2.3.3.2 Manhattan(City-Block) Uzaklığı

City-Block uzaklığı, değişkenler arasında korelasyon olmadığı durumlarda kullanılır. Burada farkların mutlak değerlerinin toplamı olarak tanımlanır. Korelasyonlu değişkenler varsa bu uzaklık kullanımı uygun değildir. Genellikle kesikli sayısal veriler için kullanılması uygundur. City-Block uzaklığı eşitlik(2.4)'te tanımlanmıştır [16].



Şekil 2.3 Manhattan ve Öklid uzaklığının grafiksel gösterimi

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

Eğer değişkenler belli bir önem derecesine göre ağırlıklandırılmışsa;

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

2.3.3.3 Minkowski Uzaklığı

Minkowski uzaklık ölçüsü Öklid ve Manhattan uzaklık ölçüsünün özel bir durumu olup genel bir uzaklık ölçüsüdür.

$$d_{ij} = \left[\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^q \right]^{\frac{1}{q}} \tag{2.5}$$

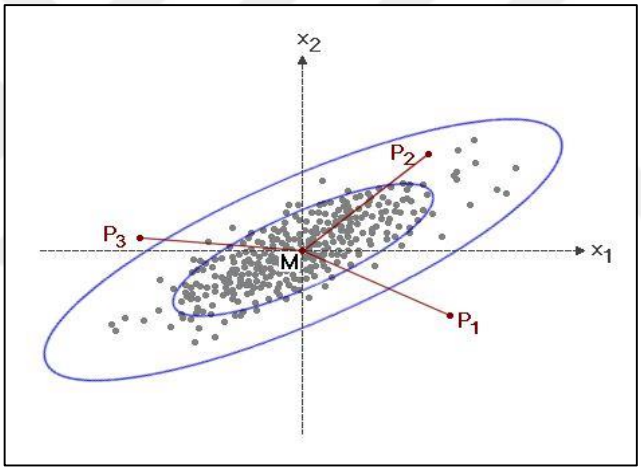
$q = 1$ değeri için sonuç City-Block (Manhattan) ölçüsüne eşit olur.

$q = 2$ değeri için ise, Öklit uzaklığına eşit olur.

Belirtilmesi gereken bir durumda formülünden de görüleceği gibi Minkowski uzaklık ölçüsü genel bir uzaklık ölçüsü, Öklid ve City-Block uzaklık ölçüleri ise Minkowski uzaklık ölçüsünün özel bir durumudur [11].

2.3.3.4 Mahalanobis Uzaklığı

Mahalanobis uzaklığı, varyans-kovaryans matrisini içine almak suretiyle değişkenler arasındaki korelasyonların da dikkate alınmasına olanak sağlar. Değişkenler arasındaki korelasyonlar sıfır olduğunda sözü edilen uzaklık, Öklid uzaklığının karesine eşit olmaktadır. Daha doğru bir ifade ile bu uzaklık ölçüsünün, Öklid uzaklığının genelleştirilmiş hali olduğu söylenebilir [17].



Şekil 2.4 Mahalanobis uzaklığının grafiksel gösterimi [18]

p değişkenli bir analiz çalışmasında *i* ve *k* gözlemleri arasındaki Mahalonobis uzaklık ölçüsü şu şekilde yazılır.

$$Md_{ij} = (x_i - x_j)' S^{-1} (x_i - x_j) \tag{2.6}$$

p değişen sayısı olmak üzere, S, $p \times p$ tipindeki kovaryans matrisini göstermektedir.

2.3.3.5 Chebychev Uzaklığı

Chebychev uzaklığı, iki nokta arasındaki, herhangi bir boyuttaki noktalar arasındaki maksimum uzaklıktır.

Chebychev uzaklığının kullanılması, noktalar arasındaki farklılık tüm boyutlar ele alındığında değil de bireysel boyutlardaki farklılık daha çok yansıtılırsa uygun olabilir. Eşitlik (2.7) de bu uzaklığın formülü verilmiştir [19].

$$d_{ik} = d(x_i, y_k) = Max_i \left| x_{ji} - y_{jk} \right| = \lim_{k \to \infty} \left(\sum_{i=1}^{n} \left| x_{ji} - x_{jk} \right|^b \right)^{1/b}, i \neq k = \overline{1,n} \qquad (2.7)$$

2.3.3.6 Hotelling T^2 Uzaklığı

İki kümenin ortalama vektörlerinin karşılaştırıldığı durumlarda Hotelling T^2 uzaklık ölçütü kullanılır. Bu uzaklık biçiminin hesaplanması için gereken formül eşitlik(2.8) de verilmiştir.

$$T^2 = \frac{n_1 n_2}{n} (\overline{x_i} - \overline{x_j})' S^{-1} (\overline{x_i} - \overline{x_j}) \qquad (2.8)$$

$\overline{x_i}, \overline{x_j}; i, j$. Kümelere ilişkin $p \times 1$ boyutunda karakter ortalamaları vektörünü,

n_1, n_2; 1. ve 2.kümenin birey sayılarıdır. S; bileşik kovaryans matrisi olmak üzere her iki gruba ilişkin kareler ve çarpımlar toplamı matrisleri için;

$S = \frac{1}{n-2}(A_1 + A_2)$ eşitliği ile bulunur. (2.9)

2.3.3.7 Korelasyon Benzerlik Ölçüsü (Correlation Similarity Measure)

Pearson uzaklık ölçüsü x_i ve x_j iki birim arasındaki uzaklık hesaplamalarında değişkenin varyansınıda dikkate almaktadır. x_i ve x_j iki birim olmak üzere, x_i ve x_j birimleri arasındaki uzaklık;

$$r_p = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n} \right] \left[\sum Y^2 - \frac{(\sum Y)^2}{n} \right]}} \qquad (2.10)$$

X ve Y gözlem değerlerini, n ise gözlem sayısını göstermektedir. Değişkenler ağırlıklandırılmamışsa formüle ağırlıklandırma katsayısı eklenerek çözüme ulaşılır [20].

2.3.4 Kategorik Değişkenler için Benzerlik Ölçütleri

Kategorik değişken içeren verilerde, benzerlik ölçüleri yaygın bir biçimde kullanılır. Bu tür kategorik sistemlerde değerleri [0,1] aralığında ölçeklendirilir. Benzerlik değeri sıfır olduğunda bütün değişkenler içinden iki birey arasında farkın maksimum olduğunu gösterir.

2.3.4.1 İkili Veriler(Binary Data) için Ölçütler

İkili sonuçların olduğu kategorik değişkenler çok değişkenli data tipinde en çok kullanılan veri tiplerinden biridir. Bu yapı için seçilen benzerlik ölçüsünün ve veri setinin yanlışlığı, bu ölçüt için önemli derecede etkiye sahiptir. Bu tür değişkenlerde çapraz sınıflandırma(cross classification) sistemi üzerinden mesafe ölçütleri hesaplanır.

İkili kategorik değişkenler için benzerlik ölçütünün hesaplanması aşağıdaki çizelgede gösterilmiştir. İkili değişkenler her iki durumda da eşit değere ve ağırlığa sahiplerse yani sonuçların hangisinin 1, hangisinin 0 ile kodlanacağı tercihi yoksa simetrik ikili değişkenlerdir.

Cinsiyet örneği üzerinde duracak olursak, bu değişkenin 2 kategorik seviyesi(kadın-erkek) vardır. Simetrik ikili değişkenlerde başkalık, simetrik ikili başkalık olarak adlandırılır. Çizelge 2.1' deki eşitlik *i* ve *j* gözlemleri arasındaki başkalık ölçüsüdür.

$$d(i, j) = \frac{r + s}{q + r + s + t} \tag{2.11}$$

q değeri iki gözlemde de 1'e eşit niteliklerin toplam sayısıdır.

t değeri ise iki gözlemde de 0'a eşit niteliklerin toplamıdır.

r ve s her iki gözlem için eşit olmayan durumların toplamıdır.

Çizelge 2.1 İkili kategorik değişken için ikili sonuçların sayımı

			i.Birey	
	Çıktı	1	0	toplam
j.Birey	1	*a*	*b*	*a+b*
	0	*c*	*d*	*c+d*
	toplam	*a+c*	*b+d*	*p=a+b+c+d*

2.3.4.2 İki Seviyeden Fazla Kategorik Değişkenler İçin Ölçütler

Değişkenlerin ikiden fazla seviyesi olduğu(eğitim durumu, göz rengi vb.) durumlarda kullanılır. Burada ikili veriler için uygulanan yönteme benzer olarak, değişkenin her bir seviyesi tek bir ikili değişken gibi kabul edilir. Çok kullanılmayan bir ölçüt olmakla beraber benzerlik katsayısı bulma formülü aşağıda verilmiştir.(s=skor)

$$s_{ij} = \frac{1}{p}\sum_{k=1}^{p} s_{ijk} \tag{2.12}$$

2.3.5 Sıralı Değişkenler için Mesafe Ölçütleri

Sıralı değişkenler objektif olarak ölçülemeyen ölçümlerin durumları hakkında kıyaslama yapabilecek yararlı bilgiler sunabilmektedir. Örneğin bir şirkette işçi, yönetici, müdür, genel müdür ardışık sıralı ifade edilebilmektedir. Sürekli sıralı değişkenler ise bilinmeyen bir ölçekteki sürekli küme gibidir. Örnek olarak spor müsabakalarındaki bronz, gümüş, altın gibi türler gösterilebilir.

Sıralı değişkenin *f* olduğunu farz edelim ve bunun M_f durumu olsun. Dolayısıyla durum sıralaması $1, \dots, M_f$ olacaktır. Sıralı değişkenlerde hesaplanan gözlemler arasındaki başkalık hesabı aralıklı ölçek değişkenlerindekine benzerdir. n gözlem bulunan f sıralı değişken kümesi olduğunu düşünürsek. Başkalık ölçütü şu aşamaları izlemektedir,

1. f nin i. gözlemi x_{if} ve f nin M_f sıralı durumu olsun. Her x_{if} yi derecelerine ilişkin$r_{if} \in \{1, \dots, M_f\}$yer değiştirilir.

2. Her bir sıralı değişken farklı durum sayılarına sahip olabileceği için, eşit ağırlığa sahip olabilmeleri için tüm değişken dizisini $[0,1]$ üzerine yansıtmak gerekebilir. f. değişkende *i*. gözlemin r_{if} sırasını

$$z_{ij} = \frac{r_{if} - 1}{M_f - 1} \tag{2.13}$$

değiştirilerek standardize değerler hesaplanmaktadır.

2.3.6 Oran Ölçekli Değişkenler için Mesafe Ölçütleri

Oran ölçekli değişkenler doğrusal olmayan ölçeklerde pozitif bir ölçüm sağlamaktadır. Üstel ölçeklerde:

Ae^{Bt} ya da Ae^{-Bt} ;

A ve B sabit ve t de zamanı göstermek üzere, bir bakteri popülasyonunun büyümesi örneği verilebilir. Bu tarz ölçümler üç yöntemle hesaplanabilir.

1. Oran ölçekli değişkenlere aralıklı ölçek değişkenlermiş gibi hesaplamaya gidilebilir. Ölçeğin deforme olabilmesi nedeniyle genellikle iyi bir tercih olmamaktadır.
2. Logaritmik dönüşüm: oran ölçekli f değişkeninin *i* gözleminin x_{if} değeri $y_{if} = \log(x_{if})$şeklinde dönüştürülür.y_{if}değeriaralıklı ölçekmiş gibi mesafe ölçütleri kullanılabilmektedir. Bazı oran ölçekli değişkenlere, değişkenin tanımı ve uygulamaya göre log-log veya diğer dönüşümler uygulanabilir.
3. x_{if}değerine sürekli sıralı değişkenmiş gibi uyarlayıp, sıra dizisine aralıklı ölçekmiş gibi davranıp mesafe ölçütleri hesaplanabilmektedir.

2.3.7 Kategorik ve Sürekli Değişkenler İçin Benzerlik Ölçütleri

Aynı anda hem kategorik hem de sürekli tipte değişkenleri içeren veriler için kullanılan çeşitli benzerlik ölçütleri vardır. Burada bir seçenek tüm değişkenleri dikomotize hale getirip, daha sonra ikili veriler için kullanılan benzerlik ölçülerinin kullanılmasıdır. Bir diğer seçenek ise bütün değişkenleri yeniden ölçeklendirerek bütün değişkenleri aynı ölçeğe getirmektir. Devam seçeneği ise her bir değişken için bir benzemezlik ölçütü(dissimilarity ölçütü) oluşturmaktır. Hem kategorik hem de sürekli tipte

değişkenleri içeren veriler için Gower'in 1971 yılında ortaya koyduğu genel benzerlik ölçüsü aşağıda verilmiştir.

$$s_{ij} = \sum_{k=1}^{p} w_{ijk} s_{ijk} \Big/ \sum_{k=1}^{p} w_{ijk} \tag{2.14}$$

s_{ij} = k.değişken tarafından ölçülen i. ve j. bireyler arasındaki benzerlik ölçütüdür.

w_{ijk} = gözlemlerin geçerli olup olmadığına göre 1 ya da 0 değerini alır. Kayıp veri olduğu durumlarda bu değer 0 olarak alınır.

Daha sonra Gower, sürekli tip değişkenler için şu formülü önermiştir.

$$s_{ij} = 1 - \left| x_{ik} - x_{jk} \right| / R_k \tag{2.15}$$

Burada R_k k. Değişken için gözlemlerin genişliğini(range of observations) ifade etmektedir [7].

2.3.8 Vektör Gözlemler

Çok fazla kullanılan bir yöntem olmamakla beraber uygulamalarda karmaşık gözlemlerin bulunduğu sembolik verilerin karşılaştırılması ve kümelenmesi gerekebilir, özellikle bilgi düzeltme, metin madenciliği (tex-mining), biyolojik taksonomi sistemlerinde, tf-idf (term frequency inverse document frequency) belge vektörleri gibi veri yapılarında geleneksel metrik mesafe ölçüleri yerine metrik olmayan benzerlik fonksiyonları ile karşılaşılmaktadır.

$\boldsymbol{x}$ ve $\boldsymbol{y}$ olan iki vektörü karşılaştırabilmek ve benzerlik fonksiyonu $s(\boldsymbol{x}, \boldsymbol{y})$ tanımlayabilmek için birkaç yol bulunmaktadır. Bunlardan biri de kosinüs(cosine) ölçüsüdür.

$$s(x, y) = \frac{x^t . y}{\|x\| \|y\|} \tag{2.16}$$

x^t , x 'in transpozunu ifade eder.

$\|\boldsymbol{x}\|$ ise, x ' in öklidyen normunu ifade eder.

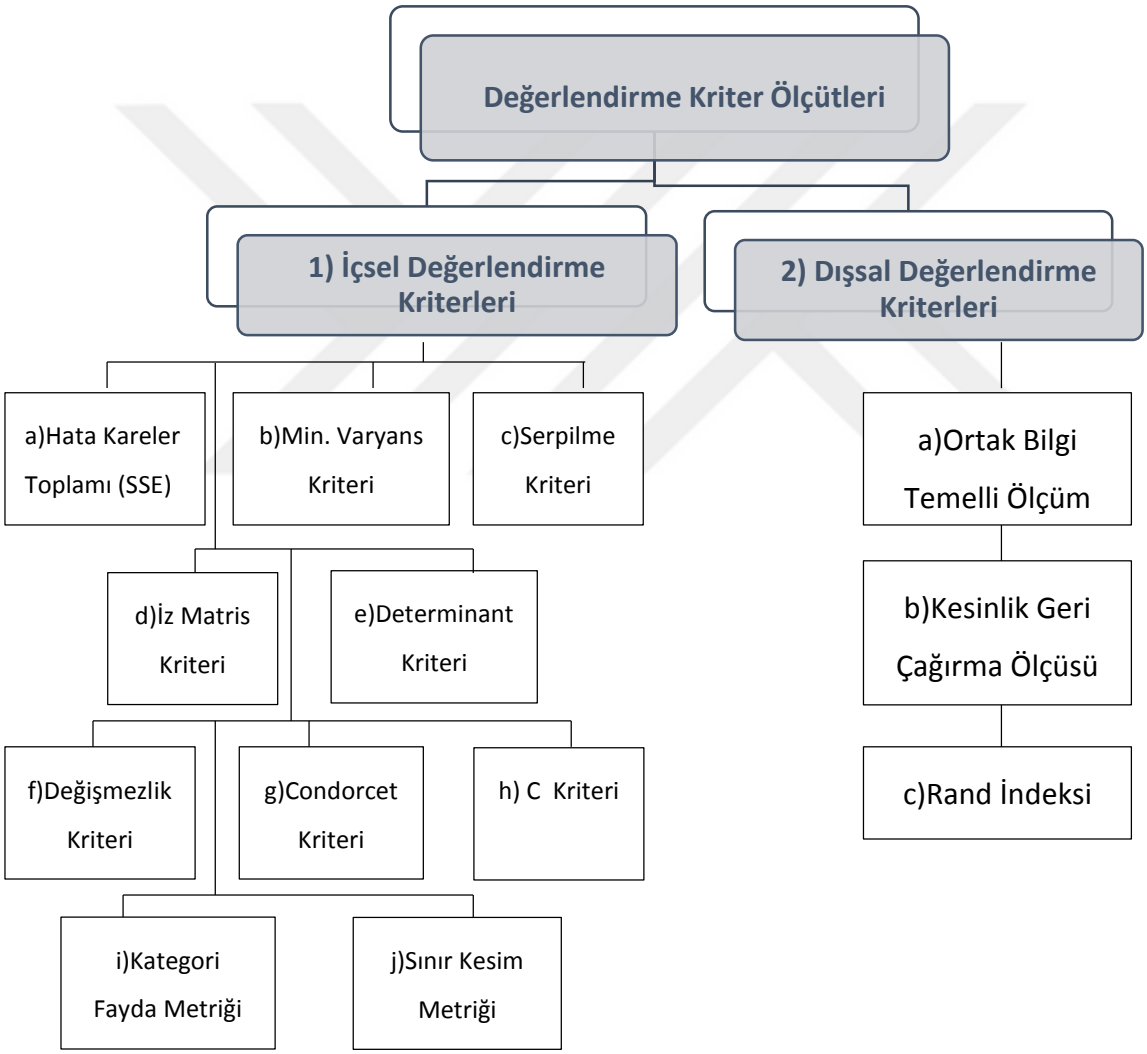
$\|\boldsymbol{y}\|$ ise , y 'nin öklidyen formunu ifade eder.

s ise $\boldsymbol{x}$ ve $\boldsymbol{y}$ olan iki vektörleri arasındaki açının kosinüsüdür. Bu değer rotasyon ve genleşmeye karşı değişim göstermez, fakat doğrusal dönüşümlerde değişim göstermektedir.

2.3.9 Değerlendirme Kriter Ölçütleri

Değerlendirme ölçütleri için kesin bir değerlendirme söz konusu olmamakla beraber, genel olarak bu ölçütleri şu şekilde gösterebiliriz.

Çizelge 2.2 Değerlendirme kriter ölçütleri



2.3.9.1 İçsel Nitelik Kriteri

İçsel nitelik metrikleri belirli benzerli ölçütleri ile kümelerin kompakt bir yapıda olup olmadıklarını ölçmektedir. Bu genellikle küme içi homejenliği ve kümeler arası

ayrılabilirliği yada bu ikisinin kombinasyonunun ölçümüdür. Verinin yanında ekstra bir bilgi gereksinimine ihtiyaç yoktur.

2.3.9.1.1 Hata Kareler Toplamı (SSE)

Sayısal sınıflandırma analizleri için önemli bir ölçüttür.

$$\mathrm{SSE} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \sum_{\forall \mathrm{x_i} \in \mathrm{C_k}} \|\mathrm{x_i} - \mu_{\mathrm{k}}\|^2 \tag{2.17}$$

C_k, k kümesindeki örnek set; μ_k ise k kümesinin vektör ortalaması. μ_k' nın bileşeni:

$$\mu_{\mathrm{k,j}} = \frac{1}{\mathrm{N_k}} \sum_{\forall \mathrm{x_i} \in \mathrm{C_k}} \mathrm{x_{i,j}} \tag{2.18}$$

$N_k = |C_k|$k kümesine ait örnek sayısı.

SSE kriterini minimize eden kümeleme yöntemi, minimum varyans bölümlendirmeleri olarak adlandırılmaktadır [21]. Cebirsel manipülasyon ile,

$$\mathrm{SSE} = \frac{1}{2} \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{N_k} \bar{\mathrm{S}}_{\mathrm{k}} \tag{2.19}$$

$$\bar{\mathrm{S}}_{\mathrm{k}} = \frac{1}{{\mathrm{N_k}}^2} \sum_{\mathrm{x_i}, \mathrm{x_j} \in \mathrm{C_k}} \left\| \mathrm{x_i} - \mathrm{x_j} \right\|^2 \tag{2.20}$$

$(C_k = k\ k\ddot{u}mesi)$olmak üzere SSE kriter fonksiyonu, birbirinden çok iyi ayrılmış kompakt bulut formundaki durumlar için uygun bir kriterdir [22].

2.3.9.1.2 Minimum Varyans Kriteri

SSE' ye ek minimum kriteri $\bar{S}_k$ değerini aşağıdaki ifade ile değiştirerek elde edilmektedir.

$$\bar{\mathrm{S}}_{\mathrm{k}} = \frac{1}{{\mathrm{N_k}}^2} \sum_{\mathrm{x_i}, \mathrm{x_j} \in \mathrm{C_k}} \mathrm{s}\left(\mathrm{x_i}, \mathrm{x_j}\right) \quad \text{ya da}$$

$$\bar{S}_k = \min_{x_i, x_j \in C_k} s\left(x_i, x_j\right) \tag{2.21}$$

Formüllerinden biri kullanılarak bu kritere bakılır.

2.3.9.1.3 Serpilme Kriteri

Serpilme kriteri küme içi ve kümeler arası dağılımın ve her ikisinin toplam dağılım matrisini yansıtan serpilme matrislerinden elde edilmektedir. k. küme için serpilme matrisi

$$S_k = \sum_{x \in C_k} (x - \mu_k)\,(x - \mu_k)^T \tag{2.22}$$

Küme içi serpilme matrisi:

$$S_W = \sum_{k=1}^{K} S_k \tag{2.23}$$

Kümeler arası serpilme matrisi:

$$S_B = \sum_{k=1}^{K} N_k (\mu_k - \mu)(\mu_k - \mu)^T \tag{2.24}$$

μgenelortalama vektör olup:

$$\mu = \frac{1}{m} \sum_{k=1}^{K} N_k \mu_k \tag{2.25}$$

Genel serpilme matrisi:

$$S_T = \sum_{x \in C_1, C_2, \ldots, C_k} (x - \mu)(x - \mu)^T \tag{2.26}$$

Bu üç S_T, S_B, $S_W$serpilme matrisinden üç skalar kriter çıkabilir:

2.3.9.1.4 İz Matris Kriteri

Bir matristeki diyagonaldaki elemanların toplamı. S_W matrisinin izini minimize etmek SSE' nin minimizasyonuna benzerdir. Bu kriter küme içi serpilmeye dayanarak [7]:

$$J_e = \mathrm{tr}[S_W] = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2 \tag{2.27}$$

Diğer bir kriter ise kümeler arasını maksimize eden kriter olabilir:

$$\mathrm{tr}[S_B] = \sum_{k=1}^{K} N_k \|\mu_k - \mu\|^2 \tag{2.28}$$

2.3.9.1.5 Determinant Kriteri

Serpilme matrisinin determinantı kabaca serpilme hacminin karesidir. Küme sayısı boyuttan küçük veya eşitse yada m-c değeri boyuttan küçükse S_B tekil matris olacağı için determinant uygun bir kriter olmayacaktır. S_W matrisinin tekil olmayan bir matris olduğu varsayımında determinant kriter fonksiyonu kullanılabilmektedir.

$$J_d = |S_W| = \sum_{k=1}^{K} S_k \tag{2.29}$$

2.3.9.1.6 Değişmezlik Kriteri

$S_W^{-1}S_B$ matrisinin öz değerleri olan $\lambda_1, \lambda_2, \dots, \lambda_d$serpilme matrisinintemel doğrusal değişmezleridir. Öz değerlerden elde edilen birkaç kriter bulunmaktadır:

$$\mathrm{tr}[S_W^{-1}S_B] = \sum_{i=1}^{d} \lambda_i \tag{2.30}$$

$$J_f = \mathrm{tr}[S_T^{-1}S_W] = \sum_{i=1}^{d} \frac{1}{1+\lambda_i} \tag{2.31}$$

$$\frac{|S_W|}{|S_T|} = \prod_{i=1}^{d} \frac{1}{1+\lambda_i} \tag{2.32}$$

2.3.9.1.7 Condorcent Kriteri

$$\sum_{C_i \in C} \sum_{\substack{x_j, x_k \in C_i \\ x_j \neq x_k}} s(x_j, x_k) + \sum_{C_i \in C} \sum_{x_j \in C_i;\, x_k \notin C_i} d(x_j, x_k) \tag{2.33}$$

$s(x_j, x_k)$ ve $d(x_j, x_k)$ ölçümleri, x_j ve x_k vektörlerinin benzerlik ve başkalık ölçüleridir [23].

2.3.9.1.8 C *Kriteri*

C-Kriteri (Fortier ve Solomon, 1996) **Condorcet Kriteri**

$$\sum_{C_i \in C} \sum_{\substack{x_j, x_k \in C_i \\ x_j \neq x_k}} \left(s(x_j, x_k) - \gamma\right) + \sum_{C_i \in C} \sum_{x_j \in C_i;\, x_k \notin C_i} \left(\gamma - s(x_j, x_k)\right) \tag{2.34}$$

γ eşik değerini göstermektedir.

2.3.9.1.9 *Kategori Fayda Metriği*

Kategori fayda (Gluck ve Corter, 1985) güvenilir bir kümeleme ile doğru bir şekilde tahmin edilen nitelik değerlerinin beklenen değerlerinin artışını tanımlanmaktadır. Bu metrik göreli az sayı içeren nominal değişkenlerin problemini içeren durumlar için faydalıdır.

2.3.9.1.10 *Sınır Kesim Metriği*

Bazı durumlarda kümeleme problemini, sınır kesim problemi olarak tasvir etmek faydalı olabilmektedir. Bazı örneklerde nitelik kalan sınır ağırlıklarının tüm sınır ağırlıklarına oranı olarak ölçülebilmektedir. Eğer kümelerin boyutları için bir kısıt yok ise, en iyi değeri bulmak çok kolaydır. Bu nedenle minimum kesim ölçüsü dengesiz yapıların kontrolü için revize edilir.

2.4. Standartlaştırma

Kümeleme analizinde değişken seçimi araştırma süreçlerinde en önemli adımlardan birini teşkil etmektedir. Fakat bu adımın önemi birçok uygulamada kullanılmamış olmasından dolayı yeterince bilinmemektedir [24].

Benzerlik kavramını temsil eden değişkenler kümesinin bulunması temel problemdir. Değişkenler sınıflandırmaya destek olma amacı ile teoriye uygun olarak seçilmelidir. Çalışmalarda kullanılmak üzere değişkenlerin rasyonel seçimi için teori temeli teşkil etmektedir. Bu durum pratikte uygulanamamaktadır. Teori uygulamada sadece kapalı araştırmalarda sınıflandırmayı destekleyici nitelikte kullanılmaktadır. Bu nedenle bu tür çalışmalardaki değişken uygunluğunu değerlendirmek güç olmaktadır [24].

Birçok istatistiksel analizde uygun yöntemler kullanılarak standardize işlemi uygulanmaktadır. Değişkenlerin normal dağılımı söz konusu olduğunda logaritmik veya diğer dönüşümler sıklıkla kullanılmaktadır. Verilerin aynı ölçekte olmadığı durumlarda ise değerler birim varyans ile standardize edilmektedir. Bazı araştırmacılar alt gruplara ayırmada kümeleme analizinde kullanılan standardize yönteminin uygunluğunu tartışmışlardır.

2.4.1 Z Değerleri Standardizasyonu

Birçok analizin genel varsayımlarından biri olan normal dağılım gösterdiği biline ya da varsayılan oransal ya da aralıklı ölçeklerden biriyle ölçülmüş verilere uygulanır. En çok kullanılan tekniklerin basında gelir. Z skorlarına şekline dönüştürme tekniğinin genel ifade ediliş biçimi aşağıda verilmiştir [7].

$$Z_i = \frac{X_i - \bar{x}}{S} \tag{2.35}$$

2.4.2 0-1 ve -1,1 Standardizasyonu

Değişken değerlerinin standartlaştırılması için bu bölümde sırasıyla $0 \leq x \leq 1$ ve $-1 \leq x \leq 1$ aralıkları için standartlaştırmayı göreceğiz.

0≤ x ≤1 Standardizasyonu: Homojen bir yapıda olmayan veya aşırı değer gibi yapıyı etkileyecek dataların olduğu bir veri sistemine uygulanabilen standardizasyon işlemlerinden biridir. Aşağıdaki formülle ifade edilir.

$$X_i = \frac{x_i - x_{\min}}{R} \tag{2.36}$$

R(Range):Değişim aralığı

Değişim aralığı değişkenin minimum değere göre farkları değişim aralığının bir yüzdesi olarak elde edilmektedir.

