

**UKUROVA NİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

YÜKSEK LİSANS TEZİ

Hasan YILDIRIM

**UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN İSTATİSTİK ALANINDAKİ
KULLANIMI**

İSTATİSTİK ANABİLİM DALI

ADANA, 2015

**ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN İSTATİSTİK ALANINDA KULLANIMI

Hasan YILDIRIM

YÜKSEK LİSANS TEZİ

İSTATİSTİK ANABİLİM DALI

Bu Tez 13/03/2015 Tarihinde Aşağıdaki Jüri Üyeleri Tarafından
Oybirliği/Oyçokluğu ile Kabul Edilmiştir.

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

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

.....
Yrd. Doç. Dr. Hüseyin GÜLER
ÜYE

Bu Tez Enstitümüz İstatistik Anabilim Dalında hazırlanmıştır.

Kod No:

**Prof. Dr. Mustafa GÖK
Enstitü Müdürü**

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

ÖZ

YÜKSEK LİSANS TEZİ

UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN İSTATİSTİK ALANINDA KULLANIMI

Hasan YILDIRIM

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

Danışman : Prof. Dr. Sadullah SAKALLIOĞLU
Yıl: 2015, Sayfa: 97
Jüri : Prof. Dr. Sadullah SAKALLIOĞLU
: Prof. Dr. Selahattin KAÇIRANLAR
: Yrd. Doç. Hüseyin GÜLER

Kümeleme analizi, doğal sınıflamaları hakkında açık bilgi bulunmayan durumlarda üzerinde çalışılan birimler topluluğunu, kendi içinde türdeş fakat diğerlerinden farklı olacak biçimde kümelere ayırma olanağı tanıyarak ana kitleye ilişkin tahminlerin yapılmasında yararlanılan çok değişkenli istatistiksel bir yöntem olarak tanımlanmaktadır (Sharma, 1996). Bu yöntemin etkinliğini belirlerken kullanılan uzaklık ölçüsü önem arz eder. Bu çalışmada da özellikle iki sonuçlu (binary) veri yapıları için uzaklık ve benzerlik ölçülerinin kümeleme analizi üzerindeki etkileri araştırılmış, geçerlilik kriteri olarak kofenetik korelasyon katsayısı kullanılmıştır. Bu amaçla literatürde tanımlanmış olan 37 benzerlik ve 10 uzaklık ölçüsü olmak üzere toplam 47 ölçüye bağlı olarak uzaklık matrisleri hem gerçek veri kümesi hem de yapay olarak üretilmiş dört farklı yapıdaki veri kümesi kullanılarak oluşturulmuş ve analizler sonuçlandırılmıştır. Ayrıca ikili uzaklık ve benzerlik ölçülerinin yapıları, ölçülerin sağlamaları önerilen birtakım teorik koşullara göre incelenmiştir. Bunun yanı sıra aşamalı kümeleme analizi yöntemleri incelenmiş, pratikte nasıl uygulanacağı gösterilmiştir. Elde edilen bulgular sonucunda Hamann, Russell&Rao, Chord ve Sokal&Sneath-1ölçüleri en iyi kümelenme yapısını ortaya koyan ölçüler olarak önerilmiştir.

Anahtar kelimeler: Kümeleme, uzaklık, benzerlik, kofenetik korelasyon

ABSTRACT

MSc. THESIS

THE USAGE OF DISSIMILARITY AND SIMILARITY MEASURES IN STATISTICS

Hasan YILDIRIM

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

Supervisor : Prof. Dr. Sadullah SAKALLIOĞLU
Year: 2015, Page: 97
Jury : Prof. Dr. Sadullah SAKALLIOĞLU
: Prof. Dr. Selahattin KAÇIRANLAR
: Asst. Prof. Dr. Hüseyin GÜLER