-1≤ x ≤1 Standardizasyonu: Bu standardizasyon işlemi veri yapısında pozitif ve negatif değerlerin var olduğu ve bununla beraber aşırı değerlerinde sistemin içinde değerlendirildiği durumlarda kullanılan bir standardizasyon işlemidir. Burada bu tekniğin nasıl formüle edildiği aşağıda gösterilmiştir.

$$x_i = \frac{X_i}{X_{\max}} \tag{2.37}$$

$X_{\max}$: dizideki en büyük değeri ifade eder.

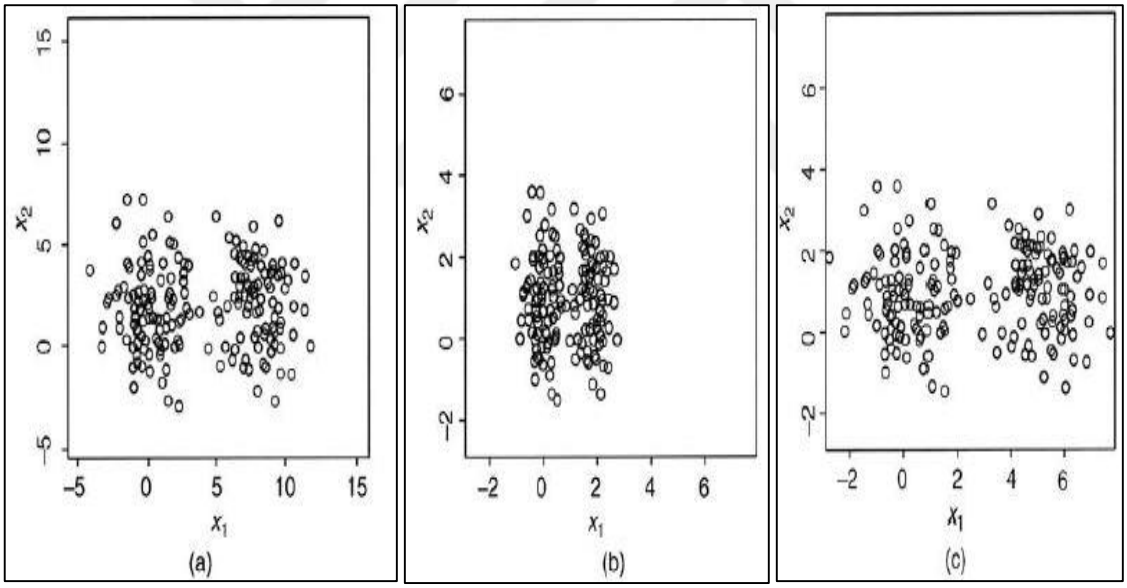
2.4.3 μ=1 ,δ=1 Standardizasyonu

μ=1 Standardizasyonu: Burada dizinin ortalamasını "1" olarak istediğimiz durumlarda kullanabileceğimiz standartlaştırma işlemidir. Matematiksel olarak ifade ediliş şekli aşağıda eşitlik 2.38'de verilmiştir($\bar{x}$: Dizinin ortalamasını gösterir).

$$x_i = \frac{X_i}{\bar{x}} \tag{2.38}$$

δ=1 Standardizasyonu *:* Standart sapmanın(S) "1" olmasını istediğimizde kullandığımız bir standartlaştırma tekniğidir. Formül şeklinde ifade edilmiş hali eşitlik 2.39'da verilmiştir.

$x_i = \frac{X_i}{S}$ şeklinde ifade edilir. (2.39)



Şekil 2.5. *Standardlaştırma işleminin görselleştirilmesi. (a) orijinal ölçekli data. (b) arzu edilmeyen standartlaştırma: total standart sapma baz alınarak işleme alınan ağırlıklandırma. (c) arzu edilen standatlaştırma: gruplar içi standart sapma baz alınarak işleme alınan ağırlıklandırma [7].*

2.5 Benzerlik ve Benzemezlik Ölçütü Seçimi

Genel olarak benzerlik ve benzeşmezlik katsayılarına baktığımızda, sayamayacağımız sayıda katsayı vardır. Bazı yazarlar bu katsayıları çeşitli özelliklerine göre(data ölçek türü,

metriklik durumu vb.) çeşitli kategorilere koyma işlemleri üzerinde çalışmışlardır. Ancak bu kategorize niteliklerinin, katsayılar arasında seçimin kesin olarak yapılmasına yetmiyordu. Bu seçim üzerinde birçok çalışma yapıldıktan sonra, bu seçime yardımcı olacak çeşitli kriterler önerilmiştir [25].

İlk olarak önerilen kriter, yakınlık ölçeği seçiminde verinin doğal yapısı bu seçimi etkileyen en önemli kriter olmalıdır. Bazı koşullar altında, örnek olarak, kantitatif bir veri yapısı ikili(binary) olarak kabul edilebilir.Bir diğer örnek olarak Gower ve Legendre (1986), uygun yenilebilir balık üretiminin yapılacağı nehir bölgelerini belirlemek için, böcek ilaçları ve ağır metallerin eşik değerleri durumlarını baz alarak nehir bölgelerini uygun benzerlik ölçütünü belirleyerek sınıflamıştır.

Yakınlık katsayısının belirlenmesinde en önemli ikinci ölçüt, seçimin data ölçeğine bağlı olarak seçilmesidir. Çizelge 3.2 dikkate alınarak benzerlik katsayısı seçilecekse, veri yapısı ikili(binary) olmalıdır. Sürekli tipte veri yapısı için, mesafe(distance) veya korelasyon tipte benzemezlik katsayısı ilgilenilen nesnenin boyutu ve yapısının uygun olup olmadığı dikkate alınarak seçilmelidir. Sürekli ve ikili tipte veri yapısının beraber olduğu karışık tipteki yapılar içinse, çeşitli sayıda katsayılar önerilmiştir. Ayrıca karışık modeller için katsayılar, sürekli ve kategorik veriler için oluşturulan yakınlık ölçüleri birleştirilerek yeniden yapılandırılabilir [25].

Farklı tipteki benzerlik ya da benzemezlik ölçütleri hesaplamaları aynı tipteki ölçütlerden hesaplanarak çıkarılabilir. Ancak farklı tipte ölçüt kullanımı veya seçimi çalışmalarda farklı sonuçlar doğuracaktır. Buna bağlı olarak bazı özel durumlarda hangi benzerlik veya benzemezlik katsayısının kullanılacağının bilinmesi, araştırmacı için çok önem arz eden bir durumdur. Katsayı seçimi bugün bile tam olarak kesin sonuçlara ulaşmış bir durumda değildir.

BÖLÜM 3

KÜMELEME ANALİZİ YÖNTEMLERİ

Kümeleme analizinde, uygun uzaklık ölçüsü seçildikten sonraki aşama, hangi kümeleme analizi tekniğinin seçileceğidir. Araştırmacı hangi benzerlik veya uzaklık ölçüsünü kullanacağına karar verdikten sonra, kümeleme işleminin nasıl olacağına karar vermek zorundadır. Birimlerin benzerliklerine göre kümelere dâhil edilmesinde kullanılabilecek çeşitli yaklaşımlar vardır. Bu yaklaşımlardan biri, en çok benzer iki birimi aynı gruba atamakla başlayıp tüm birimlerin aynı gruba atanması ile biten hiyerarşik bir yaklaşımdır. Diğer yaklaşım ise tüm verilerin ortalama değerlerine en yakın değerlere sahip birimlerin aynı kümeye atanmasını esas alan yaklaşımdır. En çok kullanılan bu iki yaklaşım dışında diğer yaklaşımlar da mevcuttur. Tüm yaklaşımlarda en önemli ölçüt, kümeler arası farklar ile kümeler içi benzerliklerin maksimum olmasını sağlamaktır.

İlk olarak 1939'da literatüre giren Kümeleme Analizinde; 1950'li yıllarda verilerin hiyerarşik olarak gruplanmasına çalışılmış, 1960'lı yıllarda ise veriyi hiyerarşik olmayan alt gruplara ayıran kümeleme yöntemleri geliştirilmiştir.

Tarih boyunca günümüze kadar kişilerin ve kişilere ilişkin karakteristik özelliklerin belirlenmesinde birçok yöntem geliştirilmiştir. En genel yöntem kümeleme analizi teknikleri olmakla beraber, bu teknikler tarihsel gelişim bakımından farklı şekilde sınıflandırılmışlardır.

Ball(1966) ise kümeleme tekniklerini şu şekilde sınıflandırmışdır [26].

1. Olasılıksal sınıflandırma
2. İşaret belirleyicisi
3. Yığma

4. Özdeğer
5. En küçük tepe arayıcı şeklinde sınıflandırmıştır.

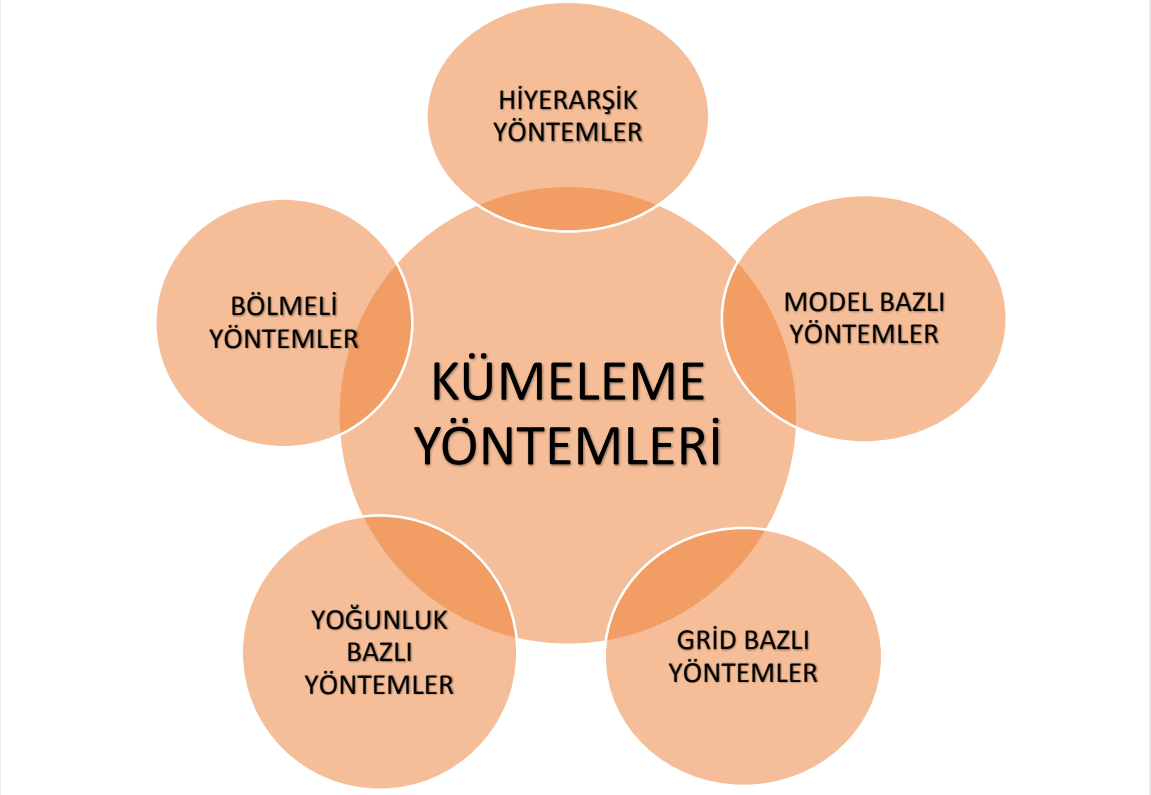
Lance ve Williams ile oluşturdukları kümeleme sınıfları şu şekildedir,

1. Basitleştirici yöntemler
2. Yığma yöntemleri
3. Bölümleme yöntemi
4. Aşamalı yöntemler olarak sınıflandırmıştır [27].

Anderberg ise genel olarak ifade edilen sınıflandırmaları dikkate alarak daha da genelleştirmiştir [11].

1. Aşamalı Kümeleme Yöntemleri
2. Aşamalı Kümeleme Yöntemleri olarak ikiye ayırmıştır.

Genel olarak kümeleme analizi yöntemleri aşağıdaki şekilde verilmiştir.



Şekil 3.1: Kümeleme Metotları

3.1 Hiyerarşik Kümeleme Metotları

Belirli özelliklere göre birimlerin benzerliklerini ortaya koyarak ve bu benzerlikleri esas kabul ederek birimleri doğru kategorilere sınıflandırmak kümeleme analizi yöntemlerinin temelini teşkil etmektedir. Kümeleme analizi teknikleri ile birimler için mevcut andaki durumlar göz önünde bulundurulmaktadır [28].

Kümeleme analizi yöntemlerinden hiyerarşik kümeleme metotları önemli bir yer kaplamaktadır. Bu yöntemin farklı çalışmalar ve araştırmacılar tarafından aşama sıralı kümeleme yöntemi olarak da bilinir.

Kümeleme analizi teknikleri önemli iki başlık altında sınıflanmaktadır. Bunlar hiyerarşik kümeleme yöntemleri ve hiyerarşik olmayan kümeleme yöntemleridir. Hiyerarşik kümeleme yöntemleri birleştirici ve parçalayıcı hiyerarşik kümeleme yöntemleri olmak üzere iki temel grupta incelenmektedir. Literatürde birleştirici hiyerarşik kümeleme yöntemleri parçalayıcı kümeleme yöntemlerine nazaran daha yaygın olarak kullanılmaktadır. Parçalayıcı hiyerarşik kümeleme yönteminde mevcut gözlemlerin tamamı tek bir küme olarak kabul edilmektedir. Küme sayısı bir eksiltilerek benzerlik matrisi yeniden düzenlenir, benzerlik ve uzaklık matrisine göre benzer birimleri bir araya getirilmektedir. n birim aşamalı olarak sırası ile 1,2,3..(n-r)…(n-3), (n-2), (n-1) ve n kümeye yerleştirilir [14, 28].

Hiyerarşik kümeleme yöntemlerinden ilki olan gruplayıcı hiyerarşik yöntemdeki sürecin tam tersi olan yönteme de parçalayıcı hiyerarşik yöntem denir. Bu yöntemde tüm birimleri bir kümede toplanır. Sonra benzer olmayan birimler başka bir kümeye atanır. Bu yolla küme sayısı bir artırılmış olur. Daha sonra her bir birim bir küme oluşturana kadar bu işlem devam ettirilerek kümeler elde edilir. Aşamalı kümeleme yöntemleri için yapılmış olan tanımlamalardan yola çıkarak en genelleştirilmiş hali olduğu kabul edilen, Andersen(1973) tarafından yapılan hiyerarşik kümeleme tanımı bir sonraki açıklamalarda verilmiştir.

Hiyerarşik kümeleme için Andersen tarafından yapılan açıklama şu şekildedir.

Bireylerin kümelenmesinde n birimin;

$$\frac{n(n-1)}{2} = \binom{n}{2} \tag{3.1}$$

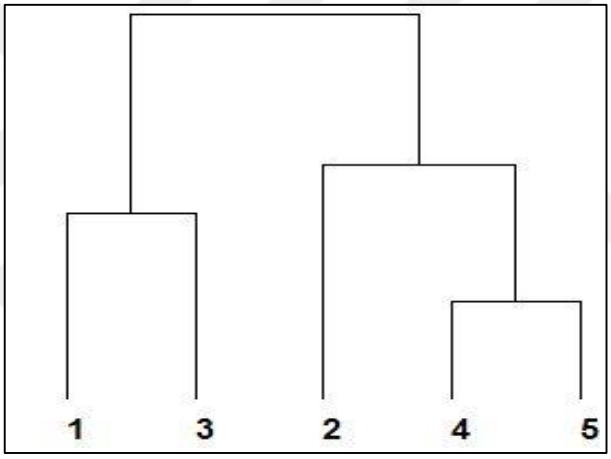
ya da karakterlerin kümelenmesinde p karakterin

$$\frac{p(p-1)}{2} = \binom{p}{2} \tag{3.2}$$

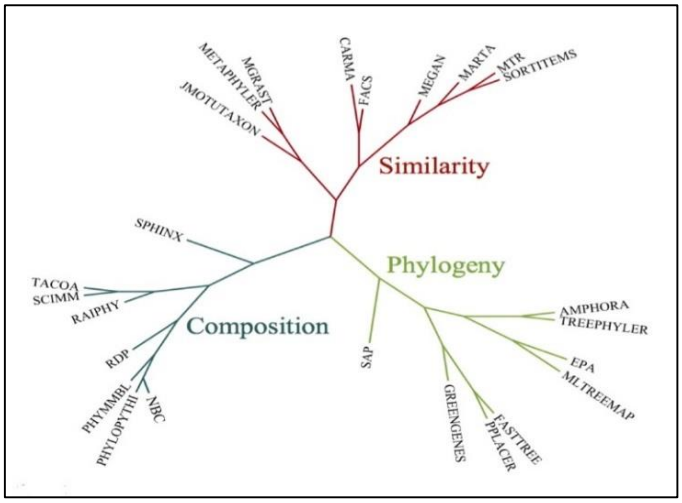
Adet mümkün olan çiftlerin arasındaki benzerlik düzeylerine göre, birbirlerine aşamalı biçimde bağlamayı amaçlayan yöntemlere aşamalı kümeleme yöntemleri denir [11].

Hiyerarşik kümeleme yönteminde, başlangıçta N küme mevcut iken sonunda 1 küme elde edilir (parçalayıcı kümelemede tam tersi). Bu yöntem uygun küme sayısının ne olduğu konusunda bilgi vermez.

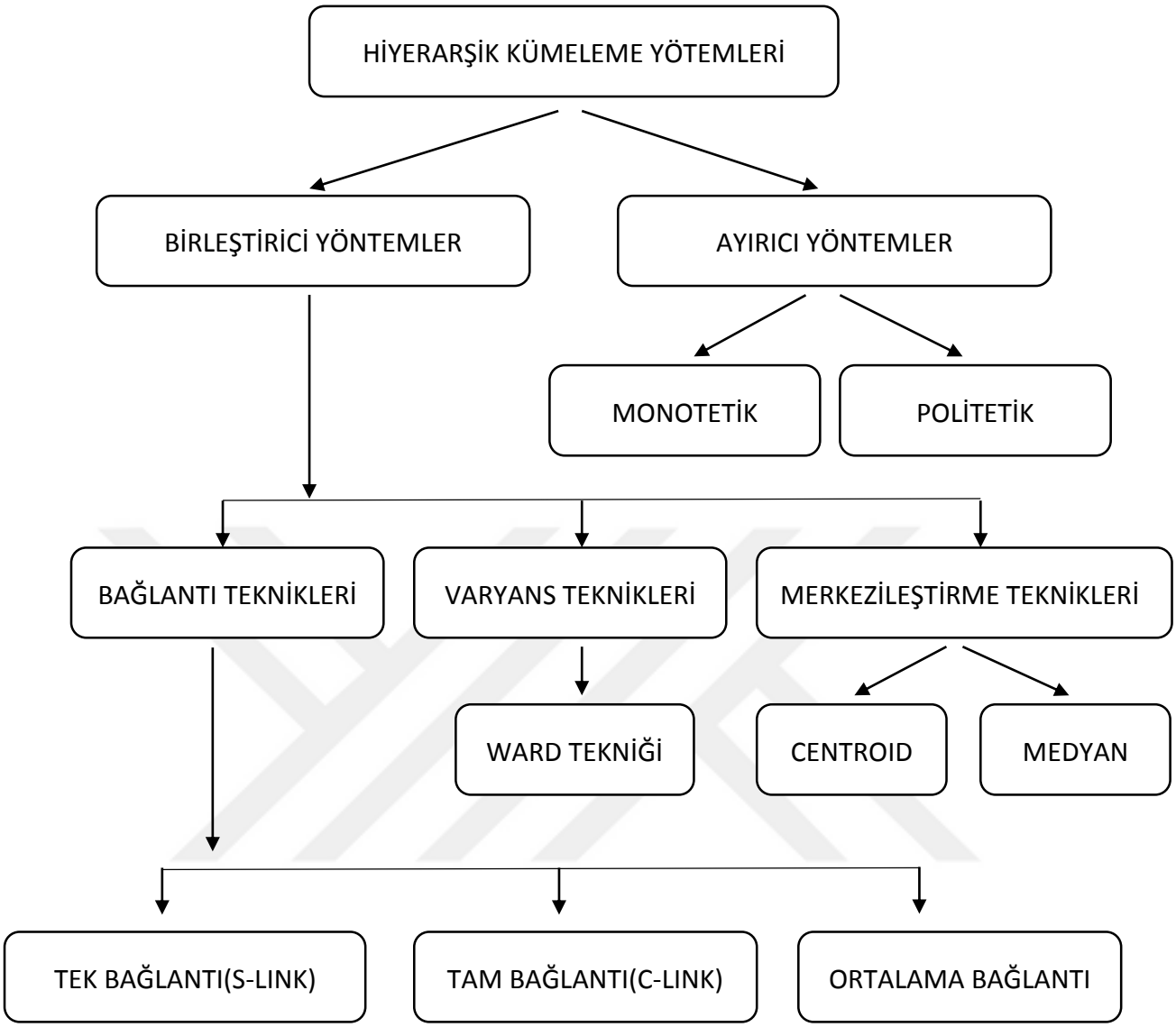
Hiyerarşik kümelemede sonuçları göstermek için sıkça kullanılan dendogram (Yunancadan *dendro* "ağaç" ve *gramma* "çizimi") ve hayat ağacı (lifetree) görselleri bir sonraki bölümlerde detaylı olarak gösterilecektir.



Şekil 3.2 Dendogram görseli



Şekil 3.3 Hayat ağacı(lifetree) görseli [29]



Şekil 3.4 Hiyerarşik kümeleme yöntemler

3.1.1 Birleştirici(Agglomerative) Aşamalı Kümeleme Yöntemleri

Birleştirici aşamalı kümeleme yöntemleri genel olarak en çok kullanılan yöntemdir. Birleştirici aşamalı kümeleme yöntemleri için genel bilgi Bölüm 3.1'de verilmişti. Bu yöntem kendi içinde tek-bağlantı(single-linkage), tam bağlantı(complete-linkage) ve ortalama bağlantı(average-linkage) olarak üç alt gruba ayrılır.

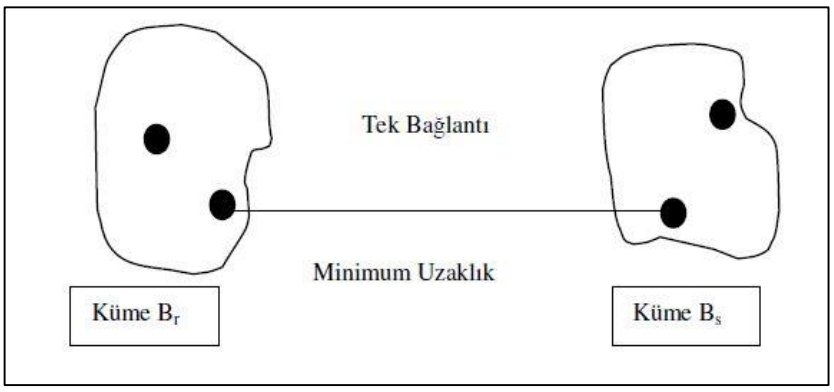
3.1.1.1 Tek Bağlantı (Single Linkage-SLINK) Yöntemi

En yakın komsuluk olarak da bilinen tek bağlantı tekniği, uzaklıklar matrisini kullanarak birbirine en yakın (uzaklık değerleri en küçük) birey ya da nesneleri birleştirmeye dayanmaktadır. Bu teknik çeşitli isimlerlede adlandırılmaktadırlar. Bunlara örnek olarak;

1. En yakın komşu(Nearest neighbor),
2. Minimum yöntem(Minimum distance),
3. Basit bağlantı gibi isimlerde kullanılmaktadır [7].

Tek bağlantı kümeleme tekniğinde her bir birey, nesne veya bunlara ilişkin nitelikler bir grup, bir küme olarak alınır. Başlangıçta n tane küme vardır. Öncelikle tek elemanlı bu kümelerden birbirine en yakın olanlar bir küme, grup haline getirilir. Bu yüzden uzaklık matrisi(D) ve ya *S*={s_{ij}} ilişki tipi benzerlik matrisi içindeki en yakın komşuluğu veya en benzer bireyleri belirleyerek küme haline getirir. Başlangıçta (*u*) ve (*v*) kümeleri gruplandırılmış ise bir sonraki adımda (*u,v*) kümesine en yakın olan (*w*) kümesi arasındaki uzaklık ve benzerlik ölçümleri aşağıdaki eşitlik 3.3 ve eşitlik 3.4 den hareketle hesaplanabilir.

Tek bağlantı tekniğinde iki küme arasındaki uzaklık mümkün tüm kümeler arasındaki minimum uzaklık ile baslar [30]. Yani ilk önce en yakın olan iki küme birleştirilmektedir.



Şekil 3.5 Tek bağlantı tekniği

Nesneler ya da bireyler arasında d_{ij} tipi uzaklık matrisi kullanılmışsa;

$$d(u,v)w = \min\{d_{uw}, d_{vw}\} \tag{3.3}$$

Nesneler ya da bireyler arasında s_{ij} gibi ilişki türü ölçüler kullanılmışsa;

$$S(u,v)w = \max\{S_{uw}, S_{vw}\} \tag{3.4}$$

Eşitlikleri ile hesaplanır. Burada d_{uw} ve d_{vw} uzaklık değerleri, S_{uw} ve S_{vw} benzerlik değerleri sırasıyla (u) ve (w) kümesi ile (v) ve (w) kümelerinin en yakın bağlantıları arasındaki uzaklık veya benzerliktir. İşlemler yapıldıktan sonra ilgili matrisin boyutu (n-1) indirgenir ve böylece matris yenilenir. Teknik tekrarlanarak devam eder [11].

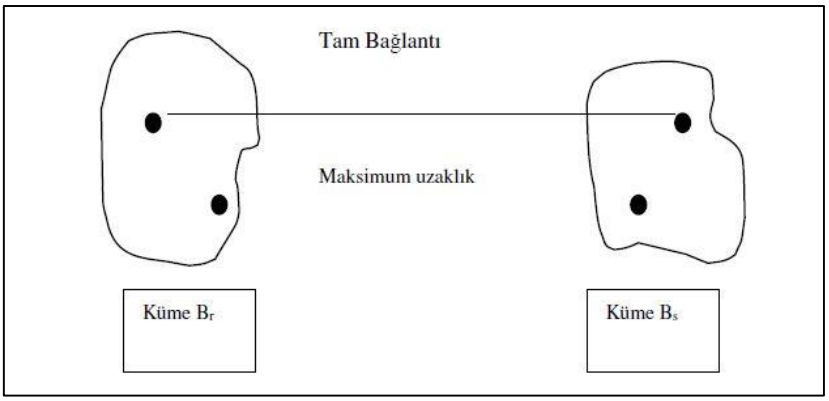
3.1.1.2 Tam Bağlantı (Complete Linkage-CLINK) Yöntemi

Geçmiş çalışmalara ve araştırmacıların çalışmalarını incelediğimizde tam bağlantı yöntemi aşağıdaki isimlerle de anılmaktadır.

1. En uzak komşu(Furthest neighbor analysis)
2. Maksimum yöntem(Maximum distance method)
3. Çap yöntemi(Diameter method)

gibi isimlerle anılmaktadır [4].

Tam bağlantı yöntemi, tek bağlantı yöntemine çok benzemekle beraber bu iki yöntemi birbirinden ayıran en önemli özellik, uzaklık veya benzerlik matrisinin en yüksek değerinin alınmasıdır. Bu yöntemde gruplar belirlenirken öncelikle en uzak komşu niteliğine sahip nesneler ya da bireyler birleştirilir [11]. Tam bağlantı tekniği çok basit bir yöntem olmasına rağmen, küme yapısını belirlemede iyi sonuçlar veren bir tekniktir. Tam bağlantı tekniğinde iki küme arasındaki uzaklık mümkün tüm kümeler arasındaki maksimum uzaklık ile baslar. Yani ilk önce en uzak olan iki küme birleştirilmektedir.



Şekil 3.6 Tam bağlantı tekniği

Nesneler ya da bireyler arasında d_{ij} tipi uzaklık matrisi kullanılmışsa;

$$d(u,v)w = \max\{d_{uw}, d_{vw}\} \tag{3.5}$$

Nesneler ya da bireyler arasında s_{ij} gibi ilişki türü ölçüler kullanılmışsa;

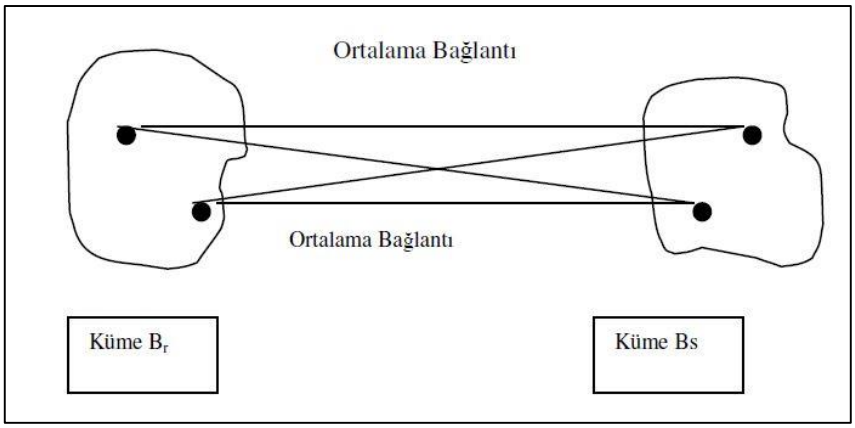
$$S(u,v)w = \min\{S_{uw}, S_{vw}\} \tag{3.6}$$

eşitliklerinden yararlanılır [11].

3.1.1.3 Ortalama Bağlantı (Average Linkage) Yöntemi

Bu yöntemde birimler arasındaki benzerlikler önem kazanmaktadır. İki küme arasındaki uzaklığın hesaplanmasında, ilk kümedeki birimlerin ikinci kümedeki birimlere olan uzaklıklarının ortalaması elde edilmektedir. Diğer yöntemlere karşı avantajlı olduğun durumlar söz konusudur. Ortalama bağlantı kümeleme yöntemi uç değerlere karşı hassasiyeti oldukça azaltmaktadır [31].

Tek bağıntı ve tam bağıntı yöntemleri ele alındığında ortalama bağlantı yönteminin diğer yöntemlerden tek farkı; uzaklıkların hesaplanmasındaki farklılıktır. Bir küme veya grup içindeki birimler, nesneler veya bireyler ile diğer bir küme,grup içindeki birimler arasındaki ortalama (yaygın olarak aritmetik ortalama) uzaklıklar kullanılır.



Şekil 3.7 Ortalama bağlantı tekniği

Ortalama bağlantı yöntemi için kümeler arası birimlerin ortalama uzaklığı eşitlik 3.7 ’de verilen formülle hesaplanır.

$$D_{k(i,j)} = \frac{\sum_{i,j}\sum_{k} d_{k(i,j)}}{N_{(i,j)} N_k} \tag{3.7}$$

Daha sonra minimum uzaklığa sahip küme çifti gruplandırılır ve işlem tekrar edilir.

Genel olarak bağlantı tekniklerinden tek bağlantı, tam bağlantı ve ortalama bağlantı yöntemlerini ele aldığımızda bu üç yöntem algoritmik olarak aynıdır. Birimlerin birbiriyle olan uzaklıkları belirtilen formüllerle hesaplanır ve benzerlikleri en fazla olan kümeler gruplandırılır. Bu yöntemleri birbirinden ayrılan tek ve en önemli yanları ise uzaklıkların hesaplanmasındaki farklılıktır.

3.1.1.4 Ward Yöntemi

"En küçük varyans yöntemi" olarak da ismi anılan Ward tekniği, hiyerarşik kümeleme yöntemlerinden biridir. Ward hiyerarşik kümeleme yöntemi; iki küme birleştirilirken "en az bilgi kaybını" sağlayacak bir yapıya oturtmayı önerir. Bu tekniğin en önemli özelliği birim sayısı(n) çok büyük olduğunda(genellikle n>50), en küçük varyanslardan hareketle diğer hiyerarşik kümeleme yöntemlerine göre daha iyi sonuçlar vermesidir.

Kümeleri oluştururken oluşabilecek bilgi kaybının en az seviyede olmasını sağlamak için "Hata Kareler Toplamı(HKT)" bilgisinden faydalanır. Bu yöntemde amaç hata kareler toplamını minimum seviyede tutmaktır.

Ward yönteminin en önemli bir diğer özelliği ise, benzerlik matrisi elemanlarının Öklid uzaklığının karesi kullanılarak oluşturulmasıdır. *n* bireyden elde edilen p niteliğe ilişkin verilere göre gözlem değerleri *k* kümeye ayrılmak isteyelim. Bu durumda X_{ijk};k. kümenin, *j.* bireyinin, *i.* değerini gösterir. *k* kümesinde n_k birey vardır ve bireylerin *i.* nitelik ortalaması $\bar{x}_{ik}$;

$$\overline{X}_{ik} = \frac{\sum_{j=1}^{j=n_k} X_{ijk}}{n_k} \tag{3.8}$$

olarak gösterilir.

Hata kareler toplamı ise eşitlik 3.9'da verilmiştir.

$$HKT = \sum_{i=1}^{k=n} HKT_k \tag{3.9}$$

Bu işlemler yapıldıktan sonra iki adımda küme birleştirilmesi yapılır.

1. HKT_k =0 olacak şekilde her birim bir küme kabul edilir.
2. İkinci adımda hata kareler toplamında enküçük aartışı sağlayan *(u) ve (v)* kümeleri birleştirilerek *(uv)* kümesi oluşturulur. HKT'daki bu artış ;

$$\Delta HKT_{uv} = HKT_{(uv)} - HKT_{(u)} - HKT_{(v)} \tag{3.10}$$

olarak hesaplanır ve n birim (n-1) kümeye ayrılmış olur.

3. Elde edilen küme sayısı *k=1* oluncaya kadar bir önceki adım(2) tekrarlanarak tüm nesne veya bireylerin hiyerarşik olarak birbirlerine bağlanmaları sağlanır [11].

Bu adımları gerçekleştirirken, her aşamada HKT'ında oluşan minimum artış, birleştirilen kümelerin merkez noktaları arasındaki Öklid uzaklığının karesi ile orantılı olarak gerçekleşmiş olur.

Ward Yöntemi kullanılarak yapılan analiz sonucunda, birimler değişik seviyelerde başarılı bir şekilde bir araya geldikleri "dendogram" adlı şemada gösterilir.

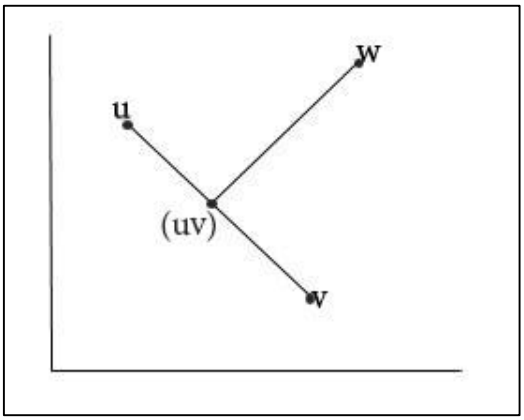
Ward yöntemi aykırı noktalara duyarlı bir yöntemdir [7].

3.1.1.5 Merkezileştirme(Centroid) Yöntemi

Merkezi bağlantı yöntemi, kümeler arası uzaklıklar ve küme merkezleri arasındaki uzaklıklar olarak tanımlanmışlardır. Çünkü kümeler merkez noktaları itibariyle ifade edilmektedirler. Burada kümelerin birleştirilmesi küme merkezi arasındaki uzaklığa göre yapılmaktadır. Kümeler arası uzaklığı en az olanlar kümeleştirilecek ilk birimlerdir [7].

(uv) kümesinin *(w)* kümesine olan uzaklığı eşitlik 3.11'deki gibi hesaplanır.

$$d_{(uv)w} = \frac{N_u.d_{uw} + N_v.d_{vw}}{N_{(uv)}} - \frac{N_u.N_v.d_{uv}}{N_{(uv)}^{\ 2}} \tag{3.11}$$



Şekil 3.8 Merkezi(Centroid) bağlantı tekniği

Anderberg, bu yöntemin en iyi sonucu için D benzerlik matrisinin Öklid uzaklığının karesinin kullanılması ile bulunabileceğini bildirmiştir. Kısaca küme merkezi olarak küme ortalamalarının ağırlıklı ortalaması alınmaktadır.

3.1.1.6 Medyan Yöntemi

Medyan bağlantı yöntemi özellikle birimler ilişkin değerler ordinal(sıralı) ölçek ile elde edildiği veya ölçüm değerleri yerine sıra puanları kullanıldığı durumlarda kümelerin belirlenmesi için kullanılır. Burada gözlem sayılarının birbirinden çok farklı olması durumunda, yeni kümenin merkezi büyük kümeye daha yakın olacak ve diğer kümenin özellikleri hemen hemen kaybolacaktır. Gower, bu durumu sakıncalı bulmuş ve bu durumun düzenlenmesi için birleştirilecek kümelerin eşit boyutlu olmasını ve bu şekilde kümenin boyuttan bağımsız kılınabileceğini söylemiştir. Ya da Gower'in diğer bir önerisi gözlem sayısına bakılmaksızın her bir kümeye birim ağırlıkları verilmesidir [32]. Medyan yöntemi için eğer; *S* ilişki türü benzerlik matrisi ise *(u) ve (v)* kümelerinin birleştirilmesi ile oluşan *(uv)* kümesi başka bir *(w)* kümesi ile birleştirilecek ise bu birleştirmede kullanılacak olan ilişki Eşitlik 3.12'de ifade edilmiştir.

Gower tarafından önerilen ve üzerinde çeşitli çalışmalar yapılan medyan(ortanca) bağlantı yöntemi için Eşitlik 3.12 aşağıda verilmiştir.

$$S_{(uv)w} = \frac{1}{2}.S_{uw} + \frac{1}{2}.S_{vw} + \frac{1}{4}.(1 - S_{uv}) \tag{3.12}$$

Eğer elimizde veriler için uzaklık tipi matris kullanılması gerekiyorsa, medyan bağlantı yöntemi için uzaklık formülü;

$$d_{(uv)w} = \frac{1}{2}.d_{uw} + \frac{1}{2}.d_{vw} - \frac{1}{4}.d_{uv} \tag{3.13}$$

Şeklindedir [8].

3.1.2 Ayırıcı Aşamalı Kümeleme Yöntemleri

Ayırıcı aşamalı kümeleme yöntemleri, grup ayıcı kümeleme yöntemlerinin tam aksine çalışır. Önce bir tane büyük kümeyi ele alır ve başarılı bir şekilde bu kümeyi küçük parçalara bölerek kümeleme işlemini gerçekleştirir.

Ayırıcı kümeleme algoritması olarak bilinen DIANA (DIvisive ANAlysis) algoritması da yine 1990 yılında Kaufman ve Rousseeuw tarafından sunulmuştur. Algoritma veri setinin tamamı ile baslar ve veri setini alt kümelere bölerek ayırır. Daha sonra alt kümeler daha da küçük kümeler ayrılarak algoritma devam eder. Bu teknikte kümeleme süreci toplayıcı kümeleme tekniğinin tersine yukarıdan aşağıya doğru bir süreçtir [33].

Kümelemenin gerçekleşmesi için genel olarak 2^{k-1} hesaplamasıyla, olası küme bölünmesi *k* değeri dikkate alınarak iki alt kümeye bölünür. Ancak, veri ikili yapı içeriyorsa, göreli olarak hesaplanması ve yapılması kolay yöntem olan monotetik ayırıcı yöntem kullanılır. Bu yöntemler genellikle her bir *p* değişkeninin varlığı veya eksikliğine göre kümeleri ayırır. Böylece, her bir bölümde oluşan küme ilgilenilen değişkenin niteliklerini içerir. Bu metotların kullanılabilmesi için kullanılacak verinin tipinin ikili matris şeklinde olması gerekir. Monotetik terimi, kümeleme için ayırmanın her bir aşamasında tek değişkene göre bölümlenmesinden gelir. Politetik terimi ise, kümeleme için ayırmanın her bir aşamasında bütün değişkenlerin kullanılmasından gelir. Gruplayıcı yöntemlere göre ayırıcı yöntemler daha az tercih edilen bir yöntemdir [7].

3.1.2.1 Monotetik Ayırıcı Aşamalı Kümeleme Yöntemi

Monotetik ayırma yönteminde en önemli adım değişken seçimidir. Burada değişken seçimi kümenin homojenliğine veya diğer değişkenlerle birlikteliklerine göre

değerlendirilir. Amaç ayırımların sayısını minimum yapmaktır. Bu yöntemde homojenlik kriteri olan bilgi içeriği (information content) C ile temsil edilir.

$$C = pn\log n - \sum_{k=1}^{p}\left[f_k \log f_k - (n - f_k)\log(n - f_k)\right] \tag{3.14}$$

Monotetik ayırıcı aşamalı kümeleme yönteminin bir avantajı olarak da kümelemenin her bir aşamasında hangi değişkenin etkili olduğunun bilinmesidir. Ancak, bu yöntemin genel problemi kümelemenin değişkenlerin belirli niteliklerinin etkisi altında kalmasıdır.

Monotetik ayırma yöntemlerine alternatif olarak yeni bir ayrıcı yöntem olan sınıflandırma ve regresyon ağaçları(Classification And Regression Tree-CART) yöntemi Piccarreta ve Billari(2007) tarafından önerilmiştir. Bu yöntemin ana amacı "saf(pure)" kümeler elde etmektir. Bu yeni yöntem için, kümeleri ayırmada en önemli etmen bu analiz algoritmasında kullanılan sıklıklardan oluşan veri altyapısıdır.

3.1.2.2 Politetik Ayırıcı Aşamalı Kümeleme Yöntemi

Politetik ayırıcı kümeleme yöntemi, monotetik ayırıcı kümeleme yöntemine göre gruplayıcı kümeleme yöntemlerine daha çok benzerdir. Burada ayırma işlemi grup içinde nesne ya da birime en uzak olan birimin seçilmesiyle başlar ve bu ayrımı başlangıç olarak seçer. Bu işlem devam ederek başlangıç noktalarına yakın olan birimler ilgili kümeye dahil olur.

Politetik ayırıcı yöntemi daha iyi anlamak için aşağıda örnek verilmiştir. Burada bir politik partinin içinde bulunan düşünce bazlı iç çatışma için mesafe matrisi verilmiştir. Düşünce bazlı bu rahatsızlıktan memnun olmayan üyeler partiden ayrılmak isteyip yeni bir parti kurmayı düşünmektedirler. Aşağıdaki veri matrisinde parti üyelerinin diğer üyelerle daha çok anlaşmazlığa düştüğü veri matrisi yapısıdır.

$$D = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{pmatrix} 0 & & & & & & \\ 10 & 0 & & & & & \\ 7 & 7 & 0 & & & & \\ 30 & 23 & 21 & 0 & & & \\ 29 & 25 & 22 & 7 & 0 & & \\ 38 & 34 & 31 & 10 & 11 & 0 & \\ 42 & 36 & 36 & 13 & 17 & 9 & 0 \end{pmatrix}$$

Ayırma işlemi ortalama uzaklığı maksimum olanların belirlenmesiyle başlamıştır. 1.üye başlangıç olarak kabul edilip diğer üyelerle(başlangıç grubu(1), diğer üyeler(2,3,4,5,6,7)) arasındaki ortalama mesafeler hesaplanmıştır. Bu hesaplama için örnek çizelge aşağıdadır.

Çizelge 3.1 Ortalamalar arası farklılıklar (1)

Parti üyeleri (ana grup birimleri)	**Ayırıcı gruba ortalama mesafe(A)**	**Ana gruba ortalama mesafe(B)**	***B-A***
2	10.0	25.0	15.0
3	7.0	23.4	16.4
4	30.0	14.8	-15.2
5	29.0	16.4	-12.6
6	38.0	19.0	-19.0
7	42.0	22.2	-19.8

Çizelge 3.1 ’den hareketle maksimum farklılığın 16.4 değeri ile 3.parti üyesine aittir. Şimdi ayırıcı grubunu birinci ve üçüncü parti üyesi oluşturmaktadır(ayırıcı grub(1,3), diğer üyeler(2,4,5,6,7)). Daha sonra aynı işlemler tekrarlanır. Çizelge 3.2 ise bir sonraki işleme ait ortalamalar arası farkı vermektedir.

Çizelge 3.2 Ortalamalar arası farklılıklar (2)

Parti üyeleri (ana grup birimleri)	**Ayırıcı gruba ortalama mesafe(A)**	**Ana gruba ortalama mesafe(B)**	***B-A***
2	8.5	29.5	21.0
4	25.5	13.2	-12.3
5	25.5	15.0	-10.5
6	34.5	16.0	-18.5
7	39.0	18.7	-20.3

Bu ayırma işleminde 2.parti üyesi de bu gruba girdi ve işlemler yine tekrarlanıp devam ederek Çizelge 3.3 ’de görülen değerler elde edildi.

Çizelge 3.3 Ortalamalar arası farklılıklar (3)

Parti üyeleri (ana grup birimleri)	Ayırıcı gruba ortalama mesafe(A)	Ana gruba ortalama mesafe(B)	*B-A*
4	24.4	10.0	-14.3
5	25.3	11.7	-13.6
6	34.3	10.0	-24.3
7	38.0	13.0	-25.0

Farklılıkların tamamı negatif olduğundan politetik ayırıcı yöntemlerine göre kümelerimiz oluşmuş oldu. Birinci kümeyi 1,2 ve 3. Parti üyeleri oluştururken ikinci küme ise 4,5,6 ve 7. Parti üyeleri olacak şeklide ayrılmıştır [6].

3.2 Bölümleyici Kümeleme Teknikleri

Bölünmeli kümeleme algoritmaları k giriş parametresini alarak n tane nesneyi k tane kümeye böler. Bu teknikler, dendogram gibi iç içe bir kümeleme yapısı üzerinde çalışmak yerine tek-seviyeli kümeleri bulan işlemler gerçekleştirir. Bütün teknikler merkez noktanın kümeyi temsil etmesi esasına dayanır. Bölümlemeli yöntemler hem uygulanabilirliğinin kolay hem de verimli olması nedeniyle iyi sonuçlar üretirler. Çalışmada kullanılan k-means ve k-medoids bölümlemeli kümeleme algoritmaları aşağıda açıklanmıştır.

3.2.1 K-Means Kümeleme

Aşamalı olmayan kümeleme metotları arasında en yaygın olarak kullanılan yöntem K-ortalamalar(K-means) tekniğidir. Bu teknikle, küme içerisindeki kareler toplamını en küçük biçimde k sayıda kümeye bölmektedir.

MacQueen " *k* – ortalama(k-means)" terimini kullanırken şunu belirtmiştir. Her bir birimin, nesnenin ya da bireyin en yakın merkezli (ortalama) kümeye girmesi süreci anlamında kullanmıştır. *k* ortalama yöntemi, gözlemleri kümelerin önceden belirlenen sayısına göre küme haline getirerek süreci başlatır. Böylece her biri tek gözlemden oluşan *k* tane küme ile işleme başlanır ve her bir yeni gözlem en yakın ortalamalı gruba ilave edilerek ait oldukları gruplar belirlenir. Gruba yeni bir gözlem ilave edildikten sonra sonra küme ortalaması yeniden hesaplanır ve süreç yanı şekilde işlemeye devam eder.

Bu süreç tüm gözlemler gruplara atanıncaya kadar devam eder. Tüm gözlemler gruplara yerleştirildikten sonra bulundukları küme ortalamasından daha yakın küme ortalaması varsa, gözlemlerin yerleri değiştirilerek daha güvenilir kümeler oluşturulmalıdır.

Genel olarak k-means yönteminin amacı diğer kümeleme yöntemlerinde olduğu gibi, gerçekleştirilen kümeleme analizi işlemi sonucunda elde edilen kümelerin, küme içi benzerliklerinin maksimum, kümeler arası benzerliklerinin ise minimum olmasını sağlamaktır. Küme benzerliği kavramı, kümenin ağırlık merkezi kabul edilen bir birim ile kümedeki diğer birimler arasındaki uzaklıkların ortalama değeri ile ölçülmektedir [34].

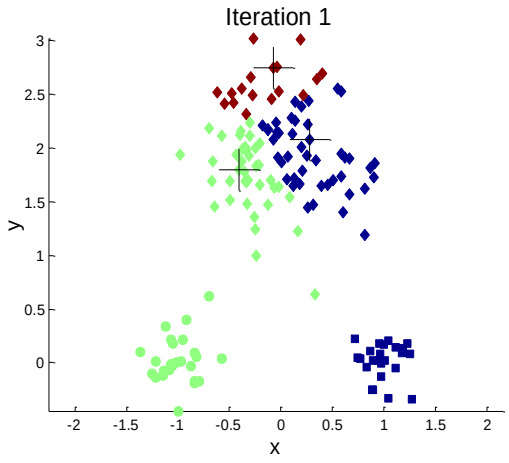
k-ortalamalar yöntemi için kısaca şunları söyleyebiliriz.

- k-means kümeleme yöntemi N adet veri nesnesinden bir veri kümesini giriş parametresi olarak verilen K adet kümeye bölümlemektedir.
- Amaç, gerçekleştirilen kümeleme işlemi sonunda elde edilen kümelerin küme içi benzerliklerin maksimum ve kümeler arası benzerliklerin ise minimum olmasını sağlamaktır.
- k-means kümeleme algoritması uygulanışı ve kullanışı kolay olduğundan en sık tercih edilen kümeleme yöntemidir.
- Büyük ölçekde verileri hızlı ve etkin şekilde kümeleyebilir.
- k-means algoritması, her verinin ait olduğu kümeye olan uzaklıkları toplamını küçültmektedir.
- k-means ile küme içi benzerlik büyük, kümeler arası benzerlik ise küçük olduğu sürece kümelemenin doğru olduğundan bahsedilebilir.

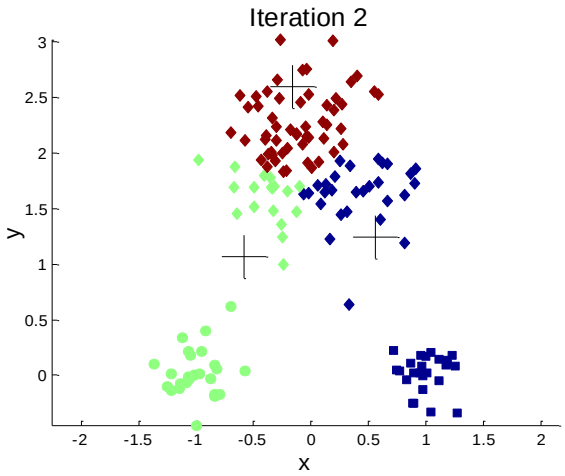
k-means kümeleme analizi için kullanılan algoritma mantığı ise aşağıdaki maddelerde kısa olarak verilmiştir.

1. Sınıf merkezlerinin belirlenmesi,
2. Örneklerin mesafelere göre sınıflandırılması,
3. Yapılan sınıflandırma sonrasında yeni merkezlerin belirlenmesi,
4. Kümeleme işlemi güvenilir hale gelinceye kadar bir önceki iki adımın tekrarlanması olarak sıralayabiliriz.

k-means kümeleme algoritmasında en önemli adımlardan biri küme başlangıç merkezlerinin seçimidir. Bu algoritma için başlangıç noktası seçimi aşağıdaki şekillerle ve sıralı iterasyonlarla gösterilmiştir.

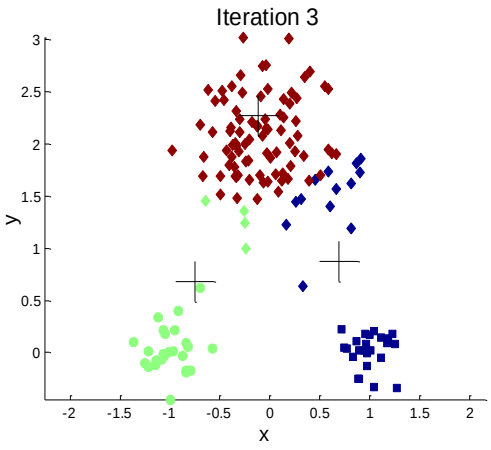


Şekil 3.9 Başlangıç noktası seçimi(1)

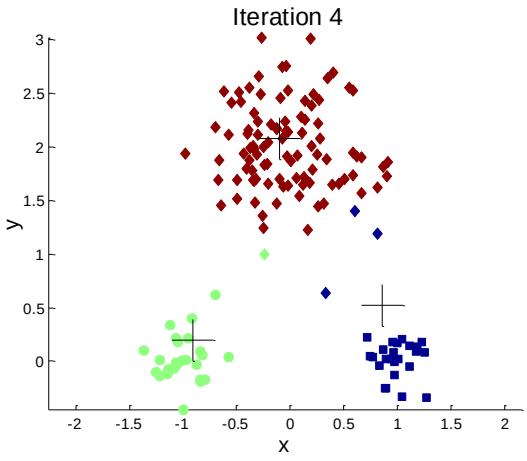


Şekil 3.10 Başlangıç noktası seçimi(2)

k-means algoritması için başlangıç noktası seçimi toplamda altı iterasyon şeklinde gösterilmiştir. Başlangıç seçimi için sıralı iterasyonların diğer adımları Şekil 3.10-11-12 ve 13'de verilmiştir.

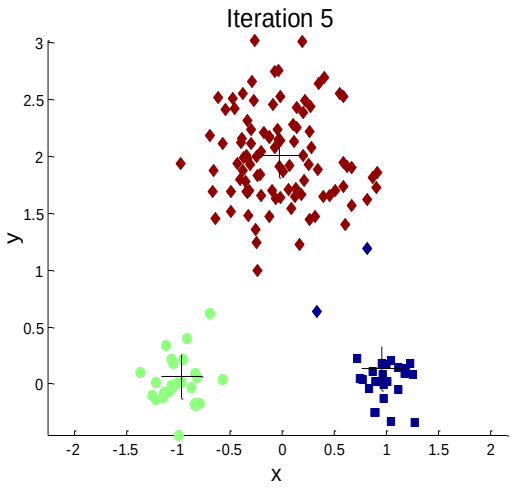


Şekil 3.11 Başlangıç noktası seçimi(3)

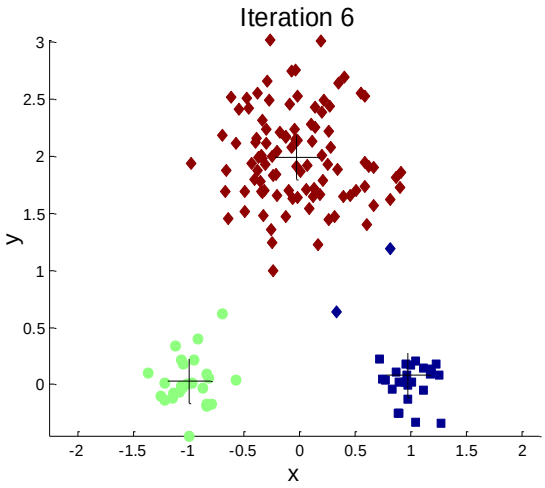


Şekil 3.12 Başlangıç noktası seçimi(4)

Başlangıç noktası seçimi k-means algoritması işleyişine göre giderek daha hızlı ve güvenilir sonuçlar verebilmesi bu yöntemin en sık kullanılan kümeleme yöntemlerinden birini olmasını sağlamıştır. Son olarak beşinci ve altıncı iterasyon adımları tamamlanınca küme başlangıç noktaları k-means algoritmasına göre seçilmiş olacaktır.



Şekil 3.13 Başlangıç noktası seçimi(5)



Şekil 3.14 Başlangıç noktası seçimi(6)

İterasyonlarla k-means kümeleme için ideal küme başlangıç noktaları, küme içi homojenlik ve kümeler arası heterojenlik ilkesi dikkate alınarak belirlenmiştir.

Bir önceki iterasyonlarda genel olarak MacQueen [35] tarafından ortaya konulan küme merkezi çalışması özetlenmeye çalışılmıştır.

Kümeleme analizi yöntemleri üzerinde çalışan Anderberg [11] tarafından Forgy adlı araştırmacının gözlemlerin kümelere atanmasında izlediği yol ise aşağıda gösterilmiştir.

1. Sürecin başında eğer çekirdek noktalar bulunmuşsa 2.adıma geçilir. İlk uygun bölümleme gerçekleşmiş ise 3.adıma geçilir.
2. Her birimi en yakın çekirdek noktaya atayarak küme oluşturulur.
3. İlk uygun bölümleme gerçekleşmiş ise çekirdek noktalar, oluşturulan çekirdek kümelerin ortalamaları olarak alınır ve yakınlık ilkesine göre atama yapılır.
4. Küme elemanlarının geçişi duruncaya kadar algoritma sürdürülür.

Anderberg ve MacQueen tarafından ortaya konulan k-means algoritmaları arasındaki önemli farklar aşağıda iki madde halinde verilmiştir.

- MacQueen tarafından ortaya konulan yöntemde her gözlemin kümeye atanmasından sonra küme merkezinin yenilenmesi,
- Anderberg tarafından ortaya konulan algoritmada ise küme merkez noktaları(çekirdekleri) hep sabit kalır.

K-means kümeleme yönteminin değerlendirilmesinde en yaygın olarak karesel hata kriteri SSE kullanılır. En düşük SSE değerine sahip kümeleme sonucu en iyi sonucu verir.

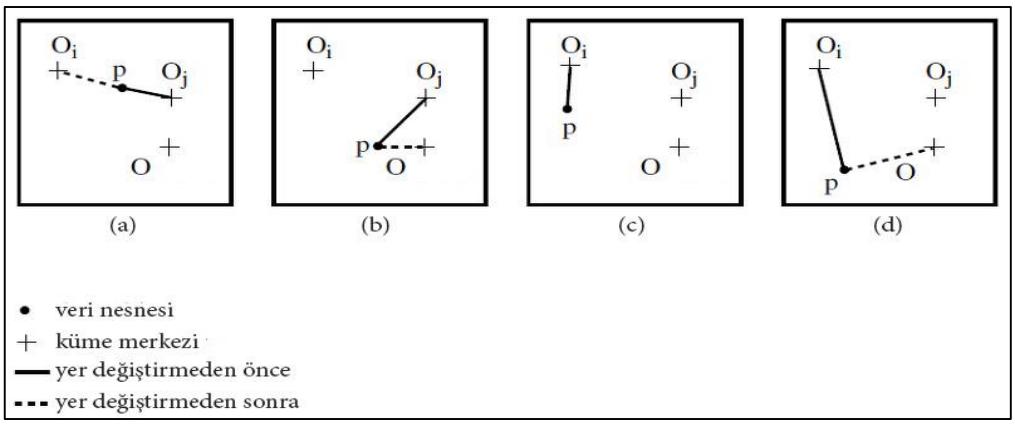
Nesnelerin bulundukları kümenin merkez noktalarına olan uzaklıklarının karelerinin toplamı (1) nolu eşitlik ile hesaplanmaktadır [37].