Cluster analysis is defined as a multivariate statistical method for grouping observations which are unknown about their classifications, into clusters such that the observations in each cluster or group are similar and the observations across groups are as different as possible and getting some estimations about populations (Sharma, 1996). Dissimilarity measures are important in determining the efficiency of this method. In this study, the effects of similarity and dissimilarity measures on cluster analysis is investigated for binary data sets and as a validity index or criteria, the cophenetic correlation coefficient is used. For this purpose, the distance matrix depending on each of 47 measures, which of 37 are similarity, 10 are dissimilarity measures, is generated by using both real data and four artificial data sets which have different features and these datasets are analyzed. Besides, structures of binary similarity and dissimilarity measures are examined in terms of some theoretical conditions which are proposed to be satisfied. In other respects hierarchical clustering analysing methods are explained and their practical applications are exemplified. As a result of the findings, Hamann, Russell&Rao, Chord and Sokal&Sneath-1 measures are suggested as the measures providing best clustering structure.

Key Words: Clustering, dissimilarity, similarity, cophenetic correlation

TEŞEKKÜR

Bu tez çalışması süresince beni destekleyen, bilgi ve görüşlerini paylaşan sayın Prof. Dr. Sadullah SAKALLIOĞLU'na, tez jüri üyeleri sayın Prof. Dr. Selahattin KAÇIRANLAR'a ve sayın Yard. Doç. Dr. Hüseyin GÜLER'e, Çukurova Üniversitesi İstatistik bölümünde görev yapan çalışma arkadaşlarıma teşekkürlerimi sunarım.

Ayrıca tüm hayatım boyunca üzerimden maddi ve manevi desteğini esirgemeyen anneme ve tüm aileme en içten şekilde teşekkürü bir borç bilirim.

İÇİNDEKİLER

SAYFA

ÖZ	I
ABSTRACT	II
TEŞEKKÜR.....	III
İÇİNDEKİLER	IV
ÇİZELGELER DİZİNİ	VI
ŞEKİLLER DİZİNİ.....	VIII
1. GİRİŞ	1
2. ÖNCEKİ ÇALIŞMALAR.....	5
3. KÜMELEME ANALİZİ.....	9
3.1. Aşamalı Kümeleme Analizi Yöntemleri	13
3.1.1. Tek Bağlantı Kümeleme Yöntemi	15
3.1.2. Tam Bağlantı Kümeleme Yöntemi	18
3.1.3. Ortalama Bağlantı Kümeleme Yöntemi	20
3.1.4. Ağırlıklı Ortalama Bağlantı Kümeleme Yöntemi.....	22
3.1.5. Merkezi Bağlantı Kümeleme Yöntemi	23
3.1.6. Medyan Bağlantı Kümeleme Yöntemi	24
3.1.7. Ward Bağlantı Kümeleme Yöntemi	25
3.1.8. Esnek Beta Yöntemi	26
3.2. Kümelerin Geçerliliği.....	27
3.2.1. Dışsal (External) İndeksler.....	28
3.2.1.1. Rand İndeksi	30
3.2.1.2. Jaccard Katsayısı.....	30
3.2.1.3. Fowlkes ve Mallows İndeksi	30
3.2.1.4. Γ İstatistiği	31
3.2.2. İçsel (Internal) İndeksler	31
3.2.2.1. Kofenetik (Cophenetic) Korelasyon Katsayısı	32
3.2.2.2. Wilks'in Lambda Test İstatistiği.....	32
3.2.3. Oransal (Relative) Kriterler	33