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_j} dist^2(m_i, x) \tag{3.15}$$

Bu işlemlerin sonucu, k tane kümenin olabildiğince yoğun ve birbirinden ayrı sonuçlanması hedeflenmeye çalışılır. Algoritma, karesel-hata fonksiyonunu azaltacak k parçayı belirlemeye gayret eder. K-means algoritması, algoritmaya kullanıcı tarafından verilen k parametresi ile n tane veriden oluşan veri setini k adet kümeye böler. Küme benzerliği kümedeki nesnelerin ortalama değeri ile ölçülür, bu da kümenin ağırlık merkezidir [36].

3.2.2 K-Medoid Kümeleme

K-medoids algoritmasının temeli verinin çeşitli yapısal özelliklerini temsil eden k tane temsilci nesneyi bulma esasına dayanır. En yaygın kullanılan k-medoids algoritması 1987 yılında Kaufman and Rousseeuw tarafından geliştirilmiştir. Temsilci nesne diğer nesnelere olan ortalama uzaklığı minimum yapan kümenin en merkezi nesnesidir. Bu nedenle, bu bölünme metodu her bir nesne ve onun referans noktası arasındaki benzersizliklerin toplamını küçültme mantığı esas alınarak uygulanır. Kümeleme literatüründe temsilci nesnelere çoğunlukla merkez tipler (centrotypes) denilmektedir (Işık, 2006). K-medoid algoritmasının işleyiş şekli aşağıdaki Şekil 3.14'te verilmiştir.



Şekil 3.15 K-medoid algoritmasının kümeleme adımları

K-medoids algoritmasının birden fazla uygulanma şekli ve faklı algoritmaları bulunmaktadır. PAM (Partitioning Around Medoids) ilk ortaya atılan K-medoids algoritmasıdır. PAM, öncelikle k-means algoritmasında olduğu gibi rastgele seçtiği k adet sayıyı küme merkezi olarak alırve daha sonra kümeye her yeni eleman katıldığında kümenin elemanlarını deneyerek kümenin gelişmesine en fazla katkıda bulunabilecek noktayı tespit edince bulduğu noktayı yeni merkez, eski merkezi ise sıradan küme elemanı olacak şekilde yer değiştirme (swap) işlemi yapar [38].

K-Medoids algoritmasının işlem basamakları aşağıdaki gibidir:

- Adım 1: K küme sayısının belirlenmesi.
- Adım 2: Başlangıç medoidleri olarak k nesnelerinin seçimi.
- Adım 3: En yakın medoid x'e sahip kümeye, kalan nesneleri atamak.
- Adım 4: Amaç fonksiyonunu hesaplamak. (Hata kareler kriteri: en yakın medoidler için bütün nesnelerin uzaklıklarının toplamı).
- Adım 5: Tesadüfî olarak medoid olmayan y noktasının seçimi.
- Adım 6: Eğer x ile y'nin yer değiştirmesi amaç fonksiyonunu minimize edecekse bu iki noktanın (x ile y) yerini değiştirmek.
- Adım 7: Değişiklik olmayana kadar Adım 3 ile Adım 6 arası işlemler tekrarlanır.

3.3 Modele Dayalı Kümeleme Yöntemleri

Gruplayıcı hiyerarşik ve k-means kümeleme metotları büyük ölçüde sezgisel düşünceye bağlıdır ve işleyiş bakımından sık tercih edilen ve güvenilebilir yöntemlerdir. Ama bu yöntemler genellikle biçimsel(formal) olarak model bazlı değildir. Bu da küme sayısına karar vermede, kümeleme yöntemine karar verme gibi çeşitli durumlarda bize zorluklar çıkarabilir. Kümeleme analizinin açıklayıcı bir araç olarak kullanıldığı durumlarda bir formel modele ihtiyaç duyulmaz. Buna karşın model bazlı kümeleme yöntemlerinin belli bazı avantajları vardır.

Model bazlı yaklaşıma en başarılı örnekler Scoot ve Symons(1971) tarafından verilmiş olup daha sonra Banfield ve Raftery(1993) tarafından bu başarılı yaklaşım daha da genişletilmiştir. Bu yaklaşımda ana kütleden farklı bir ana kütleye ve ona bağlı olarak da bu gözlemler *c* alt ana kitlelerden oluşur. Her bir alt ana kitle bir kümeye karşılık gelir ve

ϑ_j , j. alt ana kitleden q-boyutsal gözlemin yoğunluğudur. γ_i ise her gözlemin alt popülasyonunu temsil eder. Burda kümelemede karşılaştığımız zorluk $\theta=(\theta_1, \theta_2,\ldots \theta_c)$ parametresini belirlemek ve γ'nin olabilirlik fonksiyonunu maksimize etmektir. Olabilirlik fonksiyonunun maksimize edilmesi ile ilgili bilgiler aşağıda verilmiştir.

θ' ve γ'nın maksimum olabilirlik fonksiyonunun belirlenmesi Eşitlik 3.16'da verilmiştir.

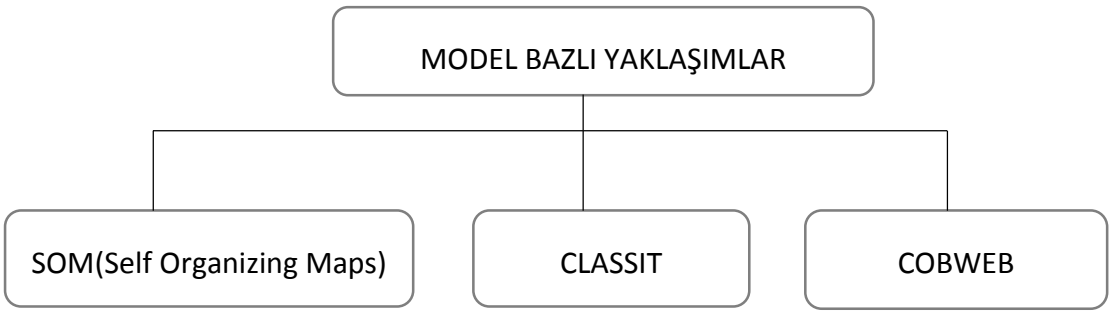
$$L(\theta,\gamma)=\prod_{i=1}^{n} f_{\gamma i}(x_i;\theta_{\gamma i}) \tag{3.16}$$

Eğer $f_j(x;\vartheta_j)$, μ_j ve kovaryans matrisi $\sum_j$ olan çok değişkenli normal yoğunluğa sahipse bu seferde olacak olabilirlik fonksiyonu aşağıda verilmiştir.

$$L(\theta;\gamma)=const\prod_{k=1}^{c}\prod_{i\in E_k}\left|\sum\nolimits_k\right|^{1/2}\exp\left\{-\frac{1}{2}(x_i-\mu_k)'\sum\nolimits_k^{-1}(x_i-\mu_k)\right\} \tag{3.17}$$

Model seçme işleminde uygun model seçimi ve optimal küme seçimi önemli birer adımlardır ve bu seçimde Bayesyen Bilgi Kriterinin(BIC) kullanılması önemli bir yaklaşımdır.

Model bazlı kümelemeye üç farklı istatistiksel algoritmik yaklaşım vardır.

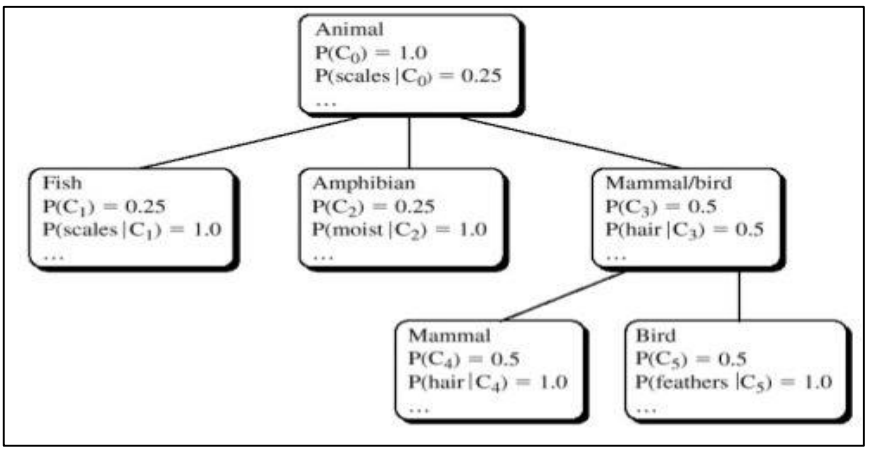


Şekil 3.16 Model bazlı yaklaşımlar

COBWEB Model Bazlı Yaklaşım

- Son zamanlarda popüler olan kavramsal bir öğrenme türüdür.
- Hiyerarşik kümeleme yönteminin formunun sınıflandırma ağaçları(classification tree) türüdür.

- Sınıflandırmada her bir düğüm bu analiz içinde bulunan olasılıksal ifadesini içerir. Karar ağaçlarından en büyük farkı da budur.



Şekil 3.17 Cobweb sınıflandırma ağaç görseli [40]

CLASSIT Model Bazlı Yaklaşım

COBWEB model bazlı yaklaşımın sürekli veri türünde kullanılan uzantısıdır.

SOM(Self Organizing Maps) Model Bazlı Yaklaşım

Kümeleme analizinde kullanılan istatistiksel tekniklere alternatif olarak SOM geliştirilmiştir. Kümelemede "benzer" veri noktaları aynı küme içerisinde yer almaktadır [39]. SOM yöntemi, kümeleme analizinde kullanılan tipik bir yapay zeka tekniğidir. Sinir ağlarına dayalı tekniklerin, istatistiksel modelleme tekniklerine tercih edilmesinin en önemli nedenlerinden biri, verilerin dağılımlarına ilişkin varsayımlar gerektirmemeleridir. SOM yüksek boyutlu verilerin daha düşük boyutlu resmedildiği bir harita sunar ve bu harita da veri madenciliği araştırmacılarına kümeleri görüntüler. SOM yöntemi bazı istatistiksel kümeleme yöntemleri gibi kümelerin ön sayısına, değişkenlerin olasılık dağılımlarına ve değişkenler arasındaki bağımsızlığa ilişkin varsayımlar gerektirmez [41].

3.4 Yoğunluğa Dayalı(Density Based) Kümeleme Teknikleri

Rasgele biçimde kümelerin keşfedilmesi için yoğunluk bazlı yöntemler kullanılmaktadır. Yoğunluk bazlı yöntemler birimlerin veri uzayındaki dağılımlarını inceleyerek kümeleri birimlerin oluşturdukları bölgelerin yoğunluklarına göre oluşturmaktadır.

Model bazlı yöntemlere örnek olarak DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identitfy), DENCLUE (Density Based Clustering) yöntemleri verilebilir.

Yoğunluk tabanlı kümeleme, kümelere, data seti içindeki nesnelerin yoğun olduğu bölgelerin gözlemlemesiyle işe başlar ve bu veri setini düşük yoğunluktaki düşük yoğunluktaki bölgelere bölmeye çalışır. Yoğunluk bazlı kümelemeye göre noktaları sonlu sayıda kümeye bölmek için, yoğunluk, bağlana bilirlik ve sınır çevresi gibi kavramların bilinmesi gerekir. Yoğunluk tabanlı kümeleme algoritmaları kendiliğinden ve rastgele şekilli kümeleri, grupları ortaya çıkarmada başarılıdır [42].

3.4.1 DBSCAN Algoritması

DBSCAN adından da anlaşılacağı üzere yoğunluk temelli bir kümeleme yöntemidir. Yoğunluk temelli yöntemler bölümlemeli yöntemlerin başarı sağlayamadığı düzensiz şekilli kümeleri bulmak için geliştirilmiştir. Bu yöntemlerin küme anlayışı genellikle düşük yoğunluktaki bölgelerden ayrılmış nispeten daha yüksek yoğunluktaki nesneler topluluğudur. Algoritma, kümeler yeterli yoğunlukta olacak şekilde bölgeler oluşturur ve gürültülü veri tabanlarından düzensiz şekilli kümeleri keşfeder. Bir kümeyi yoğunluk bağlantılı noktaların en iyi kümesi olarak tanımlar [43].

Yoğunluk temelli kümelemenin temel düşünceleri bazı yeni tanımlamalar içerir. Bu tanımlamalar:

- Verilen bir nesnenin ε yarıçapı içindeki komşuluğuna nesnenin ε *komşuluğu* denir.
- Eğer bir nesnenin ε–komşuluğu en az, en küçük sayıda, *AsgNok*, nesne içeriyorsa nesneye çekirdek nesne denir.
- Nesneler kümesi D'de eğer p, q'nun ε–komşuluğu içinde ise ve q çekirdek nesne ise p, q nesnesinden doğrudan *yoğunluk ulaşılabilirdir* denir.

- Eğer nesneler kümesi D'de $1 \leq i \leq n$, p_i D için p_i+1, ε ve *AsgNok*'a uygun olarak pi'den doğrudan yoğunluk–ulaşılabilirse ve $p_1, \ldots, p_n$, $p_1 = q$ ve $pn = p$ şeklinde bir nesne zinciri varsa p nesnesine ε ve *AsgNok*'a göre q nesnesinden *yoğunluk ulaşılabilirdir* denir.
- Nesneler kümesi D'de o D ve p ve q, ε ve *AsgNok*'a uyaraktan o'dan *yoğunluk ulaşılabilir* ise p nesnesine ε ve AsgNok'a göre q nesnesine yoğunluk bağlantılıdır denir.
- *Yoğunluk ulaşılabilirlik*, doğrudan yoğunluk ulaşılabilirliğin geçişli kapanmasıdır ve bu ilişki bakışımsızdır (asimetrik). Sadece çekirdek nesneler karşılıklı yoğunluk ulaşılabilirdir. Buna karşın *yoğunluk bağlantılılık* bakışımlıdır (simetrik).

Yoğunluk temelli küme, yoğunluk ulaşılabilirliğin en fazla sayıdaki durumuna göre oluşturulmuş yoğunluk–bağlantılı nesnelerin kümesidir. Herhangi bir küme içinde olmayan her nesne gürültü olarak ele alınır. DBSCAN, kümeleri veri tabanındaki her bir noktanın ε–komşuluğunu kontrol ederek arar. Eğer bir *p* noktasının *ε komşuluğu AsgNok*'a eşitse veya fazla nokta içeriyorsa *p*'yi çekirdek nesne olarak içeren yeni bir küme oluşturulur. DBSCAN bu çekirdek nesnelerden doğrudan yoğunluk ulaşılabilir nesneleri yinelemeli olarak toplar. Bu işlem herhangi bir kümeye eklenebilecek yeni nokta kalmadığında sonlanır [41].

DBSCAN kümeleme türünde kullanılan bazı kavramlar aşağıda verilmiştir.

Bir noktanın eps komşuluğu

Bir p noktasının Eps Komşuluğu, $N_{Eps}(p)$ ile gösterilir ve $N_{Eps}(p) = \{q \in D \mid dist(p,q) \leq Eps\}$ ile tanımlanır. Kümedeki her nokta için, o noktanın Eps Momşuluğundan en az MinPts kadar minimum sayıda nokta olması yaklaşımı gereklidir. Bir *C* kümesindeki her *p* noktası için, *p q*'nun eps Komşuluğu'nda ve $N_{Eps}(q)$ en az MinPts kadar nokta içerecek şekilde bir *C'*de *q* noktası olması önerilir.

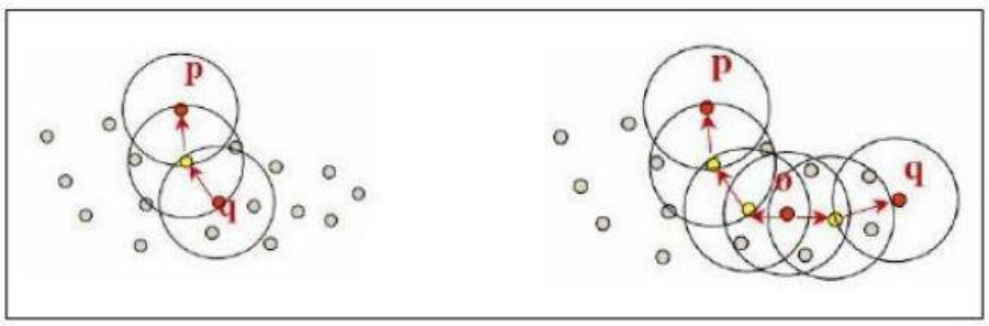
Direkt yoğunluk ulaşılabilirlik

Bir *q* noktasının *p* noktasına direkt yoğunluk ulaşılabilir olması için eps ve MinPts koşulları altında şu özellikleri sağlaması gerekir.

- $p \in NEps(q)$
- $|\ NEps(q)| \geq MinPts$

Yoğunluk bağlılık

Eğer Eps ve MinPts koşulları altında o gibi bir noktadan p ve q'nun ikisi de yoğunluk ulaşılabilir ise p noktası q noktasına yoğunluk bağlıdır. Yoğunluk bağlılık simetrik bir ilişkidir. Yoğunluk ulaşılabilir noktalar için, yoğunluk bağlılık ilişkisi aynı zamanda dönüşlüdür.



Şekil 3.18 Yoğunluk ulaşılabilirlik(a) ve yoğunluk bağlılık(b)

Bir küme bulmak için DBSCAN keyfi bir p noktası ile başlar ve Eps ile MinPts koşulları altında p'den yoğunluk ulaşılabilir tüm noktaları elde eder. Eğer p bir öz nokta ise, bu işlem Eps ve MinPts ile bir küme üretir. Eğer p bir sınır nesnesi ise p'den yoğunluk ulaşılabilir herhangi bir nokta yoktur ve DBSCAN veri tabanının bir sonraki noktasına gider. Algoritma,

- Eps ve MinPts parametreleri kullanıcı tarafından verilir.
- Veri setindeki tüm noktalar sınıflandırılmamış olarak işaretlenir.
- Eps ve MinPts koşulları altında sınıflandırılmamış bir p öz noktası bulunur. Bu p noktası sınıflandırılmış olarak işaretlenir. Mevcut küme olarak yeni bir küme yaratılır ve p bu mevcut kümeye atanır.
- p'nin Eps_Komşuluğundaki tüm sınıflandırılmamış noktalar bulunur. Geçici bir küme yaratılır ve tüm bu noktalar o kümeye konulur.
- Geçici kümedeki q noktası alınır, sınıflandırılmış olarak işaretlenir, mevcut kümeye atanır ve geçici kümeden kaldırılır.

- Eps ve MinPts parametreleri ile q‘nun öz nokta olup olmadığına bakılır. Eğer öz nokta ise, q‘nun Eps_Komşuluğundaki sınıflandırılmamış tüm noktalar geçici kümeye eklenir. 7. 5. ve 6. adımlar geçici küme boş olana kadar tekrar edilir.
- Yeni küme yaratılır ve 3 ile 7 arasındaki adımlar başka öz nokta kalmayıncaya kadar sürdürülür.
- Bulunan tüm kümeler işaretlenir, hiçbir kümeye ait olmayan nesneler sapan olarak atanır.

3.4.2 OPTICS Algoritması

Birçok gerçek veri setinin önemli bir özelliği özel küme yapılarının küresel yoğunluk parametreleri ile karakterize edilememesidir. Çok farklı bölgesel yoğunluklara, veri uzayının farklı bölgelerindeki kümeleri ortaya çıkarmak için ihtiyaç duyulabilir. Bu gibi bir küme yapısını saptamak ve analiz etmek için ilk alternatif tek bağ metodu gibi bir hiyerarşik kümeleme algoritması kullanmaktır. Ancak bu alternatifin iki dezavantajı vardır. İlki tek bağ etkisidir. Örneğin nesneler arası küçük bir uzaklık olan birkaç nokta sırası bakımından bağlı kümeler ayrılamazlar. İkincisi hiyerarşik algoritmalar tarafından üretilen sonuçların yani dendogramların anlaşılması veya analizi birkaç yüz nesneden sonra zordur.

Bu sorunların çözümlenmesi için, veri setinin her küme seviyesine ilişkin bilgiyi içeren yoğunluk tabanlı küme yapısı hakkında veri tabanının özel bir sıralamasını üreten OPTICS algoritması kullanılmalıdır [44].

3.5 Izgara Tabanlı(Grid Based) Kümeleme Teknikleri

Grid bazlı kümeleme yöntemleri ise veri uzayının gridleri kullanılarak sonlu sayıda hücrelere bölen yöntemlerdir. Bu yöntemleri kullanmanın kazancı, hızlı işlem süresidir. İşlem süresi birimlerin sayısından bağımsızdır. Fakat kümelenecek olan veri uzayında oluşturulacak her bir boyuttaki hücrelerin sayısına bağımlıdır. Bu yöntemlere örnek olarak STING (Statistical Information Grid), Wave Cluster, CLIQUE (Clustering Hıgh Dimensional Space) yöntemleri verilebilir [45].

Grid bazlı metot, kümeleme yapılacak alanın sonlu sayıda hücrelere bölünmesiyle oluşur. Bu metodun ana avantajı, genelde birbirinden bağımsız sayıda veri nesnelerinde hızlı işlem zamanıdır.

3.5.1 STING(İstatistiksel Bilgi Gridi) Algoritması

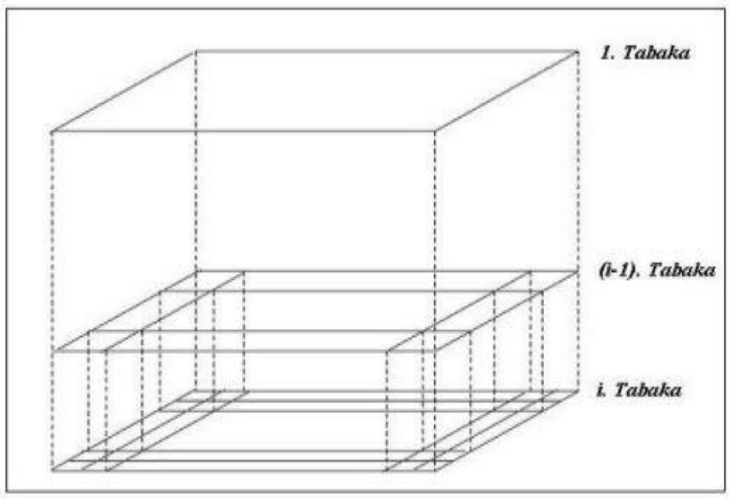
STING, grid hücrelerindeki istatistiksel bilgiyi araştırır. STING, hiyerarşik yapıyı kullanarak uzayı dikdörtgensel hücrelere bölen bir tekniktir.

STING algoritması, hücrelerin içindeki bireylerin sayısal özelliklerinin her birinin ortalama, standart sapma, varyans, minimum, maksimum gibi istatistiksel parametrelerini hesaplar. Sonra farklı seviyelerdeki kümeleme bilgisini temsil ettiği gibi, ızgara hücrelerinin hiyerarşik yapısını oluşturur.

STING, yeni bir bireyin kümelere verimli tahsisini veya sorgulama için kümeleme bilgisinin kullanımını kolaylaştırır [46].

STING'in bazı avantajları;

- Veriler hakkındaki özet bilgiyi temsil eden her bir hücrede depolanan istatistiksel bilgiden dolayı, hesaplanması sorgulamadan bağımsızdır.
- Grid yapısı, paralel işleme ve güncelleştirmelere uygundur.
- Metodun verimi asıl avantajıdır. STING, hücrelerdeki istatistiksel parametreleri hesaplamak için veri tabanına bir kere gider, kümeleri oluşturma zaman karmaşıklığı O(n)'dir, n nesnelerin toplam sayısıdır. Hiyerarşik yapıyı oluşturduktan sonra, sorgu işleme zamanı O(g), g en alt seviyedeki hücre sayısıdır (genellikle n > g olur).



Şekil 3.19 STING kümeleme için hiyerarşik yapı

STING'in kalitesi grid yapısındaki en alt seviye taneciğe bağlıdır. Tanecikler hassas ise işlem maliyeti artar, bununla beraber, en alt seviyedeki grid yapısının kalınlığı kümeleme analiz kalitesini azaltabilir. STING, ana hücrenin çocuk ve komşuları ile ilişkilerini göz önünde bulundurmaz. Kümeleme sınırları düşey veya yataydır, diyagonal sınır yoktur. Bu kalite ve doğruluğu azaltır [47].

3.5.2 Wavecluster Kümeleme

Wavecluster, uzaysal veriyi frekans alanına dönüştürmek için ve dönüştürülen uzaydaki yoğun olan bölgeleri bulmak için işaret işleme tekniklerini kullanır. İlk olarak veri uzayını çok boyutlu grid yapısına dönüştürür. Sonra wavelet dönüşümü kullanarak yoğun bölgeleri bulur ve uzayda dönüşüm yapar. Wavelet model, n defa dönüşüm yaparak bir boyutlu sinyali n boyutlu işaretlere dönüştürebilir. Farklı çözüm seviyelerindeki göreceli mesafe verileri wavelet dönüşümü sayesinde saklanır. Bu doğal kümeleri daha çok ayırt edilebilir hale getirir. İlgi alanındaki yoğun bölgeler arayarak kümeler tanımlanabilir. Avantajları;

- Bu metot danışmansız kümeleme sağlar ve noktaların yoğun olduğu bölgeleri ortaya çıkarmak için çizgi – biçimli (line – shaped) filtre kullanır. Bu özelliği sayesinde aynı zamanda küme sınırları aynı zamanda dışındaki zayıf bilgilerin ortadan kalkmasını sağlar.

- Çoklu – ayrıştırma (multi – resolution) özelliği, farklı doğruluk seviyelerindeki kümeleri ortaya çıkarmada yardımcı olur.
- n veri tabanındaki bireylerin sayısı olmak üzere, zaman karmaşıklığı O(n)'dir.
- Keyfi şekilli kümeleri ortaya çıkarmada ve büyük veri setlerini işlemede başarılıdır.
- Aykırı gözlemler olması durumunda da başarılı sonuçlar verir.
- Küme sayısı ya da komşuluk yarıçapı gibi girdi parametrelerine gerek duymaz. Wavelet tabanlı kümeleme çok hızlıdır. Algoritmanın paralel uygulaması mümkündür.

3.5.3 CLIQUE Kümeleme Yöntemi

CLIQUE, yoğunluk tabanlı ve grid tabanlı kümelerin birleşmesinden oluşan bir algoritmadır. Büyük veri tabanlarında yüksek boyutlu veri kümelemek için yararlıdır. Verilen büyük kümeli çok boyutlu veri noktalarında, veri alanı genelde tek biçimli olmaz. CLIQUE kümeleme, dağınık veri kümelerinde seyrek ve kalabalık bölgeleri tanımlar. Toplam veri noktalarının parçası giriş model parametresini aşarsa bölüm yoğundur. CLIQUE'da, küme bağlantılı yoğun bölümün maksimal bir kümesi olarak tanımlanır.

3.6 Küme Sonuçlarının Değerlendirilmesi

Kümeleme analizlerinde, analiz işlemleri yapıldıktan sonra sonuç kümelemelerinin değerlendirilmesi ve buna bağlı sonuçların çıkarımı kümeleme analiz modeli işleminin ayrılmaz bir parçasıdır. Diğer yandan bir veri grubu kümesinde küme yapısı olmasa bile kümeleme algoritmaları bu veri seti içerisinde istenilen sayıda küme bulacaktır ve bu bizim istediğimiz bir sonuç değildir. Ancak elimizdeki veri kümesinde herhangi bir küme yapısı bulunmayabilir. Bundan dolayı kümeleme yöntemlerinde algoritmalarının değerlendirilmesine yönelik çeşitli küme doğruluk ve geçerlilik (cluster validity) yöntemleri geliştirilmiştir ve bu yöntemlerden yararlanılmaktadır. Bu sayede kümeleme analizlerinde küme doğruluğu, geçerliliği ve uygun küme sayısı belirlenerek kümeleme işlemlerinin çok daha uygun, kaliteli sonuçlar vermesi sağlanabilir [48].

Küme sonuçlarının değerlendirilmesi için literatürde yaygın olarak kullanılan endeksler Şekil 3.19'da verilmiştir. Bu endeksler kullanılırken küme merkezlerine(center of cluster) ve birimlerin ait oldukları küme indislerine ihtiyaç duyulmaktadır.



Şekil 3.20 Küme Sonuçlarının Değerlendirilmesi

3.6.1 Silhouette Endeksi

Kümelenecek her bir birimin gönderildiği kümeye uygunluğunu göstermek amacıyla [49] Silhouette indeksini önermiştir.

a(i); i. birimin kendi kümesindeki tüm noktalara olan ortalama uzaklıklarını,
b(i); i. birimin diğer kümelerdeki tüm noktalara olan ortalama uzaklıklarını gösterir.

Minimum değerini göstermek üzere, Silhouette indeksi;

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \tag{3.17}$$

Gözlemlerin tamamı için ise Silhouette endeksinin ortalaması aşağıdaki gibi bulunur.

$$\overline{S} = \frac{1}{n}\sum_{i=1}^{n} S_i \tag{3.18}$$

Silhouette endeksi -1 ile 1 arasında bulunan değerler alabilir. Negatif değer alması istenilen bir durum değildir. Bu durum bu $a_i > b_i$ olduğu zamandır ve bu nedenle de küme içindeki noktalara olan ortalama uzaklık farklı bir kümedeki noktaların minimum ortalama uzaklıklarından daha büyük olur. Silhouette endeksinin 1' e yakın pozitif büyük değerler alması, *i.* gözlemin doğal kümesine diğer kümelere oranla daha yakın kümelendiğini, -1' e yakın küçük değerler alması i. gözleminin iyi kümelenmediğini yani yanlış kümelendiğini gösterir. Silhouette endeksi, a_i'nin "0" olması durumunda maksimum 1 değerini alabilir [49].

3.6.2 Davies-Bouldin Endeksi

Tüm gözlemler n adet kümeye bölündükten sonra her bir küme için hesaplanan Davies-Bouldin endeksi formülleri aşağıdaki gibidir.

$$DB_{ij} = \frac{\left\{sc(C_i) + sc(C_j)\right\}}{cd(C_i, C_j)} \tag{3.19}$$

$$DB_i = \max_{j=1\ldots N, i \neq j} (DB_{ij}) \tag{3.20}$$

C_i kümesinin içindeki birimlerin küme merkezine olan ortalama uzaklığını, *cd*(*x*, *y*) iki kümenin merkezlerinin birbirlerine olan uzaklığını ifade etmektedir. Tüm veriye ait indeks işe Eşitlik 3.21'de verilmiştir.

$$DB = \frac{1}{n}\sum_{i=1}^{n} DB_i \tag{3.21}$$

Davies-Bouldin indeksi için herhangi bir aralık verilememektedir. Ancak Davies Bouldin indeksi küme kalitesiyle ters orantılıdır. Davies-Bouldin endeksine göre en uygun küme sayısı D(k) değerini minimum yapan k küme sayısıdır [50].

3.6.3 Dunn Endeksi

Veri n adet kümeye bölündükten sonra endeks Eşitlik 3.22'deki gibi hesaplanır.

$$D_n = \min\left\{\min\left(\frac{d(C_i, C_j)}{\max_{k=1,\ldots n} diam(C_k)}\right)\right\} \tag{3.22}$$

Dunn endeksi [0,+∞) açık aralığında değerler alabilir. Kümeleme işlemleri sonucu veri seti iyi ayrılmış kümelere ayrılmışsa kümeler arası uzaklıkların büyük ve küme içi diameterin küçük olması beklenir. Dolayısıyla Dunn endeksi büyük pozitif değerler alır. Sonuç olarak Dunn endeksinin mümkün olduğunca büyük pozitif değerler alması istenir [48].

3.6.4 Calinski ve Harabasz Endeksi

Veri n adet kümeye bölündükten sonra endeks Eşitlik 3.23'deki gibi hesaplanır.

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)} \tag{3.23}$$

W(k) küme içi uzaklıkların kareleri toplamını ifade eder.
B(k) ise kümeler arası uzaklıkların kareleri toplamıdır.

W(k)	*B(k)*
$\frac{1}{2}\sum_{j=1}^{k}\sum_{i,j\in C_i} d(i,j)$ (3.24)	$\frac{1}{2}\sum_{j=1}^{k}\sum_{i,j\in C_i} d(i,j)$ (3.25)

3.6.5 Wilk's Lambda İstatistiği

Yapılan kümeleme analizi çalışmasında analize giren birim sayısı 30'un üzerinde olduğu zman, küme sayısının belirlenmesinde kullanılan ölçütlerden Wilk's Lamda ölçütüne baktığımızda bu ölçütün duyarlılığı diğer ölçütlere göre en yüksek düzeye ulaşmaktadır . Gruplar arası varyansının grup içine göre maksimum olduğunun analiz edilmesinde Wilk's Lambda ölçütü çok duyarlı olmaktadır [7].

İki veya daha çok değişkenli verilerin aynı ortalama vektörlü çok değişkenli normal dağılımdan gelip gelmediği, Çok Değişkenli Varyans Analizi (MANOVA) ile test edilebilmektedir [51].

Kümeleme analizi neticesinde oluşturulan kümelerin ortalamalarının birbirinden istatistiksel olarak farklılığı MANOVA analiziyle test edilebilir ve hesaplanan Wilk's Lamda ölçütü küme sayısının belirlenmesinde kullanılabilir [14].

$$\triangle = \frac{|W|}{|W+B|} = \frac{|W|}{|T|} \tag{3.26}$$

T: çarpımlar ve kareler toplam matrisi,

B: gruplar arası çarpımlar ve kareler toplam matrisi,

W: grup içi çarpımlar ve kareler toplam matrisini göstermektedir.

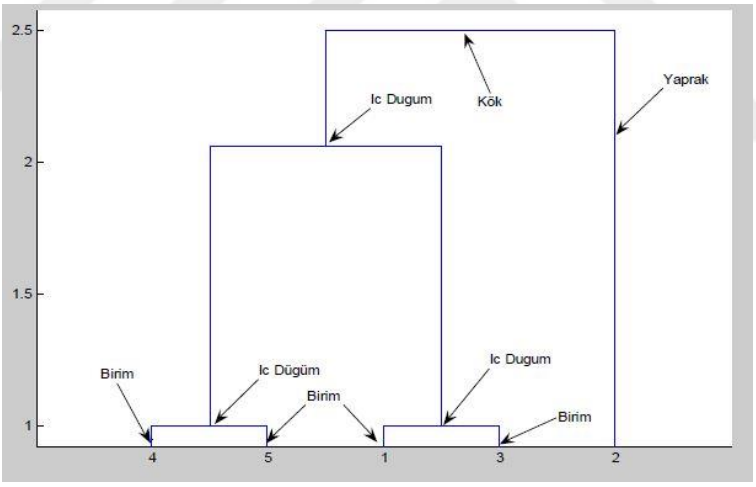
Lamda ölçütü sıfır ile bir arasında değer alan bir ölçüttür. Bu ölçütün değeri tam olarak sıfır değeri almamakla beraber, küme sayısının artırılması Wilk's Lamda değerini sıfıra yaklaşmaktadır.

3.7 Kümeleme Analizi Sonuçlarının Görselleştirilmesi

3.7.1 Dendogram

Kümeleme analizinde hiyerarşik yöntemlerde kullanılan en yaygın görselleştirme tekniğidir.

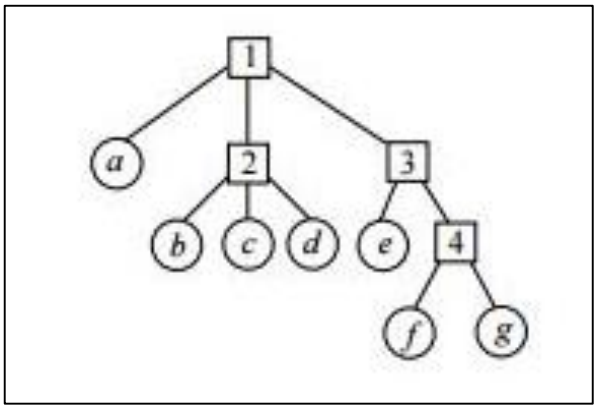
Dendogram kelimesi Yunancadan *dendro* "ağaç" ve *gramma* "çizimi" şeklinde ortaya çıkmıştır.



Şekil 3.21 Dendogram görseli

3.7.2 Ağaç Haritaları (Treemaps)

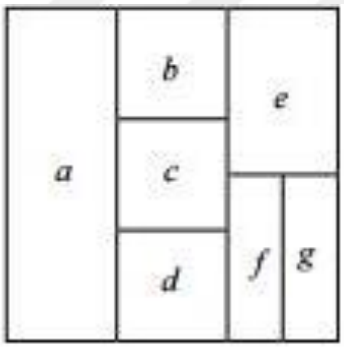
Analiz sonuçlarını görselleştirmede kullanılan en yaygın araçlardan biridir.



Şekil 3.22 Treemaps görseli [52]

3.7.3 Rectangle Grafigi (Dikdörtgen plot)

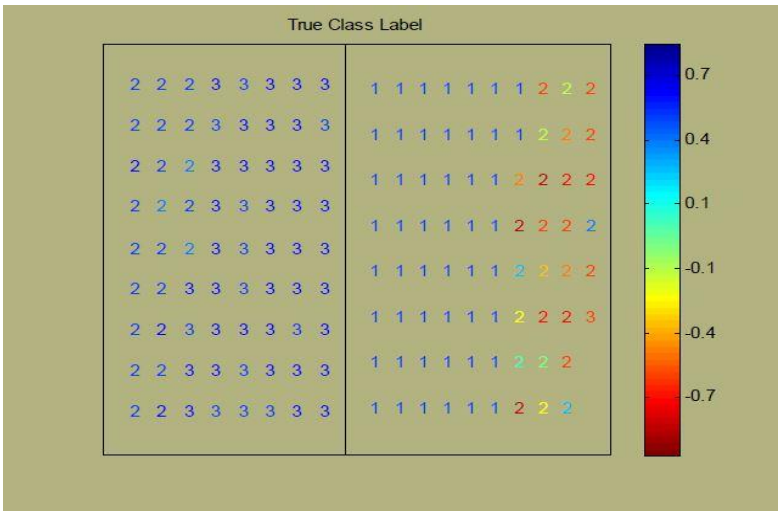
Hayat ağaçları grafiğinden yola çıkarak dendogramlara alternatif bir görsellik aracı olarak kullanılır.



Şekil 3.23 Rectangle plot görseli

3.7.4 ReClus Grafigi (ReClus Plots)

Lifetree ve rectangle plot görselinin en önemli dezavantajlarından biri sadece aglomeratif kümelemede kullanılmasıdır. Reclus grafiği diğer yöntemlerin sonuçlarınında görselleştirmesinde kullanılan bir araçtır [53].

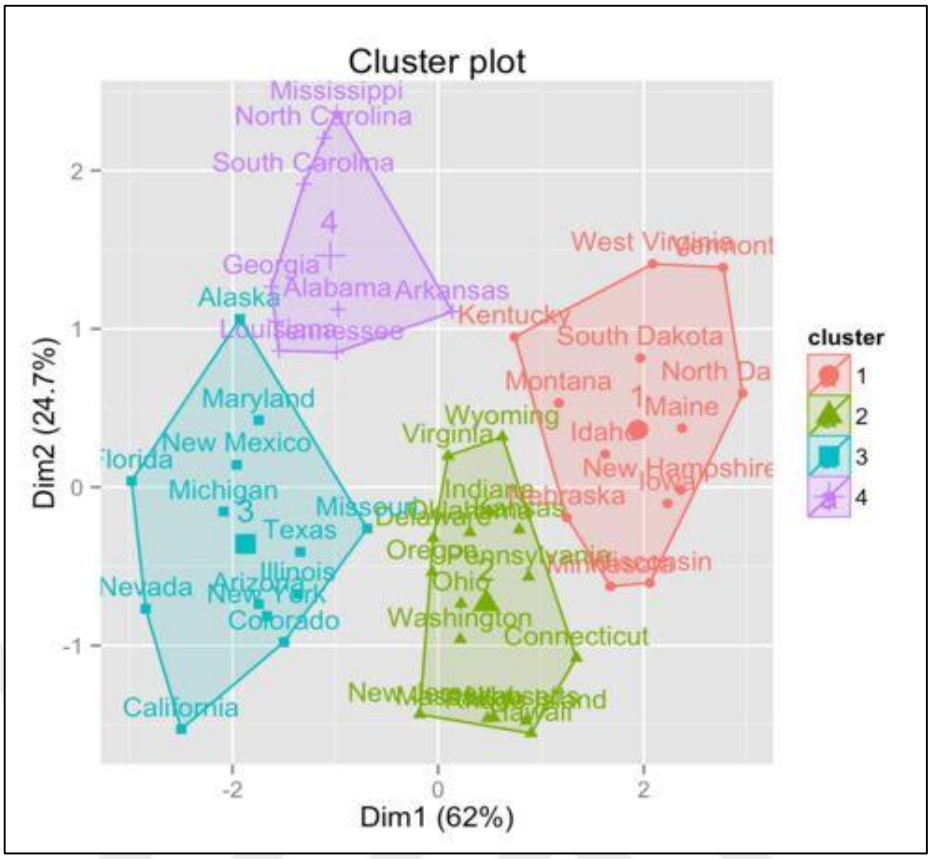


Şekil 3.24 Reclus grafiği görseli [53]

3.8 Çok Değişkenli Yöntemler Kullanılarak Kümelerin Grafiksel Olarak Tespiti

3.8.1 Temel Bileşenler Analizi

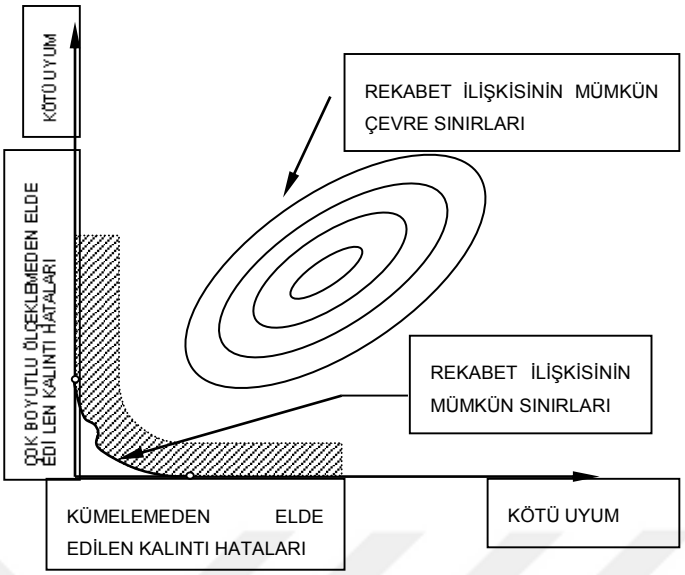
Çok değişkenli yöntemlerde incelenen durumlar ele alındığında, değişkenler arasındaki yüksek korelasyonlar olduğu sistemlerde bu durumun değişkenlerin bağımsızlık varsayımını bozduğu bilinmektedir. Çok sayıda değişkenin analiz sonuçlarının yorumlanmasının zor olduğu durumlarda, değişimin büyük çoğunluğunu açıklayan az sayıda değişkenin olması bu yorumlamayı ve analizi kolaylaştırmaktadır. Bununla beraber son zamanlarda TBA ile birlikte kümeleme analizi yoğun bir şekilde kullanılmaktadır. Bu yolla hem değişken sayısı azaltılmakta hem de çok değişkenli analiz yapısını etkileyecek bağımlılık durumu ortadan kaldırılmaktadır. Temel bileşende oluşan iki boyut üzerinden kümeleme analizinin sonuçlarının doğruluğu ve yorumu hakkında fikir alınılabilir. TBA ve kümeleme arasındaki ilişkiyi gösteren Amerika Birleşik Devletleri eyaletleri üzerinde yapılan çalışmaya ait şekil aşağıda verilmiştir (USArrest versiseti,değişkenler: cinayet,saldırı,nüfus,tecavüz,1973) [54].



Şekil 3.25 Temel Bileşen Analizi ve kümeleme arasındaki ilişki

3.8.2 Çok Boyutlu Ölçekleme

Çok boyutlu ölçekleme, yakınlık matrisleri ve onların görüntülenmesinin sağlanması ile ilgili olduğundan, bu iki karakteristiğe sahip olan kümeleme yöntemleri ile karşılaştırılması uygun olacaktır. Çok boyutlu ölçekleme ile kümeleme analizi arasındaki temel farklılık, çok boyutlu ölçekleme yakınlıkların uzaysal görüntülenmesini sağlarken, kümelemenin yakınlıkların ağaç biçiminde görüntülenmesini sağlamasıdır [55]. Kümeleme analizi her zaman kesin bir ağaç görünümü sağlamasa da, pek çok durumda elde edilen görünüm bir ağacı andırmaktadır. Bu nedenle, kümeleme ve çok boyutlu ölçekleme arasındaki matematiksel ilişki oldukça basit ve açıktır. Bu iki yöntem arasındaki istatistiksel ilişki ise, matematiksel ilişkiden çok daha karmaşıktır. Pek çok araştırmacı, bu iki yöntemin aynı yakınlıklara uygulanmasını kullanışlı bulmaktadır. Şekil 2, çok boyutlu ölçekleme ve kümeleme analizi arasındaki bir ilişkiyi gösteren izlenimci bir diagramdır [55].



Şekil 3.26 ÇBÖ ve kümeleme arasındaki ilişkiyi gösteren diagram

3.8.3 Diskriminant Analizi

Birçok fen ve sosyal bilimi için geniş uygulama sahası bulan diskriminant analizi, bilinen özelliklere göre birimleri sınıflara ayırmaktadır. Fakat bu yönüyle kümeleme analizi ile aynı işlevi gerçekleştirdiği algısı yanıltıcı olacaktır. Diskriminant analizi istatistiksel temellere dayanan güçlü bir karar biçimidir. Kümeleme analizinde değişkenlerin özelliklerine göre sınıflandırma işlemi uygulanırken diskriminant analizinde amaç hatalı sınıflandırma yapılmasını önlemek için birimleri gerçek gruplarına ayırmak ve birimlerin gelmiş olduğu yığınları tespit etmektir [56].

Diskriminant analizi tam olarak birimleri sınıflandırma işlemini en az hata ile gerçekleştirmek için mevcut istatistiksel yöntemlerdir. Bu yöntemde esasında gruplar arasındaki farklılığın ortalamalar cinsinden en yüksek değerlere ulaşmasını hedeflemektedir. Bunun için diskriminant analizi birimleri tasnif edebilecek bir fonksiyona ihtiyaç duymaktadır [56].

BÖLÜM 4

HİDROLOJİ ÜZERİNE UYGULAMALAR

Bu bölümde kümeleme analizinin istatistiksel hidroloji alanında uygulamalarına yer verilecektir. Türkiye'de havzaların akım gözlem istasyonlarınca nasıl kümelendiği görmek amaçlanmaktadır. Oluşturulacak kümelerin sayısının önceden bilinmemesi ve daha özet bilgi vermesi nedeniyle illerin kümelenmesi için, yöntem olarak kümeleme analizi tercih edilmiştir.

Havzalara temel teşkil eden suyun, ne kadar hayati ve toplumsal öneme sahip bir kaynak olduğu artık çok daha yüksek sesle dile getirilen bir gerçektir. Su açısından dünyanın yarı-kurak bir bölgesinde bulunan Türkiye'nin yağış rejimi, mevsimlere ve bölgelere göre büyük farklılıklar göstermektedir Bazı akarsu havzalarında su ihtiyaçlarının, kaynakların potansiyelini aşmış durumda olduğu görülmektedir.



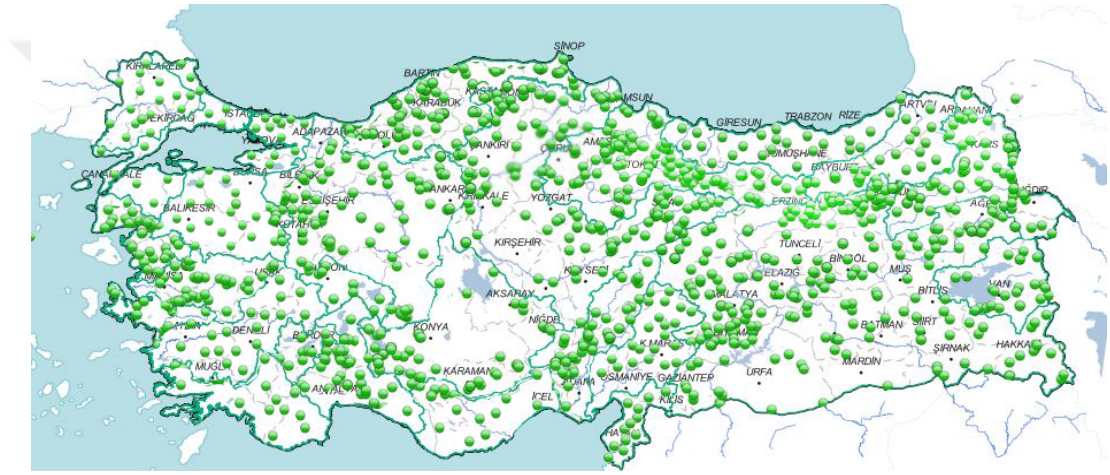
Şekil 4.1 Türkiye Nehir Havzaları Haritası (DSİ, 2012)

Hızlı nüfus artışına bağlı olarak artan su ihtiyacına karşın, uygun kaynak varlığının azlığı ve gün geçtikçe gelişen sanayi ve tarımsal faaliyetlere paralel olarak ortaya çıkan ve aşırı

kullanım, yer altı su rezervlerindeki düşüşler ve kirlilik oluşumu nedenleriyle yaşanan sorunlar, havza bazında su kaynakları yönetiminin önemini bir kat daha arttırmıştır.

Türkiye'de, su kaynakları akarsular bazında hidrolojik açıdan 25 havzaya bölünmüştür ve idaresi DSİ tarafından yapılmaktadır. Tez çalışmasında kullanılan veri kaynakları aşağıda belirtilmiştir [57].

- Devlet Su İşleri(DSİ)
- Meteoroloji Genel Müdürlüğü(MGM)
- Hollanda Ulusal İklim Ajansı(KNMI)
- World Bank İklim Veri Tabanı



Şekil 4.2 Türkiye için kullanılan akım gözlem istasyonları

Bu çalışmada örnek teşkil etmesi amacıyla Türkiye geneli ve Dicle-Fırat Havzası kaliteli ve uzun dönemli veri altyapısı nedeniyle seçilmiştir. Havza üzerinde alternatif kümeleme yöntemleri denenerek, bu havzaya bağlı olan istasyonların nasıl kümelendiği incelenecektir.

Kümeleme Analizi ile İlgili Yapılmış Çalışmalar

Kümeleme analizi, inşaat mühendisliğinde ve özellikle yağış, akış, sıcaklık gibi hidrolojik ve meteorolojik değişkenlerin sınıflandırılmasında ve homojen hidrolojik havzaların belirlenmesinde uygulanmaktadır. Yapılan bu çalışmalarda, geleneksel kümeleme yöntemleri olarak da adlandırılan hiyerarşik olmayan kümeleme (K-Ortalamalar kümeleme yöntemi) ve hiyerarşik kümeleme (Ward kümeleme yöntemi) yöntemlerinin yaygın bir şekilde kullanıldığı görülmektedir. Unal yaptıkları çalışmada, aylık sıcaklık ve

yağış verilerini kullanarak 5 farklı hiyerarşik kümeleme yöntemi ile Türkiye iklim bölgelerinin yeniden tanımlanmasını amaçlamışlardır. Demirel [59], Türkiye akarsu havzalarının sınıflandırılması ve benzer özelliklere sahip bölgelerin oluşturulması için hiyerarşik kümeleme analizi yöntemini kullanmıştır. Bunun için Türkiye genelinde ölçüm yapan 80 adet akım gözlem istasyonundan alınan aylık ortalama, mevsimsel ve yıllık ortalama akım verileri dikkate alınmıştır [57].

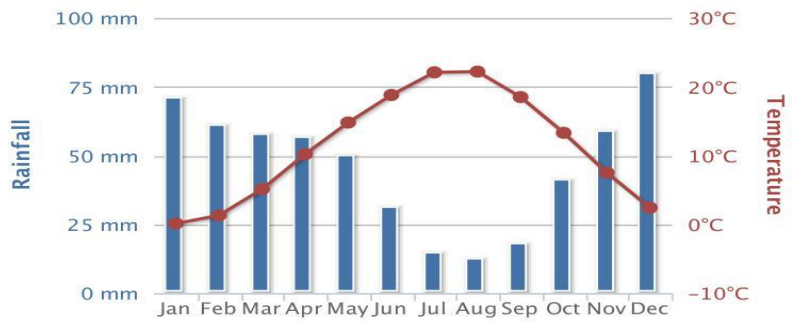
Kahya [58], Türkiye genelinde 23 akarsu havzasında aylık minimum akımların sınıflandırılmasında Principal Component Analysis ve K-Ortalamalar kümeleme yöntemlerinin birlikte kullanılmasını önermiştir. Kahya [58], yıllık ortalama akımları kullanarak K-Ortalamalar kümeleme yöntemi ile Türkiye'de akarsu akımlarının konumsal olarak sınıflandırılmasını amaçlamıştır. Acreman ve Sinclair (1986) tarafından yapılan çalışmada, taşkın frekans analizi için kümeleme analizi yöntemlerini kullanarak havzaları fiziksel karakteristik özelliklerine göre sınıflandırmıştır. Burn (1989) bölgesel taşkın frekans analizi için homojen hidrolojik havzaların belirlenmesinde hiyerarşik olmayan kümeleme yöntemini kullanmıştır. Gutmann (1993) hidrolojik olarak homojen yapıya sahip yağış bölgelerinin tanımlanması için Ward ve L-Momentler yöntemlerini önermiş ve ABD'de 104 adet drenaj havzasına ait verileri kullanarak uygulama yapmıştır.

Kulkarni Ve Krıpalanı (1998) Bulanık kümeleme yöntemi ile Hindistan'da mevsimsel yağış verilerinin konumsal olarak dağılımını incelemişlerdir. Çalışma sonunda, Bulanık kümeleme yönteminin meteorolojik parametrelerin sınıflandırılmasında başarılı bir şekilde uygulanabileceği rapor edilmiştir.

Lauzon, SOM kümeleme yöntemi ile günlük yağış verilerinin sınıflandırılmasını, homojen ve heterojen yağış alanlarının belirlenmesini amaçlamıştır. Elde edilen sonuçlar SOM kümeleme yönteminin yağış alanlarının sınıflandırılmasında uygulanabileceğini göstermiştir. Lın ve Chen (2006), bölgesel frekans analizleri için homojen hidrolojik havzaların belirlenmesinde SOM yöntemini uygulamıştır. Sonuçlara bakıldığında, SOM kümeleme yönteminin küme üyeliklerini geleneksel yöntemlere göre daha doğru bir şekilde belirlediği görülmüştür.

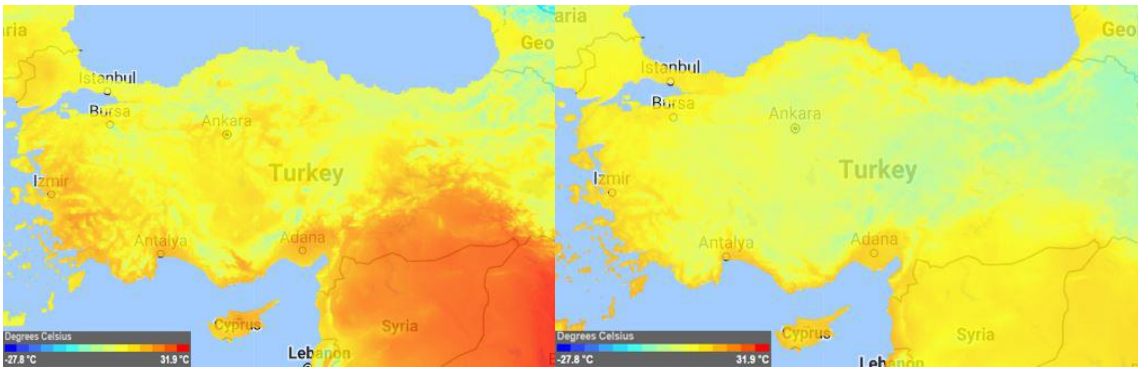
Türkiye ve Fırat Dicle Havzasının Genel Durumu ve Betimleyici İstatistikler

Türkiye'den çeşitli kollar alan Fırat Nehri'nin Türkiye-Suriye sınırındaki yıllık ortalama su miktarı 31,6 milyar m³'tür. Suriye sınırları içinde Habur ve Sacır sularının katılımı ile Fırat Nehri'nin Suriye-Irak sınırındaki yıllık ortalama su miktarı 35 milyar m³'e ulaşmaktadır (Çizelge 1). Irak topraklarından ise hiçbir katkı olmamaktadır. Belirtilen rakamlara göre Türkiye'nin katkısı %90, Suriye'nin ise %10 mertebesindedir. Fırat Nehri akımları gerek yıllar arasında gerekse bir yıl içinde mevsimler olarak büyük değişimler göstermektedir. Örneğin, 1961 yılında sınırda yıllık akım 14,9 milyar m³'e kadar düşmüştür, bu değer uzun yıllar ortalamasının %47'sine tekabül etmektedir. Dicle Nehrin'e Türkiye'nin katkısı 21,3 milyar m³ iken Irak'tan 31.4 milyar m³, diğer bir ifade ile Fırat Nehri'nin taşıdığı su kadar bir katkı gelmektedir. Aşağıdaki tabloda ortalama yıllık akımlar ve kıyıdaş ülkelerin katkıları gösterilmektedir. Fırat-Dicle su kaynaklarının Türkiye'deki geliştirilmesinin belkemiği niteliğindeki Güneydoğu Anadolu Projesi (G.A.P.), sulama ve hidroelektrik enerji üretimi amaçlarına yönelik olarak dünyanın en önemli su kaynaklarını geliştirme planlarından biridir. Bu proje kapsadığı yörenin sosyal ve ekonomik gelişmesinin de başlıca itici gücü niteliğindedir. Fırat su kaynaklarının geliştirilmesi kapsamında, yaklaşık 100 baraj, 70 su kuvveti tesisi Yukarı Fırat'ta, 53 baraj, 18 su kuvveti tesisi Aşağı Fırat'ta, olmak üzere toplam 153 baraj ve 88 su kuvveti tesisi planlanmış olup, bu düzenlemelerle 1.6 milyon ha alan sulanacak ve ortalama 39 TWh/yıl elektrik enerjisi üretimi gerçekleşecektir [57].