4. UZAKLIK VE BENZERLİK ÖLÇÜLERİ.....	35
4.1. Uzaklık ve Benzerlik Ölçülerine İlişkin Temel Kavramlar.....	36
4.2. İki Sonuçlu (Binary) İçin Kullanılan Benzerlik Ölçüleri	36
4.2.1. İkili Uzaklık ve Benzerlik Ölçülerinin Teorik Olarak İncelenmesi.....	54
4.2.2. Ölçü Seçimine İlişkin Kabul Edilebilirlik Koşulları	55
4.2.2.1. Benzerlik Ölçülerinin Teorik Olarak İncelenmesi.....	55
4.2.2.2. Uzaklık Ölçülerinin Teorik Olarak İncelenmesi	60
5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ.....	63
5.1. Aşamalı Kümeleme Analiz Yöntemlerinde İkili Uzaklık ve Benzerlik Ölçülerinin Etkisi	63
5.1.1. Tamamen Rastgele Üretilmiş Veri Kümesinin Aşamalı Kümeleme Analizi Üzerindeki Etkisinin Araştırılması	67
5.1.2. %90 oranında ‘1’ İçeren Veri Kümesinin Aşamalı Kümeleme Analizi Üzerindeki Etkisinin Araştırılması	69
5.1.3. %10 oranında ‘1’ İçeren Veri Kümesinin Aşamalı Kümeleme Analizi Üzerindeki Etkisinin Araştırılması	72
5.1.4. %50 oranında ‘1’ İçeren Veri Kümesinin Aşamalı Kümeleme Analizi Üzerindeki Etkisinin Araştırılması	74
5.1.5. Gerçek Veri Kümesinin Aşamalı Kümeleme Analizi Üzerindeki Etkisinin Araştırılması.....	76
6. SONUÇLAR VE ÖNERİLER	85
KAYNAKLAR	89
ÖZGEÇMİŞ	97

Çizelge 3.1. Eklemeli aşamalı kümelemede kullanılan bağlantı yöntemlerinin gösterimi için oluşturulan örnek veri kümesi.....	15
Çizelge 3.2. Birimler arasındaki Lance&Williams uzaklığına göre elde edilen uzaklık matrisi.....	15
Çizelge 3.3. Tek bağlantı kümeleme yöntemine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri	17
Çizelge 3.4. Tam bağlantı kümeleme yöntemine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri.....	19
Çizelge 3.5. Ortalama bağlantı kümeleme yöntemine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri.....	21
Çizelge 3.6. Ağırlıklı ortalama bağlantı kümeleme yöntemine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri.....	22
Çizelge 3.7. Lance ve Williams'ın önerdiği genel formüldeki parametrelerin bağlantı yöntemine göre değerleri.....	27
Çizelge 3.8. Şekil 3.12'daki örnek C ve P kümelenmesi için hesaplanan muhtemel dört duruma karşılık gelen değerler ve açıklamaları.....	29
Çizelge 3.9. Küme geçerliliğinin ölçülmesinde kullanılan dışsal kriterler ve değer aralıkları	31
Çizelge 4.1. İkili yapıdaki verilere ilişkin frekans tablosu.....	38
Çizelge 4.2. İkili benzerlik ölçüleri ve aralık değerleri.....	51
Çizelge 4.3. İkili uzaklık ölçüleri	54
Çizelge 4.4. Benzerlik ölçüleri için elde edilen sonuçlar	59
Çizelge 4.5. Uzaklık ölçüleri için elde edilen sonuçlar	61
Çizelge 5.1. Aşamalı kümeleme yöntemlerinde etkisi incelenen benzerlik ölçüleri.....	63
Çizelge 5.2. Aşamalı kümeleme yöntemlerinde etkisi incelenen uzaklık ölçüleri.....	65
Çizelge 5.3. Tamamen rastgele üretilmiş veri kümesi için kofenetik korelasyon katsayıları	67

Çizelge 5.4. Tamamen rastgele üretilmiş veri kümesi için bağlantı yöntemlerine göre en yüksek kofenetik korelasyon katsayı değerine sahip 5 ölçü	69
Çizelge 5.5. %90 oranında ‘1’ içeren veri kümesi için kofenetik korelasyon katsayıları	70
Çizelge 5.6. %90 oranında ‘1’ içeren veri kümesi için bağlantı yöntemlerine göre en yüksek kofenetik korelasyon katsayı değerine sahip 5 ölçü	71
Çizelge 5.7. %10 oranında ‘1’ içeren veri kümesi için kofenetik korelasyon katsayıları	72
Çizelge 5.8. %10 oranında ‘1’ içeren veri kümesi için bağlantı yöntemlerine göre en yüksek kofenetik korelasyon katsayı değerine sahip 5 ölçü	73
Çizelge 5.9. %50 oranında ‘1’ içeren veri kümesi için kofenetik korelasyon katsayıları	74
Çizelge 5.10. %50 oranında ‘1’ içeren veri kümesi için bağlantı yöntemlerine göre en yüksek kofenetik korelasyon katsayı değerine sahip 5 ölçü	75
Çizelge 5.11. Jaccard ve Dice ölçüleriyle elde edilen kofenetik korelasyon katsayıları	76
Çizelge 5.12. Bitki verisi için elde edilen kofenetik korelasyon katsayıları	76
Çizelge 5.13. Bitki verisi için bağlantı yöntemlerine göre en yüksek kofenetik korelasyon katsayı değerine sahip 5 ölçü.....	78
Çizelge 5.14. Bitki Verisine Uygulanan Aşamalı Kümeleme Analizi Sonucu Elde Edilen Kümeler ve Birimleri	83