Şekil 4.3 Dicle-Fırat havzası için ortalama aylık yağış ve sıcaklık grafiği(1900-2012)

Bu şekil 1900-2012 yılları arasında Dicle-Fırat havzasının yağış ve sıcaklık durumunu vermektedir.(Climatic Research Unit (CRU))



(a) (b)

Şekil 4.4 Türkiye ve havza için maksimum ve minimum sıcaklık değişimleri(NCDC)

Dicle su kaynaklarının geliştirilmesi kapsamında, G.A.P. kapsamındaki Batı ve Orta Dicle dahil, Türkiye'de 37 baraj, 42 su kuvveti tesisi planlanmış olup, bu düzenlemelerle 0,6 milyon ha alan sulanacak ve ortalama 17 TWh/yıl elektrik enerjisi üretimi gerçekleşecektir.

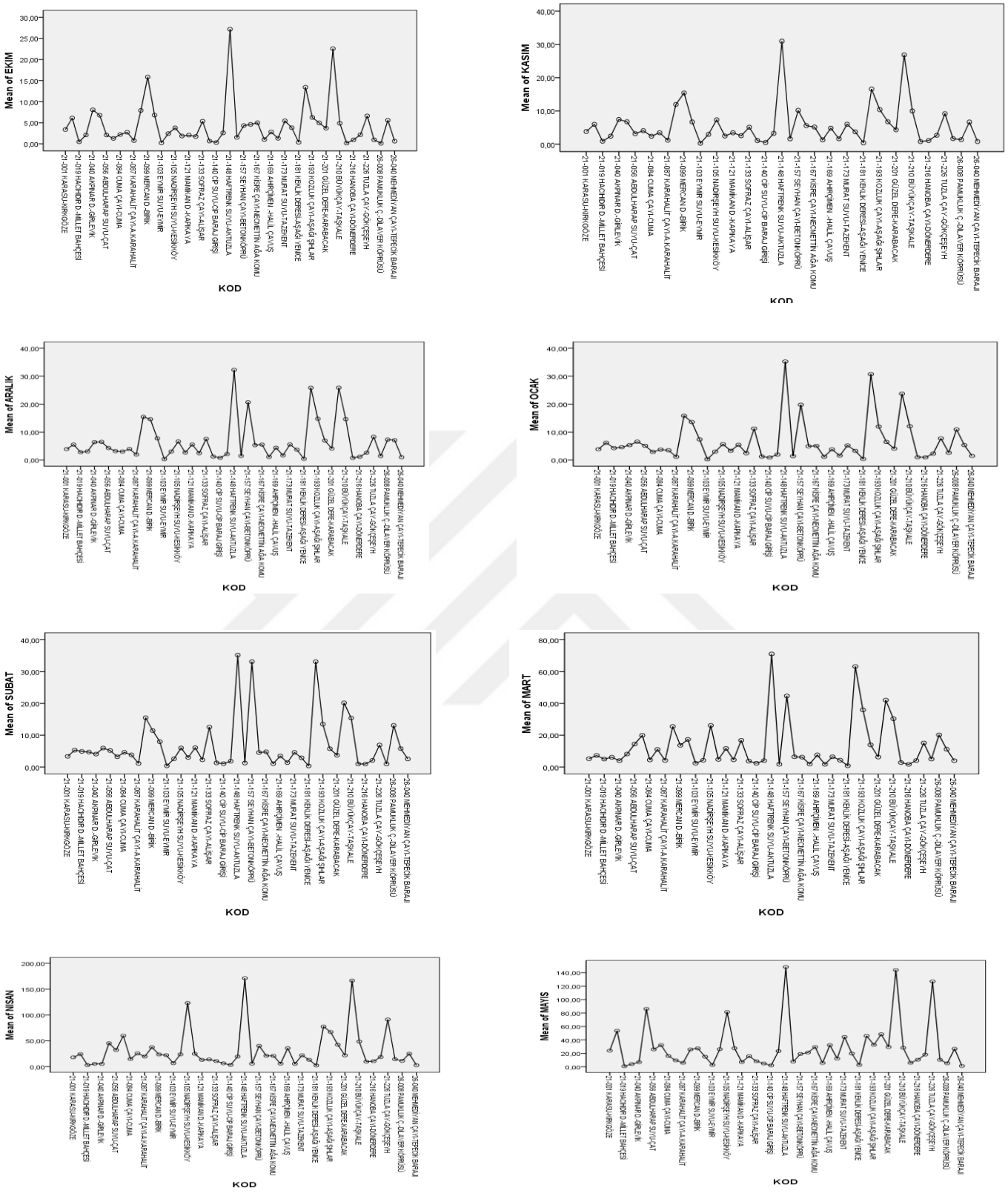
Ölçüm istasyonlarına göre aylık değerlendirmeler

1960-2000 yıllarına ait aylık ortalama akışlar ölçüm istasyonlarına aşağıdaki gibi gerçekleştirilmiştir.

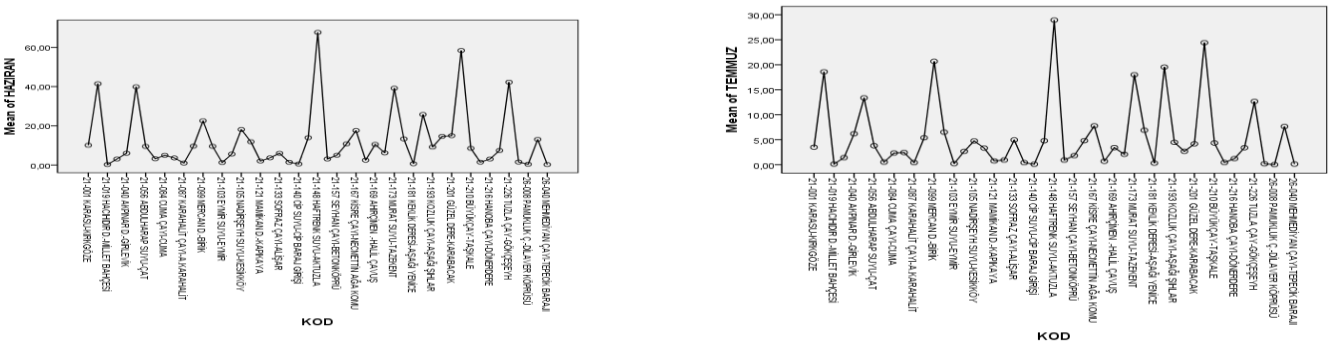
Çizelge 4.1 Varyansların Homojenliği ve Anova

	Levene S.	Sig.
EKIM	13,293	,000
KASIM	10,824	,000
ARALIK	14,044	,000
OCAK	8,801	,000
SUBAT	11,737	,000
MART	13,370	,000
NISAN	18,616	,000
MAYIS	20,506	,000
HAZIRAN	26,560	,000
TEMMUZ	11,620	,000
AĞUSTOS	12,928	,000
EYLÜL	18,126	,000

	Mean S.	F	Sig.
EKIM	554,413	89,147	,000
KASIM	721,267	50,881	,000
ARALIK	918,844	36,786	,000
OCAK	1018,388	36,074	,000
SUBAT	1209,263	46,515	,000
MART	4332,814	32,983	,000
NISAN	26564,740	63,300	,000
MAYIS	22111,819	47,904	,000
HAZIRAN	4562,151	43,227	,000
TEMMUZ	936,836	48,370	,000
AĞUSTOS	489,614	76,368	,000
EYLÜL	393,589	84,726	,000



Şekil 4.5 Aylık Ortalama Grafikleri



Şekil 4.5 Aylık Ortalama Grafikleri(devamı)

İstasyonlar genel olarak değerlendirildiğinde kendi aralarında varyansların homojen olmadığı görülmektedir. Ayrıca benzeştiği varsayımı altında bile istasyonların birbirinden farklı akış rejimleri oldukları söylenebilir. Genel olarak ortalama akış ölçümlerinin istasyonlar bazında birbirlerine yakın olduğu görülmektedir.

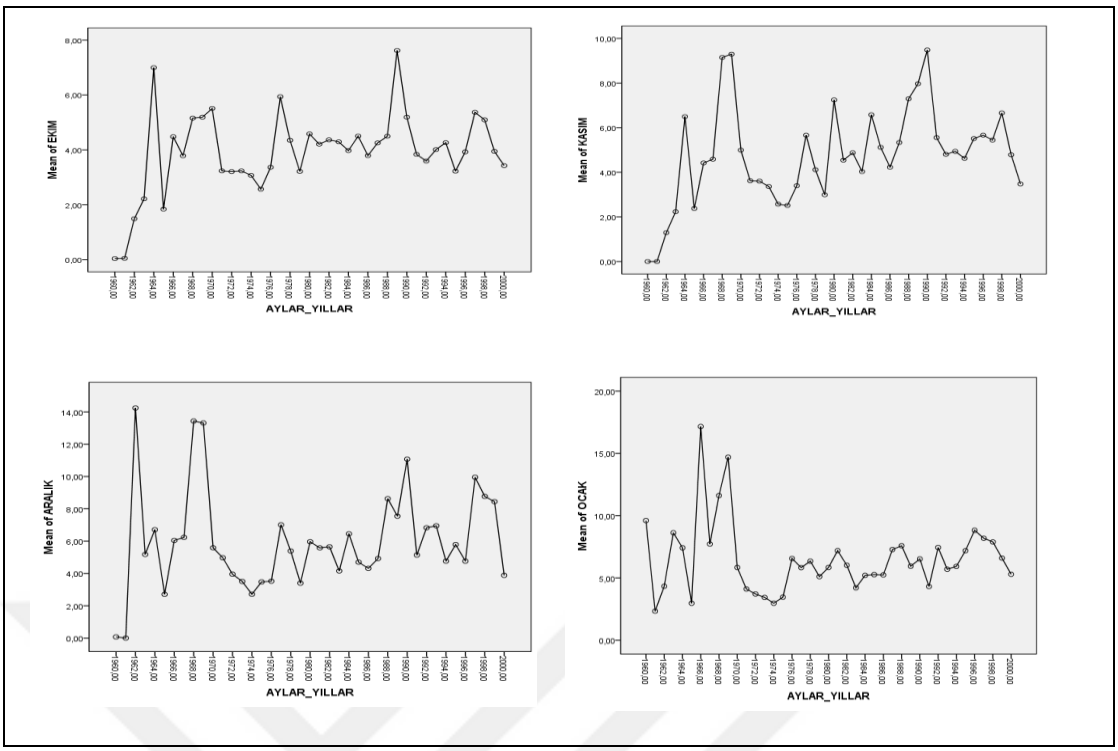
Ölçüm istasyonlarına göre yıllık değerlendirmeler

1961 -2000 yılları için aylık ortalama akım ölçümlerine göre sonuçlar aşağıdaki gibidir.

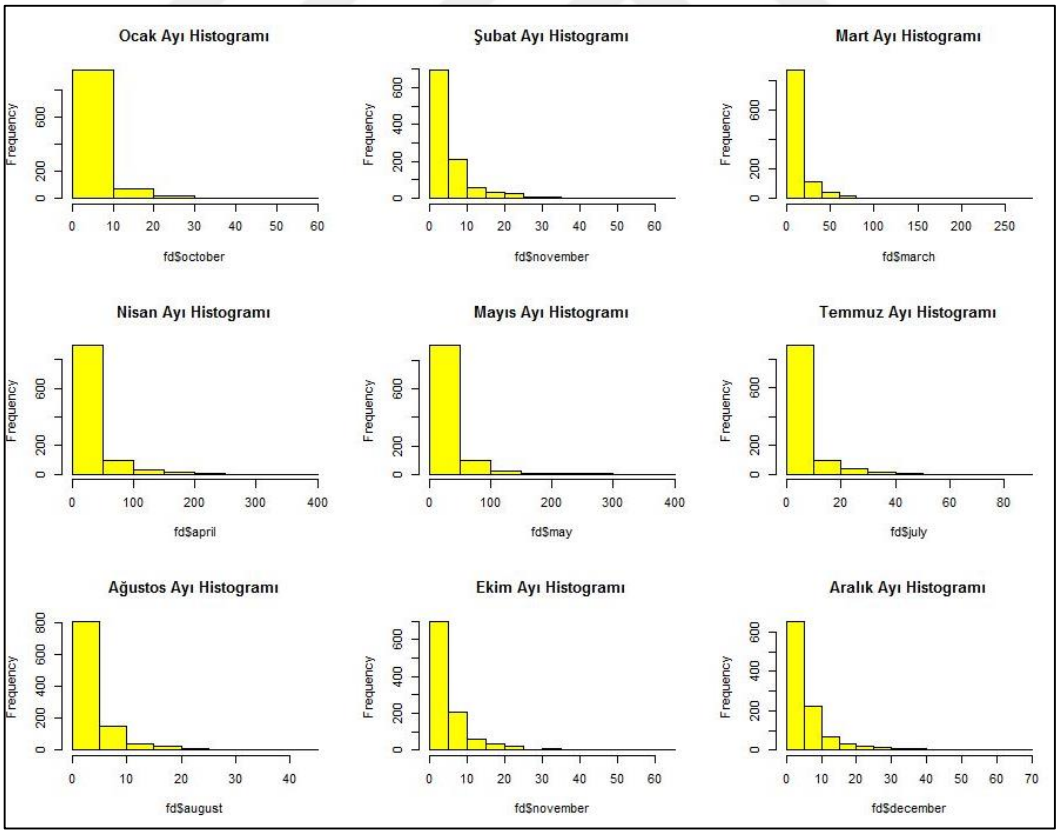
Çizelge 4.2 Varyansların Homojenliği ve Anova

	Levene S.	Sig.
EKIM	1,115	,290
KASIM	2,131	,000
ARALIK	3,535	,000
OCAK	2,179	,000
SUBAT	1,768	,003
MART	2,404	,000
NISAN	2,352	,000
MAYIS	3,277	,000
HAZIRAN	4,515	,000
TEMMUZ	3,632	,000
AĞUSTOS	2,484	,000
EYLÜL	1,766	,003

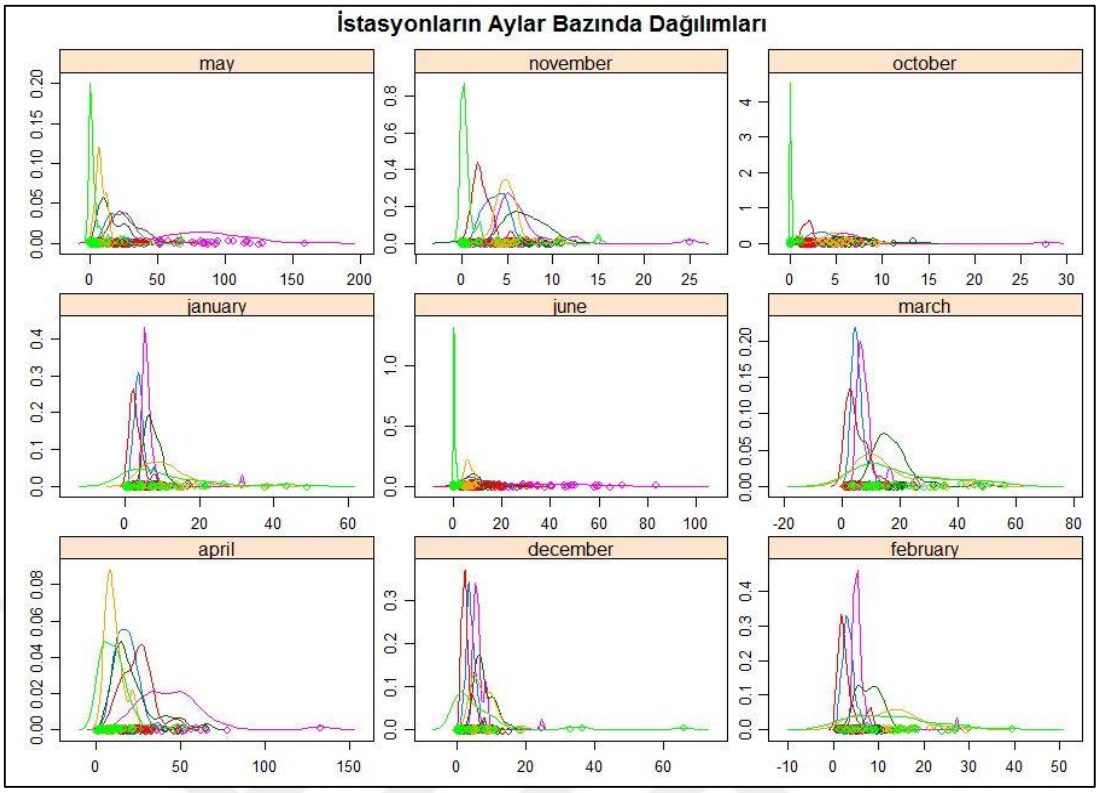
	Mean S.	F	Sig.
EKIM	29,983	,949	,563
KASIM	79,422	1,748	,003
ARALIK	150,852	2,407	,000
OCAK	125,466	1,753	,003
SUBAT	82,393	1,028	,424
MART	444,851	1,394	,055
NISAN	2557,639	1,618	,010
MAYIS	3985,960	2,945	,000
HAZIRAN	901,146	3,149	,000
TEMMUZ	136,461	2,335	,000
AĞUSTOS	38,165	1,354	,072
EYLÜL	21,597	,959	,545



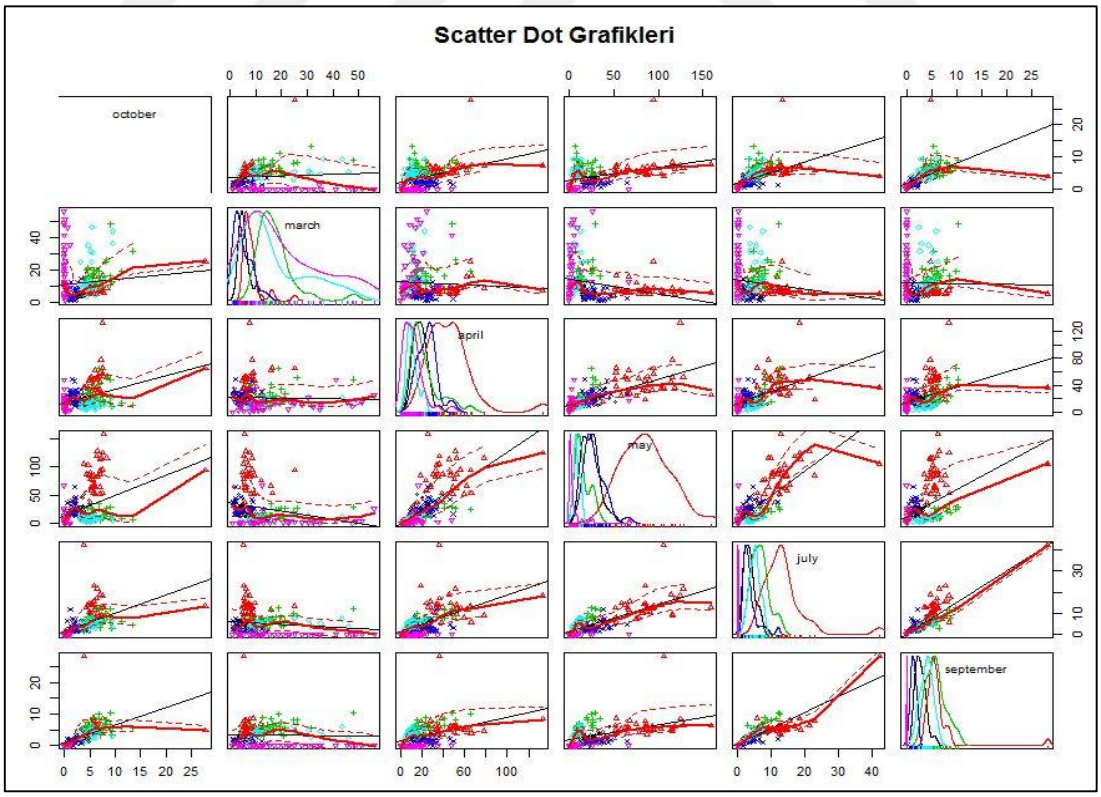
Şekil 4.6 Yıllık Ortalama Grafikleri



Şeki 4.7 Fırat Dicle Havzası için aylık betimleyici histogramlar



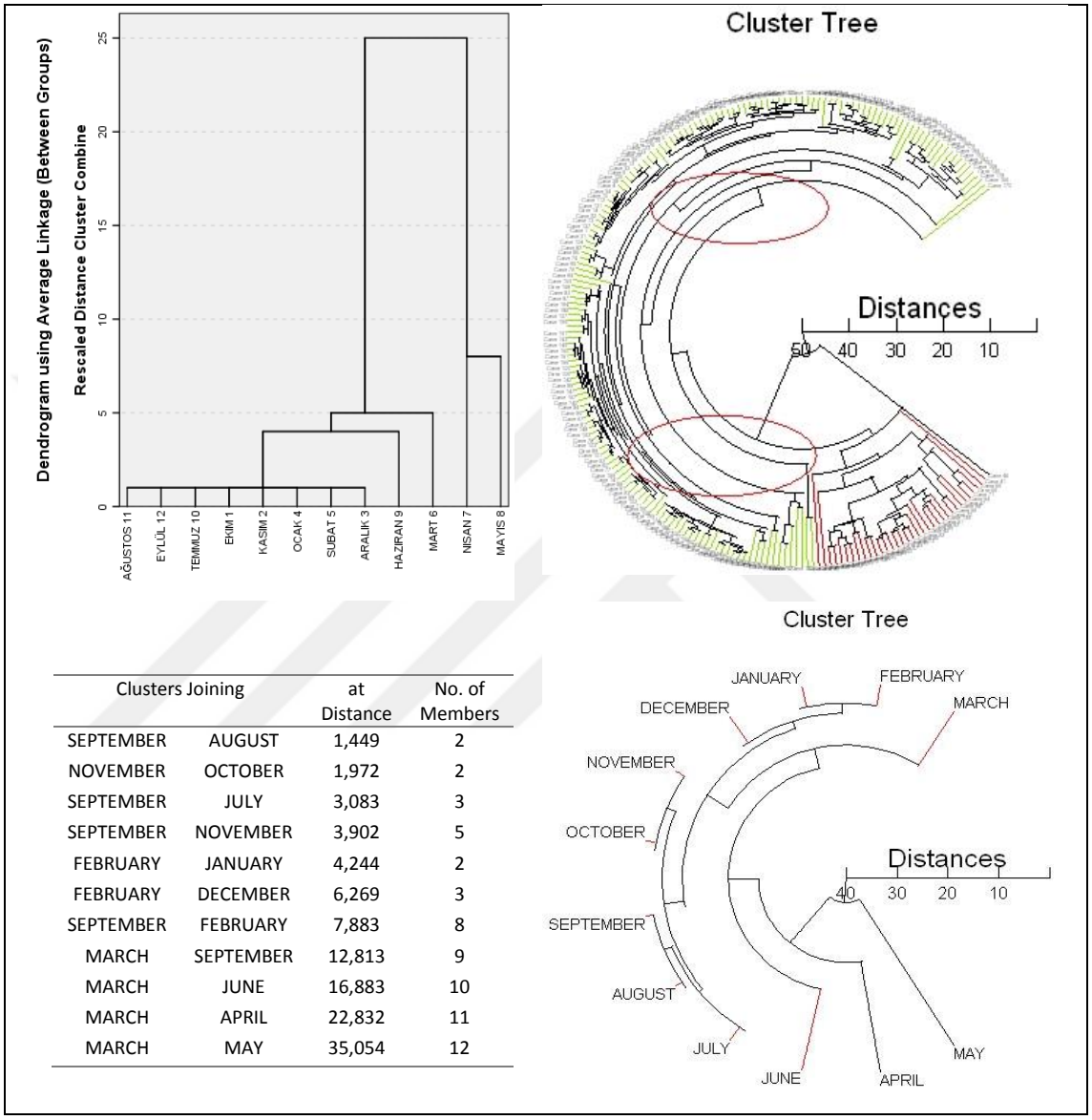
Şekil 4.8 Fırat Dicle Havzası için istasyonların aylar üzerinde genel dağılımı



Şekil 4.9 Fırat Dicle Havzası için istasyonların aylar üzerinde scatter dot grafik görseli

Hiyerarşik Kümeleme Analizi

40 yıllık bir perspektifte aylar bazında benzeşen özelliklerin olup olmadığı kümeleme analizine tabi tutularak araştırılmıştır.



Clusters Joining		at Distance	No. of Members
SEPTEMBER	AUGUST	1,449	2
NOVEMBER	OCTOBER	1,972	2
SEPTEMBER	JULY	3,083	3
SEPTEMBER	NOVEMBER	3,902	5
FEBRUARY	JANUARY	4,244	2
FEBRUARY	DECEMBER	6,269	3
SEPTEMBER	FEBRUARY	7,883	8
MARCH	SEPTEMBER	12,813	9
MARCH	JUNE	16,883	10
MARCH	APRIL	22,832	11
MARCH	MAY	35,054	12

Şekil 4.10 Fırat Dicle Havzası için istasyonların aylar üzerinde hiyerarşik kümeleme görselleri ve çıktıları

Yukarıda elde edilen dendogramlara göre nisan ve mayıs aylarının akım rejimlerinin benzer düzeylerde olması. 1. Kümeyi belirlemiştir. 2. Kümede kalan diğer aylar olarak görülmektedir. Bu da özellikle mevsimsel farklılıkların akım rejimleri açısından oldukça azaldığı sonucunu ortaya koymaktadır. İstasyonlar genel olarak değerlendirildiğinde

kendi aralarında varyansların homojen olmadığı görülmektedir. Ayrıca benzeştiği varsayımı altında bile istasyonların birbirinden farklı akış rejimleri oldukları söylenebilir. Genel olarak ortalama akış ölçümlerinin istasyonlar bazında birbirlerine yakın olduğu görülmektedir.

Hiyerarşik Ward Yöntemi

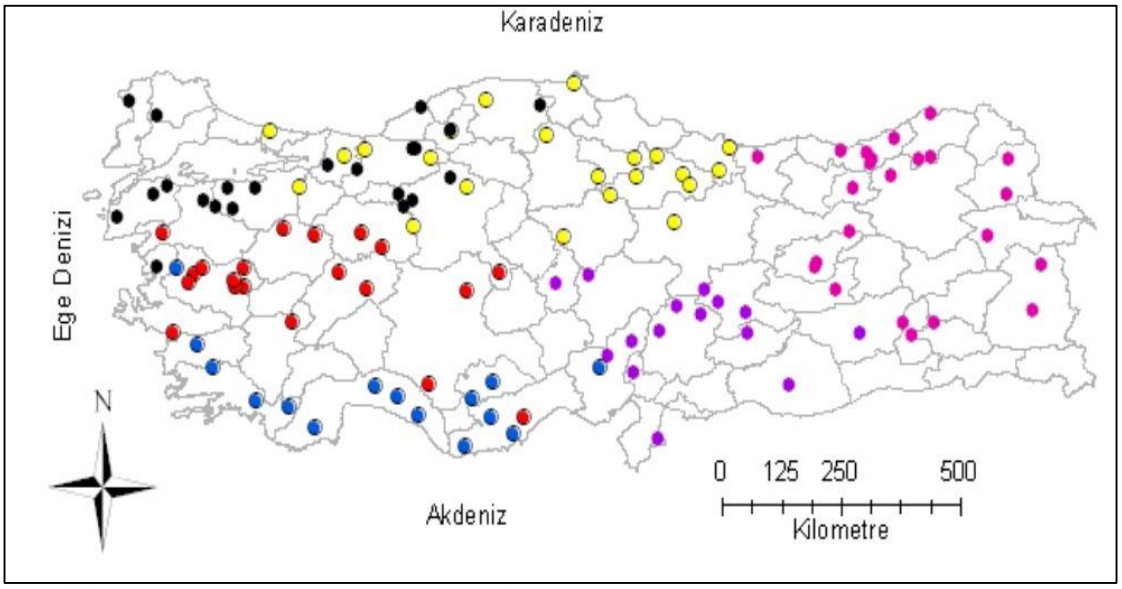
Yıllık maksimum akımların sınıflandırılmasında geleneksel kümeleme yöntemlerinden hiyerarşik kümeleme yöntemi (Ward kümeleme yöntemi) uygulanmıştır. Veriler normalize edildikten sonra kümeleme analizi yapılmış ve daha önceden verilen ölçütler yardımıyla en uygun küme sayısına karar verilmiştir.

En uygun veriseti kullanılarak Ward kümeleme yöntemi ile sınıflandırma yapılmış ve belirlenen bölgeler için uyumsuzluk ve bölgesel homojenlik testi uygulanmıştır[60].

Çizelge 4.3 Ward kümeleme yöntemi ile belirlenen bölgeler için uyumsuzluk ve bölgesel homojenlik testi sonuçları

Bölge	**Bölgede Bulunan İstasyon Sayısı**	**İstasyon Yıl Sayısı**	**H testi**			**İstasyon No. (D değeri)**
			H_1	H_2	H_3	
1	22	682	1,71	0,44	0,00	-
2	20	620	1,49	-0,17	-0,20	713(4,1)
3	15	465	1,62	1,87	1,48	-
4	22	682	2,41	0,63	0,00	1323(4,47)
5	15	465	2,04	0,21	-1,14	2132(3,03)
6	23	713	1,47	1,55	0,98	2232(3,44)

Yapılan değerlendirmelere göre Ward kümeleme yöntemi ile yıllık maksimum akımların sınıflandırılmasında en uygun sonuçlar elde edilmiştir. Bu sonuçlara göre veri seti kullanılarak belirlenen bölgelerin bölgesel frekans analizinde, dağılımların belirlenmesi ve büyüklüklerin tahmininde kullanılabileceği söylenebilir. Veri seti kullanılarak Ward kümeleme yöntemi ile belirlenen bölgelere ait istasyonların dağılımı aşağıdaki şekilde gösterilmektedir [60].



Şekil 4.11 Ward kümeleme yöntemi ile belirlenen bölgeler

K-Ortalamalar Yöntemi

K-Ortalamalar 1. Örnek

Aşamalı olmayan kümeleme metotları arasında en yaygın olarak kullanılan yöntem K-ortalamalar(K-means) tekniğidir. Bu teknikle, küme içerisindeki kareler toplamını en küçük biçimde k sayıda kümeye bölmektedir.

Analiz sürecinde tüm gözlemler gruplara yerleştirildikten sonra bulundukları küme ortalamasından daha yakın küme ortalaması varsa, gözlemlerin yerleri değiştirilerek daha güvenilir kümeler oluşturulmalıdır.

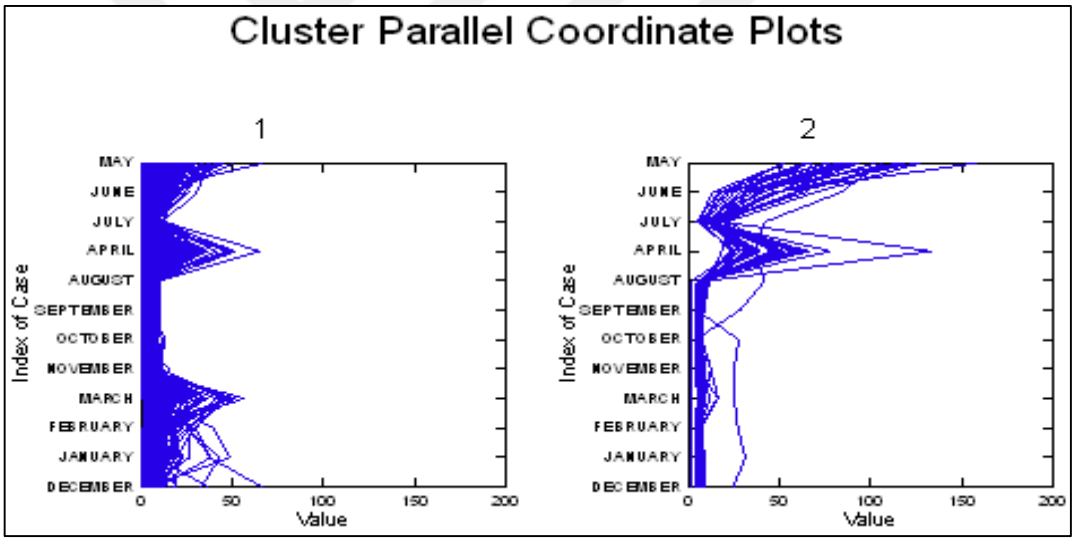
Genel olarak k-means yönteminin amacı diğer kümeleme yöntemlerinde olduğu gibi, gerçekleştirilen kümeleme analizi işlemi sonucunda elde edilen kümelerin, küme içi benzerliklerinin maksimum, kümeler arası benzerliklerinin ise minimum olmasını sağlamaktır. Küme benzerliği kavramı, kümenin ağırlık merkezi kabul edilen bir birim ile kümedeki diğer birimler arasındaki uzaklıkların ortalama değeri ile ölçülmektedir.

Aşağıda hem kategorik hem de sürekli değişkenlerin kullanılabileceği bir yöntemler içeren k-means kümeleme analizi algoritmasının Türkiye geneli ortalama akım dağılımlarının betimleyici istatistikleri verilmiştir.

Çizelge 4.4 k-means çıktıları

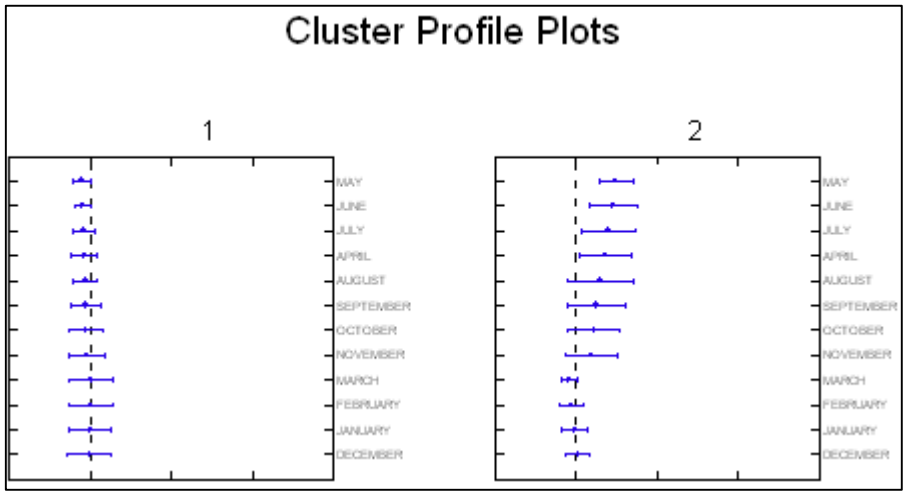
Özet istatistikler					
Variable	**Between SS**	**df**	**Within SS**	**df**	**F-Ratio**
Ekim	305,037	1	1.727,508	198	34,962
Kasım	228,160	1	1.771,911	197	25,367
Aralık	10,645	1	7.837,238	199	0,270
Ocak	16,818	1	9.864,573	199	0,339
Şubat	105,867	1	8.325,943	199	2,530
Mart	607,673	1	22.926,731	199	5,274
Nisan	22.310,195	1	33.670,725	199	131,857
Mayıs	139.314,628	1	47.092,482	199	588,706
Haziran	28.552,744	1	15.816,872	199	359,236
Temmuz	2.650,805	1	2.776,963	199	189,959
Ağustos	906,759	1	2.204,450	199	81,855
Eylül	352,178	1	1.464,282	199	47,862

Bu çizelgeden hareketle en dikkati ayların mayıs, nisan, temmuz ayları olduğu dikkat çekmektedir. Bu aylar daha sonraki kümeleme çıktılarında incelencektir.



Şekil 4.12 Ward kümeleme yöntemi ile belirlenen bölgeler

Paralel plot çıktıları her bir oluşan küme için oluşturulur. Kümeler arasında farklılığa neden değişkenlerin belirlenmesinde önemli bir görsellik aracı olarak kullanılır. Daha öncede belirttiğimiz gibi özellikle mayıs ve nisan ayları en dikkat edilmesi gereken aylardır.



Şekil 4.13 Ward kümeleme yöntemi ile belirlenen bölgeler

Küme profilleri ise üstteki şekilden rahatlıkla elde edilebilir ve skor çıktıları kullanılarak diğer çok değişkenli analizleri yapmamıza olanak sağlar. Burada özellikle ikinci küme üzerinde ciddi farklılıklar incelenerek, bu farklılıkların entegre havza sistematiğine kattığı avantaj ve dezavantajlar tartışılmalıdır.

K-Ortalamalar 2. Örnek

Tez çalışması kapsamında yağış serilerinin sınıflandırılmasında K-Ortalamalar kümeleme yöntemi kullanılmıştır. K-Ortalamalar kümeleme yöntemi ile yağış serileri sınıflandırılmış ve bölgesel homojenlik testi uygulanmıştır. Elde edilen sonuçlar bulanık kümeleme yöntemi sonuçları ile karşılaştırılarak değerlendirilmiştir[60].

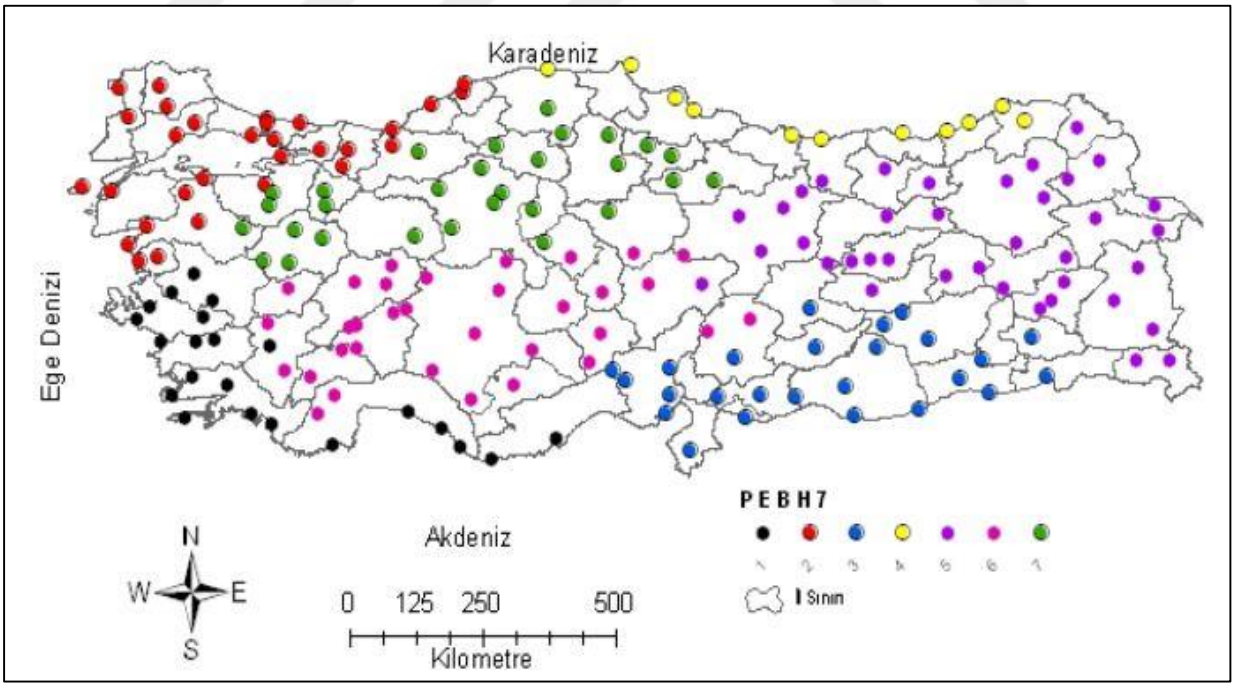
Çizelge 4.5 K-Ortalamalar kümeleme yöntemi ile belirlenen bölgeler için uyumsuzluk ve bölgesel homojenlik testi sonuçları

Bölge	Bölgede Bulunan İstasyon Sayısı	İstasyon Yıl Sayısı	H testi			İstasyon No. (D değeri)
			H_1	H_2	H_3	
1	22	682	-1,53	-2,33	-1,87	17232(3,40)
2	30	930	-0,46	-1,05	-0,88	17110(3,08)
3	25	775	-0,60	-1,67	-1,58	-
4	11	341	1,85	-0,78	-0,72	-
5	38	1178	1,17	0,66	0,15	17088(3,65)
6	33	1023	-0,09	-0,79	-0,91	17926(4,69)
7	29	899	-3,30	-3,55	-4,02	17748(4,05)

Çizelgede uyumsuzluk sonuçları incelendiğinde genel olarak benzer istasyonlara ait D değerlerinin sınır değerden (D > 3) büyük çıktığı görülmektedir. Veri Seti 1 için hesaplanan H değerleri incelendiğinde, bölge 7 en büyük H1 (-3.300) değerine sahip olup bu değer kabul edilebilir sınırdan daha büyük olduğu görülmektedir. Diğer bölgeler için hesaplanan H1 değerleri ise 2'den küçük çıkmış ve "kesinlikle heterojen kabul edilir" şeklinde değerlendirilmiştir[60].

Çizelge 4.6 K-Ortalamalar kümeleme yöntemi ile belirlenen bölgeler için uyumsuzluk ve bölgesel homojenlik testi sonuçları

Bölge(T.C)	Bölgede Bulunan İstasyon Sayısı	Maksimum			Minimum		
		Yağış (mm)	Değiş. Kats.	Std. Sapma	Yağış (mm)	Değiş. Kats.	Std. Sapma
1	22	1167,99	0,29	254,8	745,2	0,23	169,5
2	30	1211,48	0,30	216,9	729,3	0,20	145,7
3	25	1099,73	0,34	267,5	645,0	0,26	166,7
4	11	2203,28	0,21	286,1	1204,0	0,14	158,8
5	38	1287,51	0,29	311,1	562,1	0,20	117,7
6	33	745,00	0,31	151,3	452,9	0,20	90,2
7	29	1490,18	0,23	274,9	517,9	0,18	93,0



Şekil 4.14 Yağış serilerinin sınıflandırılmasında K-Ortalamalar kümeleme yöntemi ile belirlenen bölgelerde istasyonların dağılımı

K-Ortalamalar 3. Örnek

gCLUTO(Graphical Clustering Tool Kit) Yöntemi

gCLUTO birinci ve ya ikinci vektörü uzaklık ölçüleriyle beraber kullanarak k-means kümeleme görsellerine bir alternatif oluşturmuştur. Aynı zamanda bu method, içinde repeated bisections, direct, nearest neighbor, agglomerative ve biased agglomerative k-means algoritmalarını içerir ve uygulamalarda alternatiflik sağlar. Bu algoritmada cosine benzerlik ölçüsüyle beraber I2 kriter fonksiyonları kullanılmıştır.

Cosine benzerlik ölçütü:

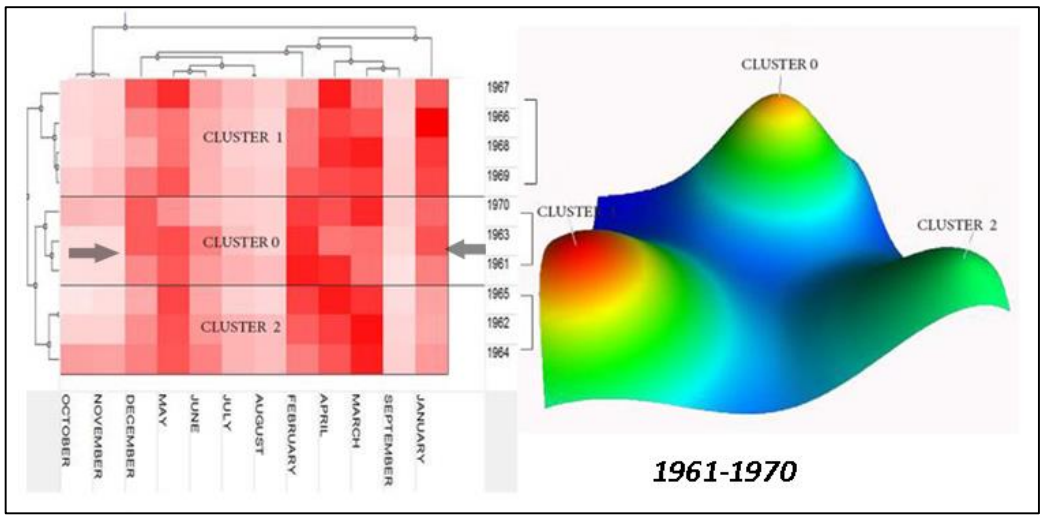
$$\cos(c_1, c_2) = (c_1 . c_2) / \|c_1\| \|c_2\| \quad (4.1)$$

I2 kriter fonksiyonu:

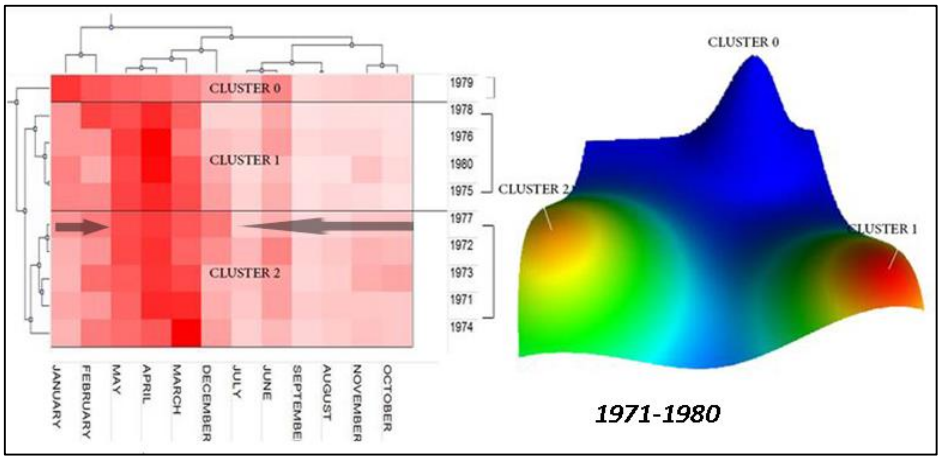
$$\sum_{r=1}^{k} \sum_{d_i \epsilon S_r} cos\,(d_i, C_r) = \sum_{r=1}^{k} \|D_r\| \quad (4.2)$$

gCLUTO K-Means yöntemleri içinde ileri derecede kaliteli ve uygun çıktılar veren bir türdür. Çıktılarına sırasıyla matris ve dağ görünümü (matrix and mountain visualization) adı verilir [61].

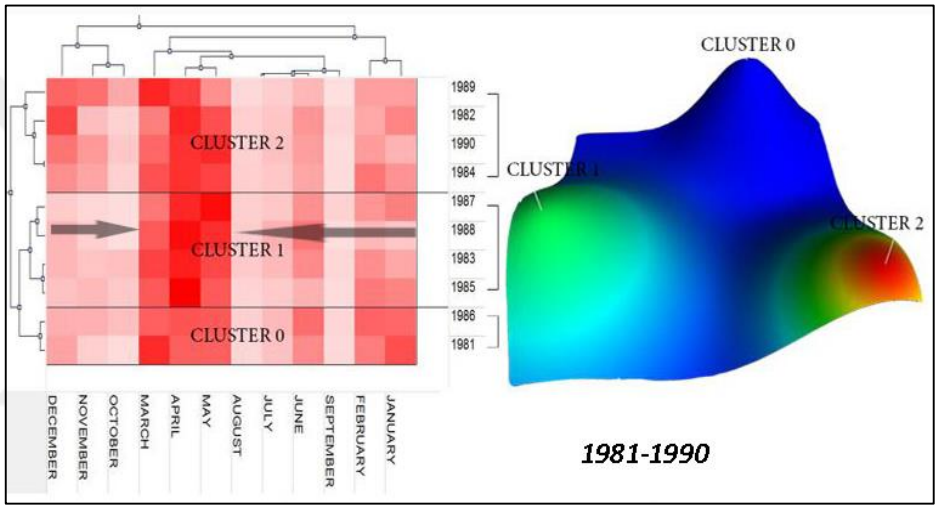
Burada gCLUTO yöntemiyle akımın tüm Türkiye bazında çapraz dendogramlar yardımıyla nasıl oluştuğunu inceleyeceğiz. 40 yıllık (1960-2000) akım verisi DSİ den alınarak analize dahil edilmiştir [61].



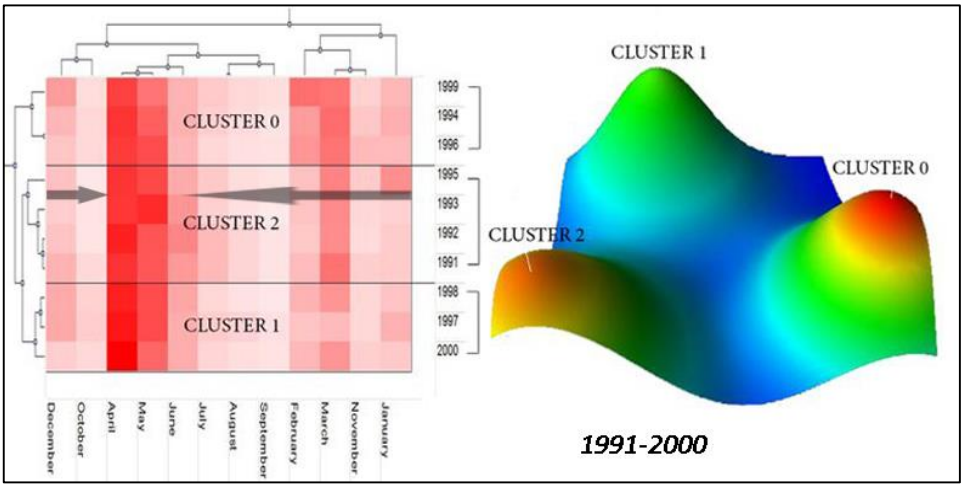
Şekil 4.15 gCLUTO kümeleme çıktısı (1961-1970)



Şekil 4.16 gCLUTO kümeleme çıktısı (1971-1980)



Şekil 4.17 gCLUTO kümeleme çıktısı (1981-1990)



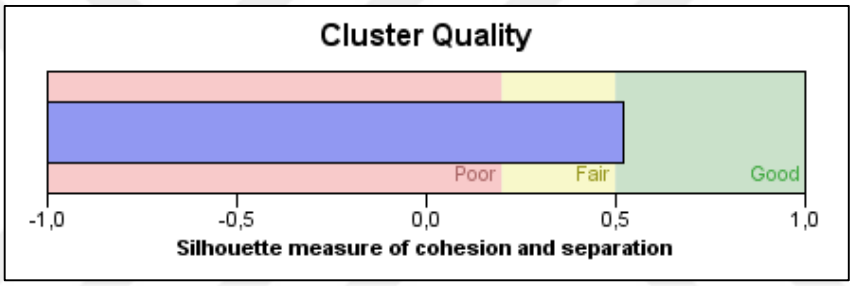
Şekil 4.18 gCLUTO kümeleme çıktısı (1991-2000)

Bu şekillerde oklar yönleriyle beraber mevsimsel değişkenliğin giderek yok olduğunu göstermektedir. Koyu kırmızı bölgeler değişkenden en yoğun şekilde etkilenen yerleri belirtir. Renklerin giderek yok olması değerin sıfıra yaklaştığını gösterir.

Çıktılarından da görüldüğü gibi mevsimsel değişim kırmızı bölgelerde havzaları en çok etkilemektedir. Mart ve Nisan aylarının Türkiye için en önemli dönem noktaları olduğunu gCLUTO k-means algoritmasından söyleyebiliriz.

Two Step Cluster Uygulaması

Bu uygulamada two step algoritması kullanılmıştır. Dicle Fırat havzasının istasyonları kümeleme analizine kategorik değişken olarak girmiştir. Uygun küme sayısını belirleme kriterlerinden bir diğeri de Silhouette indeks değeridir.

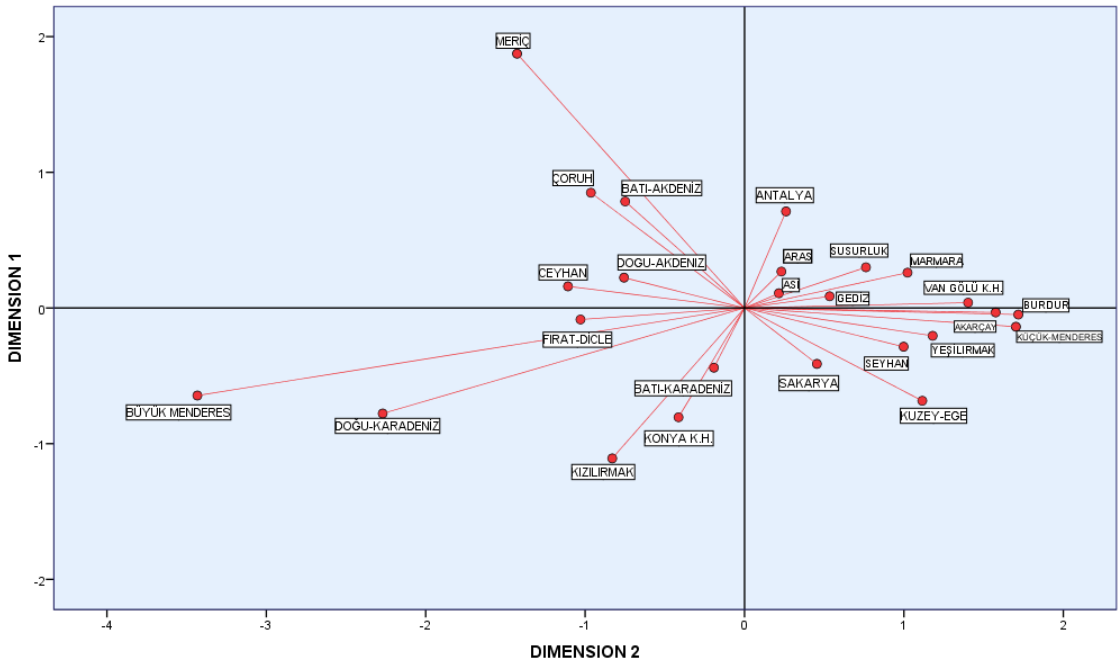


Şekil 4.19 Dicle- Fırat Havzası İçin Silhouette indeksi çıktısı

Oluşan kümelerin değerlendirmesinde Silhouette indeksi kullanılmış ve bu değerin iyi derecede olduğu görülmüştür. Yani kümeler Dicle-Fırat havzası için iyi bir şekilde oluşmuştur.

Kümeleme Analizinin Çeşitli Çok Değişkenli Analizlerle İlişkisinin Hidrolojik Veriler Üzerinden İncelenmesi

Küme analizi sonuçlarını görselleştirmek için yapılan en önemli istatistiksel analizlerden biri de Çok Boyutlu Ölçekleme ‘dir. Tüm Türkiye için yapılan hiyerarşik ve k-means algoritmalarından mevsimsel boyutların iki küme oluşturduğu görülmüştür. Toplam akımlar MDS analizi için iki boyutu oluşturarak kümeleme analizi sonuçlarının yorumlanmasına katkı sağlamıştır.



Şekil 4.20 Kümeleme sonuçlarından oluşan kümelerin MDS ile gösterimi

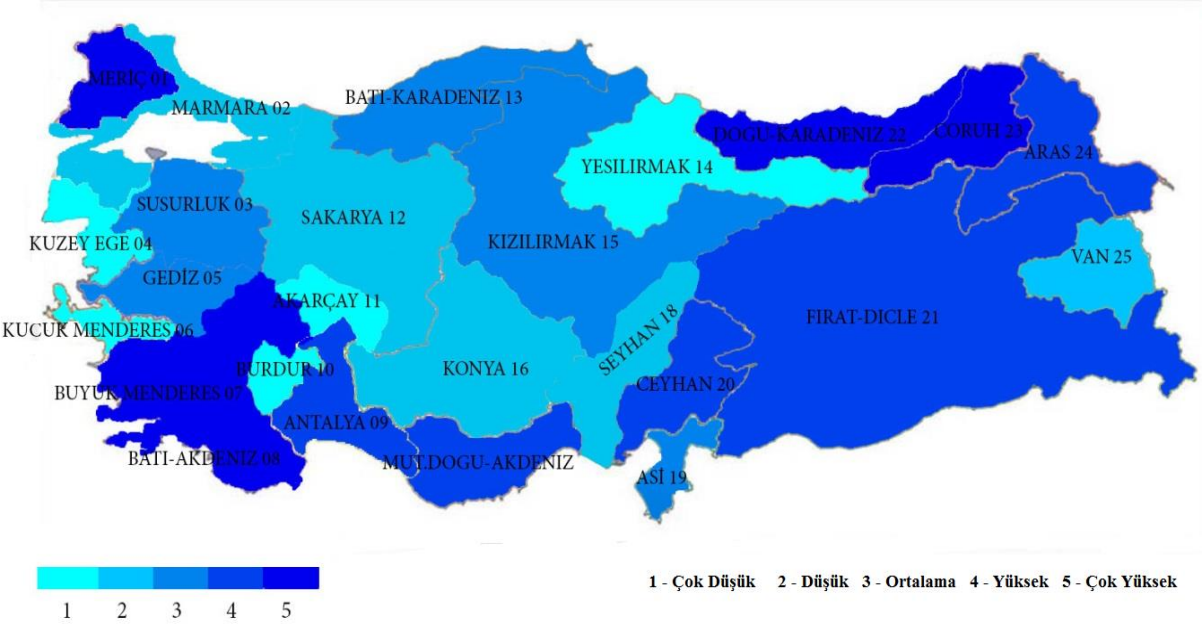
Bu grafikten hareketle birinci boyut yüksek akım bölgesini, ikinci boyut ise düşük akım bölgesini temsil eder. Merkezden uzaklaştıkça havzalar arası heterojenliğin arttığı görülmektedir. Büyük Menderes, Doğu Karadeniz, Meriç, Kuzey-Ege havzalarının en heterojen, Asi, Aras ve Gediz havzalarının ise homojen olduğu görülmüştür.

Kümeleme analizi ile birlikte en çok kullanılan analizlerden biriside Temel Bileşen Analizi(TBA) 'dir. Kümeleme analizinde oluşan kümeler Temel Bileşen Analizi ile beraber boyut indirgeme prosedürüne sokulmuştur. Kümeleme çıktısı olan yüksek ve düşük akım boyutlarının kendi içinde açıklanma oranları TBA ile görülebilir. Çizelge bize ilk bileşenin açıklanma oranının %84 olduğunu söyler.

Çizelge 4.7 TBA için oluşan ana iki kümenin açıklanma oranı

Bileşenler	**Açıklanan Toplam Varyans**					
	Toplam	% of Varyans	Kümülatif	Toplam	% of Varyans	Kümülatif
Yüksek Akım	1,67	83,923	83,923	1,67	83,923	83,923
Düşük Akım	,322	16,077	100,000			

Kümeleme analizinin Çok Boyutlu Ölçekleme ve Temel Bileşen Analizi ile beraber kullanılmasıyla beraber Türkiye için oluşturulabilecek yeni havza bölgeleri oluşturulabilir. Türkiye Cumhuriyeti entegre havza yönetimi için kümeleme analizi baz alınarak aşağıda haritadan yararlanılarak yeni su havza haritaları oluşturulabilir.



Şekil 4.21 Türkiye bazında kümeleme analiziyle beraber oluşabilecek havza bölgeleri

Bu harita Türkiye'de havzalar için yüksek akışlar bazında nasıl entegre olabileceğinin bir sonucudur. Marmara, Sakarya, Konya Kapalı ve Seyhan havzalarının sınır olması ve ortak klimatik özelliklere sahip olması bu havzalar için acil havza yönetiminin gerekli olduğunu göstermektedir. Fırat-Dicle, Doğu Karadeniz, Çoruh, Büyük Menderes ve Meriç havzalarının yüksek akış özelliğiyle ana havza gruplarının başında olabileceği ya da oluşturulacak entegre havzalarda öncelik sahibi olabileceği rahatlıkla söylenebilir.

BÖLÜM 5

SONUÇLAR ve YORUM

Türkiye bulunduğu bölge itibariyle küresel ısınmanın ve iklim değişikliği tehdidinin ciddi etkisi altındadır. Bundan dolayı entegre havza yönetimi politikalarının önemli derecede bu değişmelerin etkisinde olacağını öngörebiliriz. Entegre havza yönetimini direk olarak etkileyen bu ölçümler gerekli birimlerce taşkın, kuraklık, kuru dönemler ve bunların açacağı sosyal sorunlar dikkate alınarak ölçülmeli ve halkın farkındalığına sunulmalıdır. Bu değişimler sadece su üzerine değil, suyla bağlantısı olan her alan üzerinde bir tehdit oluşturmaktadır.

Birleşmiş Milletler İklim Değişikliği Konferansında(United Nations Climate Change Conference in Paris, December 2015) iklim değişikliği dünya üzerindeki en büyük tehdit olarak kabul edilmiştir. Bu kararın entegre havza yönetimi politikalarının oluşturulmasında dikkat edilmesi gereken en önemli husustur. Bu çalışmada iklim değişikliğine neden olan etkenlerin Türkiye üzerindeki etkileri çeşitleri istatistiksel yöntemlerle incelenmiş ve ileri düzey istatistiksel metotlarla da entegre havza yönetimi politikalarına yardımcı olacak çıkarımlarda bulunulmuştur.

Bu tez çalışmasında Devlet Su İşleri (DSİ), Meteoroloji Genel Müdürlüğü(MGM) ve diğer uluslararası kaynaklardan elde edilen ve analizler için düzenlenen kaliteli ve güçlü veriler doğrultusunda “Entegre Havza Yönetimi” alanında çalışmalar ve analizler yapılmıştır. Havzalara etki eden en önemli ayın Nisan ve Mayıs olduğu görülmüştür. K-ortalamalar yönteminde kullanılan yeni bir görsel metot olan gCLUTO algoritmasının iklim değişikliğini ve mevsimsel hareketleri çok iyi bir biçimde göstermesi alınacak önlemler açısından önemli bir çıktıdır.

Türkiye' de havzalar üzerinde mevsimsel farklılığın anlaşılması için kullandığımız gCLUTO görsel kümeleme analizi aracı bize Türkiye' de mevsim çeşitliliğinin iki mevsime doğru hızlı bir şekilde gittiğini göstermektedir. Bu sonuç Türkiye'nin havzalarının direk olarak iklim değişikliği tehdidi altında olduğunu net bir şekilde göstermektedir.

Entegre havza yönetiminde kümeleme analizi yöntemleriyle beraber su havzaları üzerinde mevsimsel geçişliliğin sıcaklık, yağış ve akım değişimleri ile beraber açıkça ortaya çıktığı görülmüştür. Bu değişimin direk olarak havzalar üzerinde su güvenliğinin, tarımın, endüstri hammaddelerinin, enerjinin ve ilişkili etmenlerin etkileneceği unutulmamalıdır.

Kümeleme analizlerinden çıkarabileceğimiz ana sonuç Türkiye bazında havzalar üzerinde dört mevsim etkisinden iki mevsim etkisine doğru hızlı bir gidiş olduğudur. Çeşitli kümeleme tekniklerinden hareketle Türkiye su havza sayıların günümüzde çok fazla olduğu ve bu nedenle havza yönetimlerinde ciddi sorunlar ortaya çıktığı görülmektedir. Hiyerarşik, k-means ve diğer kümeleme yöntemleri kullanılarak alternatif homojen ve heterojen havza bölgeleri belirlenmiştir. Bu bölgelerden yararlanılarak yeni havza bölgeleri oluşturulabilir ve entegre havza politikaları, havzaların bölgeselleştirilmesinde kümeleme analizi ana yöntem olarak kullanılabilir.

KAYNAKÇA

[1] Kara, İ., (1994). Olasılık, Bilim Teknik Yayınevi, Ankara.

[2] Kendall, M. G., Buckland, W. R., & Schneider, C. S. (1972). "A dictionary of statistical terms", American Journal of Physics, 40(11): 1720-1720.

[3] Everitt S.B., Dunn G., (1991). Applied Multivariate Data Analysis, John Wiley and sons, New York.

[4] Özdamar, K., (1988). "Hastalık Olgularının İncelenmesinde Kümeleme Çözümlemesinin Kullanılması", Anadolu Üniversitesi Yayınları, No:295.

[5] Edwards, A. W., & Cavalli-Sforza, L. L. (1965). "A method for cluster analysis", Biometrics, 9: 362-375.

[6] Scott, A. J., & Knott, M. (1974). "A cluster analysis method for grouping means in the analysis of variance", Biometrics, 10: 507-512.

[7] Everitt, B., Landau, S., Leese, M., Stahl, D., (2011). Cluster analysis, Wiley.

[8] Littmann, T. (2000). "An empirical classification of weather types in the Mediterranean Basin and their interrelation with rainfall", Theoretical and Applied Climatology, 66(3): 161-171.

[9] Kendall, D. G. (1969). "Some problems and methods in statistical archaeology", World Archaeology, 1(1): 68-76.

[10] Hietala, H. (1984). Intrasite spatial analysis in archaeology, CUP Archive.

[11] Anderberg, M.R., (1973). Cluster analysis for applications, Academic Press, New York.

[12] Bryan F.J., (1994). Multivariate Statistical Methods, Second Edition, Chapman – Hall, Londra.

[13] Alpar, R., (2013). Uygulamalı Çok Değişkenli İstatistiksel Yöntemler, Syf:318-319.

[14] Alpar, R., (2013). Uygulamalı Çok Değişkenli İstatistiksel Yöntemler, Ankara.

[15] Danielsson, P. E. (1980). “Euclidean distance mapping. Computer Graphics and image processing”, 14(3): 227-248.

[16] de Souza, R. M., De Carvalho, F. D. A. (2004). “Clustering of interval data based on city–block distances”, Pattern Recognition Letters, 25(3): 353-365.

[17] De Maesschalck, R., Jouan-Rimbaud, D., Massart, D. L. (2000). “The mahalanobis distance”, Chemometrics and intelligent laboratory systems, 50(1): 1-18.

[18] Lohninher, H., Grundlagen der Statistik ,Mahalanobis-Distanz, www.statistics4u.info/fundstat_germ/ee_mahalanobis_distance.html, 20 Şubat 2013.

[19] Teknomo, K. (2015). Chebyshev Distance.

[20] Roche, A., Malandain, G., Pennec, X., & Ayache, N. (1998). “The correlation ratio as a new similarity measure for multimodal image registration”, Medical Image Computing and Computer-Assisted Interventation, 11:1115-1124.

[21] Büyüköztürk, Ş. (2002). “Faktör analizi: Temel kavramlar ve ölçek geliştirmede kullanımı”, Kuram ve uygulamada eğitim yönetimi, 32(32): 470-483.

[22] Duda, R. O., Hart, P. E., Stork, D. G. (2001). Pattern classification, 2nd Edition, New York.

[23] Truchon, M. (1998). An extension of the Condorcet criterion and Kemeny orders, Cahier, 9813.

[24] Aldenderfer, M.S., Blashfield R.K., (1984). Cluster Analysis, Sage Publications.

[25] Everitt, B., (1974). Cluster Analysis. Heinemann, London.

[26] Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C., (1995). Multivariate data analysis.Prentice-Hall, A Simon&Schuster Company, New Jersey.

[27] Lance, G.N., Williams,W.T., (1967). “A General Theory of classificatory sorting strategies:1-Hierarcihical systems”, Computer Journal, 9:373-380.

[28] Çokluk, Ö., Şekercioğlu G., Büyüköztürk, Ş., (2010). Sosyal Bilimler için Çok Değişkenli İstatistik Spss ve Lisrel Uygulamaları, Pegem Akademi, Ankara.

[29] Bazinet, A. L., & Cummings, M. P. (2012). “A comparative evaluation of sequence classification programs”, BMC bioinformatics, 13(1): 92.

[30] Gower, J. C., & Ross, G. J. S. (1969). “Minimum spanning trees and single linkage cluster analysis”, Applied statistics, 6: 54-64.

[31] Çokluk, Ö., Şekercioğlu, G. Büyüköztürk, Ş., (2010). Sosyal Bilimler için çok Değişkenli İstatistik Spss ve Lisrel Uygulamaları, Pegem Akademi, Ankara.

[32] Gower,J.C.,(1967). “Comparison of some methods of cluster analysis”, Biometrics, 23:623-628.

[33] Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis, John Wiley & Sons.

[34] Jiawei, H., & Kamber, M. (2001). Data mining: concepts and techniques, Morgan Kaufmann San Francisco, CA.

[35] MacQueen, J., (1967). “Some methods for classification and analysis of multivariate observations”, In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, June 1967, 1:281-297.

[36] Xu, R., & Wunsch, D. (2005). “Survey of clustering algorithms”, IEEE Transactions on neural networks, 16(3): 645-678.

[37] Tan, P. N. (2006). Introduction to data mining, Pearson Education India.

[38] Stetsyura, S. V., Klimova, S. A., Wenig, S. B., Malyar, I. V., Arslan, M., Dincer, I., & Elerman, Y. (2012). “Preparation and probe analysis of Langmuir–Blodgett films with metal-containing dendritic and cluster structures”, Applied Physics A, 109(3): 571-578.

[39] Dasu, T., & Johnson, T. (2003). Exploratory data mining and data cleaning, John Wiley & Sons.

[40] Woodfine, K., Fiegler, H., Beare, D. M., Collins, J. E., McCann, O. T., Young, B. D., & Carter, N. P. (2003). “Replication timing of the human genome”, Human molecular genetics, 13(2): 191-202.

[41] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques, Elsevier.

[42] Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). “Density-based clustering”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3): 231-240.

[43] Birant, D., & Kut, A. (2007). “ST-DBSCAN: An algorithm for clustering spatial–temporal data”, Data & Knowledge Engineering, 60(1): 208-221.

[44] Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). “OPTICS: ordering points to identify the clustering structure”, In ACM Sigmod record, 28(2): 49-60.

[45] Park, N. H., & Lee, W. S. (2004). “Statistical grid-based clustering over data streams”, ACM SIGMOD Record, 33(1): 32-37.

[46] Joshi, A., & Kaur, R. (2013). “A review: Comparative study of various clustering techniques in data mining”, International Journal of Advanced Research in Computer Science and Software Engineering, 3(3).

[47] Wang, W., Yang, J., & Muntz, R. (1997). “STING: A statistical information grid approach to spatial data mining”, In VLDB, 97:186-195.

[48] Toledo, M. D. G. (2005). A comparison in cluster validation techniques, University Of Puerto Rico.

[49] Rousseeuw, P. J. (1987). “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”, Journal of computational and applied mathematics, 20: 53-65.

[50] Amasyali, M. F., & Ersoy, O. (2008). “Cline: A new decision-tree family”, IEEE Transactions on Neural Networks, 19(2): 356-363.

[51] Johnson, R. A., & Wichern, D. W. (1988). Multivariate statistics: a practical approach.

[52] Abello, J., Kobourov, S. G., & Yusufov, R. (2004). “Visualizing large graphs with compound-fisheye views and treemaps”, In International Symposium on Graph Drawing , Springer Berlin Heidelberg, 443-441.

[53] Martinez, A. R., & Martinez, W. L. (2004). Model-based clustering toolbox for MATLAB, Naval Surface Warfare Center, Dahlgren Division.

[54] Jolliffe, I. (2002). Principal component analysis, John Wiley & Sons, Ltd.

[55] Kruskal, J. B. (1977). “Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics”, Linear algebra and its applications, 18(2): 95-138.

[56] Tatlıdil, H., (2002). Uygulamalı Çok Değişkenli İstatistiksel Analiz, Akademi Matbaası, Ankara.

[57] Yıldız Teknik Üniversitesi, Bilimsel Araştırma Projeleri Koordinatörlüğü, 2014-01-05-KAP01, “Türkiye’de Havzalar Bazında Su Arz Ve Talebinin Sosyal Ekonomik Ve Hidrolojik Etkilerinin Belirlenmesi, Modellenmesi Ve Gelecek Tahmini” Haziran, 2016.

[58] Kahya, E. ve Kalaycı, S., (2004). “Trend analysis of streamflow in Turkey”, Journal of Hydrology, 289: 128-144.

[59] BİLEN, Özden., (2008). Türkiye’nin Su Gündemi Su Yönetimi ve AB Su Politikaları, Ankara.

[60] Bulanık C-Ortalamalar Yöntemi ile Türkiye Havzalarının Sınıflandırılması, Homojen Hidrolojik Havzaların Belirlenmesi ve Bölgesel Frekans Analizi, Proje No: 107Y318, TÜBİTAK.

[61] Rasmussen, M, ve George K., (2004). "gcluto: An interactive clustering, visualization, and analysis system", UMN-CS, 04: 21.

EK A

HAVZA İSİMLERİ VE İSTASYON NUMARALARI

Havza	İstasyon Numaraları
1.MERIC	003-008-011-012-013-014-020-031-039-046-052-055-057-058
2.MARMARA	003-004-010-015-028-030-031-038-041-047-057-065-069-070-84-084a
3.SUSURLUK	004-008-010-013-017-018-020-021-027-028-034-038-046-049-054-064-072-076-084-085-086-087-089-090-095
4.KUZEY-EGE	002-005-006-013-014-016-017-019-021-023-024-025-026-028-031-032-034-035-036-037
5.GEDIZ	007-015-018-021-025-026-028-031-033-034-035-039-041-042
6.KUCUK-MENDERES	001-006-007-010-011-012-013
7.BUYUK-MENDERES	003-004-007-008-009-010-014-026-030-032-035-037-039-043-049-052-053-059-061-062-065-071-073-075-081-082-083-087-100
8.BATI-AKDENIZ	001-009-011-013-018-018-019-028-049-054-055-058-061-065-067-070-072-075-076-079-081-083-085-087-089-091-092-095-096-103-105-108
9.ANTALYA	002-006-011-013-016-039-042-053-056-065-067-068-070-084-086-088-089-091-100
10.BURDUR	002-003-007-010-013-023-027
11.AKARCAY	001-002-007-012-013-015-017-020-021
12.SAKARYA	006-017-030-033-034-049-055-063-067-075-087-090-101-120-134-135-138-139-146-152-153-155-156-159-170-172-173-175-181-185-188-189-197
13.BATI-KARADENIZ	014-015-019-020-021-022-030-031-032-033-036-037-039-040-041-044-045-049-050-053
14.YESILIRMAK	011-014-015-020-023-024-026-032-033-042-046-047-062-080-092-093-095-106-108-111-117-119
15.KIZILIRMAK	010-015-016-017-021-025-026-029-041-043-046-047-048-049-053-054-055-056-057-059-062-066-067-068-069-071-074-078-088-091-092-093-095-098-099-107-108-112-117-126-131-133-134-137-148-152-153-162-163-167-168-169-171-172-177-178-179-180-182-185-186-187-188-190-191-194-201-206-208-211-212-218-236-238
16.KONYA	03a-013-015-017-026-028-037-041-048-056a-065-072-078-080-081-083-087-088-088-089-092-094-095-096-098-100-101-104-109-114-115-117-123-126-135-136-137-138-140-141-143-144-145-148
17.M.DOGU-AKDENIZ	007-011-012-014-016-017-024-027-029-031-032-033-035-038
18.SEYHAN	005-008-012-016-017-018-019-022-023-027-028-031-032
19.ASI	004-006-009-012-014-016-017-018-020-021-022-023
20.CEYHAN	002-004-005-006-007-008-010-013-015-016-017-031-035-036-040-042-043-044-045-046-051-052-053-054-055-058-059-065
21.FIRAT-DICLE	001-004-019-026-040-052-056-074-084-085-087-096-099-101-103-104-105-116-121-132-133-138-138-140-141-148-152-157-166-167-168-169-170-173-175-181-193-200-201-203-210-213-216-220-226-243-008a-028a-040a

22.DOGU-KARADENIZ	002-006-007-009-013-034-040-044-049-052-055-056-058-059-061-062-063-064-066-067-072-073-074-080-082-085-087-090-092
23.CORUH	001-003-006-007-015-016-024-026-028-029-030-031-032
24.ARAS	011-016-018-022-023-025-029-032-040-041-046-048-049-050-052-053-054-058-060-063-065-069-071-073
25.VAN	002-012-013-021-024

EK B

AYLAR VE OLUŞTURDUĞU KÜMELER

Havzalar	Küme 1	Küme 2
Akarcay	MAR,MAY,APR	JUN,JUL,AUG,SEP,OCT,NOV,FEB,JAN, DEC
Antalya	MAR,APR,MAY,DEC,JAN,FEB	JUN,JUL,AUG,SEP,OCT,NOV
Aras	APR,MAY	JUN,JUL,AUG,SEP,OCT,NOV,DEC,JAN,FEB,MA
Asi	JAN,FEB,MAR	APR,MAY,JUN,JULY,AUG,SEP,OCT,NOV,DEC
Bati-Akdeniz	DEC,JAN,FEB,MAR,APR	MAY,JUN,JUL,AUG,SEP,OCT,NOV
Bati-Karadeniz	MAR,APR	MAY,JUN,JUL,AUG,SEP,OCT,NOV,DEC,JAN,FE
Burdur	JAN,FEB,MAR,APR	MAY,JUN,JUL,AUG,SEP,OCT,NOV,DEC
Buyuk Menderes	JAN,FEB,MAR	APR,MAY,JUN,JULY,AUG,SEP,OCT,NOV,DEC
Ceyhan	MAR,APR,MAY	JUN,JUL,AUG,SEP,OCT,NOV,DEC,JAN,FEB
Coruh	APR,MAY,JUN	JUL,AUG,SEP,OCT,NOV,DEC,JAN,FEB,MAR
Dogu-Karadeniz	APR,MAY	JUN,JUL,AUG,SEP,OCT,NOV,DEC,JAN,FEB,MA
Firat-Dicle	APR,MAY	JUN,JUL,AUG,SEP,OCT,NOV,DEC,JAN,FEB,MA
Gediz	JAN,FEB,MAR,APR	MAY,JUN,JUL,AUG,SEP,OCT,NOV,DEC
Kizilirmak	APR	JUN,JUL,AUG,SEP,OCT,NOV,DEC,JAN,FEB,MA
Konya	SEP,OCT,NOV,DEC,JAN,FEB,	APR,MAY,JUN,JUL,AUG
Kuçuk Menderes	JAN	MAY,JUN,JUL,AUG,SEP,OCT,NOV,DEC,FEB,MA
Kuzey Ege	JAN	MAY,JUN,JUL,AUG,SEP,OCT,NOV,DEC,FEB,MA
Marmara	DEC,JAN,FEB,MAR	APR, MAY,JUN,JUL,AUG,SEP,OCT,NOV
Meric	DEC,JAN,FEB,MAR,APR,MAY	JUN,JUL,AUG,SEP,OCT,NOV
Dogu-Akdeniz	MAR,APR,MAY	JUN,JUL,AUG,SEP,OCT,NOV,DEC,JAN,FEB
Sakarya	MAR,APR	MAY,JUN,JUL,AUG,SEP,OCT,NOV,DEC,JAN,FE
Seyhan	APR,MAY	JUN,JUL,AUG,SEP,OCT,NOV,DEC,JAN,FEB,MA
Susurluk	JAN,FEB,MAR,APR	MAY,JUN,JUL,AUG,SEP,OCT,NOV,DEC
Van	APR,MAY,JUN	JULY,AUG,SEP,OCT,NOV,DEC,JAN,FEB,MAR
Yesilirmak	APR,MAY	JUN,JUL,AUG,SEP,OCT,NOV,DEC,JAN,FEB,MA

EK C

MDS İÇİN BOYUTLAR

HAVZA	Boyut 1	Boyut 2
Akarcay	1.5757	-0.0330
Antalya	0.2605	0.7121
Aras	0.2313	0.2685
Asi	0.2155	0.1085
Bati-Akdeniz	-0.7487	0.7853
Bati-Karadeniz	-0.1920	-0.4399
Burdur	1,7175	-0.0483
Buyuk Menderes	-3,4326	-0.6452
Ceyhan	-1,1078	0.1593
Coruh	-0.9640	0.8496
Dogu-Karadeniz	-2,2700	-0.7777
Firat-Dicle	-1,0290	-0.0850
Gediz	0.5341	0.0865
Kizilirmak	-0.8300	-1,1084
Konya	-0.4143	-0.8056
Kuçuk Menderes	1,7026	-0.1380
Kuzey Ege	1,1151	-0.6838
Marmara	1,0224	0.2605
Meric	-1,4270	1,8741
Dogu-Akdeniz	-0.7560	0.2236
Sakarya	0.4541	-0.4120
Seyhan	0.9981	-0.2855
Susurluk	0.7617	0.2999
Van	1,4027	0.0393
Yeşilirmak	1,1803	-0.2047

EK D

ANALİZLERDE KULLANILAN R KODLARI

fd, fd1, fd2: verisetleri ismi ; kod: istasyon kodları

####türe bağlı olasılık dağılımları

```
library(ggplot2)
library(lattice)
library(randomForest)
library(caret)
library(corrplot)
featurePlot(x = fd1[,2:12], y = as.factor(fd1$kod),
        plot = "density", scales = list(x=list(relation="free"), y=list(relation="free")), auto.key = T)
scatterplotMatrix(~october+march+april+may+july+september|kod, data=fd1)
```

paralel plot

```
parallelplot(~fd1[1:12], fd1, groups = fd1$kod, auto.key = F,
         ylab = "Aylık Ölçümler", lty="2",
         main ="Akım Gözlem İstasyonlarına Göre Paralel Koordinat Çizimleri")
```

####Hiyerarşik Kümeleme

```
idx <- sample(1:dim(fd1)[1],30)
fd1sample <- fd1[idx, ]
fd1sample $ kod <- NULL
hc <- hclust(dist(fd1sample), method = "ave")
plot(hc, hang = -1, labels = fd1$kod[idx])
```

####algoritmalar arası benzerlik ölçüleri ilişkileri ve dendextend kütüphanesi

```
hclust_methods <- c("ward.D", "single", "complete", "average", "mcquitty",
              "median", "centroid", "ward.D2")
library(dendextend)
fd1_dendlist <- dendlist()
for(i in seq_along(hclust_methods)) {
```

```
 hc_fd1 <- hclust(fd1, method = hclust_methods[i])
 fd1_dendlist <- dendlist(fd1_dendlist, as.dendrogram(hc_fd1), NA) }
names(fd1_dendlist) <- hclust_methods
fd1_dendlist
fd1_dendlist_cor <- cor.dendlist(fd1_dendlist)
fd1_dendlist_cor , corrplot::corrplot(fd1_dendlist_cor, "pie", "lower")

#### K-means  yöntemi
fd2$kod<-NULL
(kmeans.result <- kmeans(fd2, 3))
table(fd2$kod, kmeans.result$cluster)
#####The k-Medoids Clustering
library(fpc)
pamk.result <- pamk(fd)
pamk.result$nc  #
table(pamk.result$pamobject$clustering, fd$kod)
layout(matrix(c(1, 2), 1, 2)) # 2 graphs per page
plot(pamk.result$pamobject)
layout(matrix(1))
###plot küme merkezleri k-means devam
plot(fd2[c("ocak", "kod")], col = kmeans.result$cluster)
points(kmeans.result$centers[, c("ocak", "kod")],
col = 4:6, pch = 8, cex = 2)# plot cluster centers
####  CIRCLIZE GRAFİK
d_fd <- dist(fd2) # method="man"
hc_fd <- hclust(d_fd, method = "complete")
fd_kod <- rev(levels(fd[,5]))
library(dendextend)
dend <- as.dendrogram(hc_fd)
dend <- rotate(dend, 1:5)
# Color the branches based on the clusters:
dend <- color_branches(dend, k=3) #, groupLabels=fd_kod)
labels_colors(dend) <-
  rainbow_hcl(3)[sort_levels_values(
    as.numeric(iris[,5])[order.dendrogram(dend)]
  )]
```

```
labels(dend) <- paste(as.character(fd[,12])[order.dendrogram(dend)],
                      "(",labels(dend),")",     sep = "")
dend <- hang.dendrogram(dend,hang_height=0.1)
# reduce the size of the labels:
# dend <- assign_values_to_leaves_nodePar(dend, 0.5, "lab.cex")
dend <- set(dend, "labels_cex", 0.5)

# And plot:
par(mar = c(3,3,3,7))
plot(dend, horiz =  TRUE,  nodePar = list(cex = .007))
legend("topleft", legend = fd_kod, fill = rainbow_hcl(3))
library(circlize)
par(mar = rep(0,4))
circlize_dendrogram(dend)
```

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Mehmet Şamil GÜNEŞ

Doğum Tarihi ve Yeri :21.09.1992/ Malatya

Yabancı Dili :İngilizce

E-posta : mehmetsamilgunes@gmail.com

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Y. Lisans	İstatistik	Yıldız Teknik Üniversitesi	2017
Lisans	İstatistik	Kocatepe Üniversitesi	2014
Lise	Fen Bilimleri	Sağmalcılar A. L.	2009

YAYINLARI

Makale

1. Yildiz, D., Yildiz, D. ve Gunes M. S., (2016). “Comparison Of The Long Term Natural Streamflows Trend Of The Upper Tigris River”, Imperial Journal of Interdisciplinary Research (IJIR), Vol-2, Issue-8, Page:1174.

2. Yildiz, D., Yildiz, D. ve Gunes M. S., (2016). “ Africa In Motion: Why The Rush To Build Grand Renaissance And Inga 3 Dams In Africa?”, International Advanced Research Journal in Science, Engineering and Technology, Vol. 3, Issue 9.

3 Yildiz, D., Yildiz, D. ve Gunes M. S., (2016). “New Security Concept And Analytical-Transdisciplinary Approaches To Hydropolitics”, International Journal of Scientific & Engineering Research, Volume 7, Issue 7.

4 Yildiz, D., Yildiz, D. ve Gunes M. S., (2016). “Internationally Planned Corner Stone Of The New Nile Hydropolitics: Grand Renaissance Dam”, International Journal of Management (IJM), Volume 7, Issue 5.

Bildiri

1. Gunes, M. S., Yildiz, D., (2016). “An Interactive Visualization Approach to Hydrological Statistics with Gcluto”, Xth International Statistics Days Conference, 7-9 October, Giresun, TURKEY.

Proje

1. Türkiye’de Havzalar Bazında Su Arz Ve Talebinin Sosyal Ekonomik Ve Hidrolojik Etkilerinin Belirlenmesi, Modellenmesi Ve Gelecek Tahmini, Yıldız Teknik Üniversitesi, BAPK, Proje yürütücüsü Yrd. Doç. Dr. Doğan YILDIZ, Araştırmacı Mehmet Şamil GÜNEŞ, 2016.