ŞEKİLLER DİZİNİ

SAYFA

Şekil 1.1. Yıllara göre içerisinde “ikili uzaklık ve benzerlik ölçüleri” kavramını içeren çalışma sayıları	3
Şekil 3.1. 12 sayının kümelenmesinden elde edilebilecek farklı bazı kümeler.....	12
Şekil 3.2. Kümeleme analizi yöntemlerinin sınıflandırılması	12
Şekil 3.3. Aşamalı kümeleme analiz yöntemlerinin ağaç diyagramı ve Venn şeması ile gösterimi	14
Şekil 3.4. Tek bağlantı kümeleme yöntemine göre iki küme arasındaki uzaklığın belirlenmesi	16
Şekil 3.5. Tek bağlantı kümeleme yöntemi ile birimlerin bağlanmasının ağaç grafiği ile gösterimi	17
Şekil 3.6. Tam bağlantı kümeleme yöntemine göre iki küme arasındaki uzaklığın belirlenmesi	18
Şekil 3.7. Tam bağlantı kümeleme yöntemi ile birimlerin bağlanmasının ağaç grafiği ile gösterimi	19
Şekil 3.8. Ortalama bağlantı kümeleme yöntemine göre iki küme arasındaki uzaklığın belirlenmesi	20
Şekil 3.9. Ortalama bağlantı kümeleme yöntemi ile birimlerin bağlanmasının ağaç grafiği ile gösterimi.....	21
Şekil 3.10. Ağırlıklı ortalama bağlantı kümeleme yöntemi ile birimlerin bağlanmasının ağaç grafiği ile gösterimi	23
Şekil 3.11. Merkezi bağlantı kümeleme yöntemine göre iki küme arasındaki uzaklığın belirlenmesi	24
Şekil 3.12. C ve P kümelenmesi için birim çiftlerinin muhtemel dört durumu için örnek gösterim.....	29
Şekil 5.1. Tamamen rastgele veri kümesi.....	65
Şekil 5.2. %90 oranında ‘1’ içeren veri kümesi	66
Şekil 5.3. %10 oranında ‘1’ içeren veri kümesi	66

Şekil 5.4. %50 oranında ‘1’ içeren veri kümesi	66
Şekil 5.5. Bitki verisi kullanılarak ortalama bağlantı yöntemi ile elde edilen ağaç diyagramı	79

1. GİRİŞ

Uzaklık ve benzerlik ölçüleri, birçok alanda önemli bir role sahiptir. Sınıflandırma çalışmalarının verimi uygun bir ölçünün seçimine doğrudan bağlı olduğundan, birçok araştırmacı bulundukları alanda en uygun ve anlamlı benzerlik ve uzaklık ölçüsünü belirlemek amacıyla yıllardır yoğun çabalar yürütmektedirler. Bu süreçte tanımlanan ölçülerin etkinliklerinin sınanmasının yanı sıra, çok sayıda ölçü ortaya konmuş ve tanımlanmıştır.

Günümüzde hiç şüphe yoktur ki, teknolojinin gelişmesine paralel olarak artan bilimsel çalışma sayısı, devamlı veri akışını sağlamış ve beraberinde verinin içerdiği bilginin ortaya çıkarılması ihtiyacını doğurmuştur. Elbette ki karşılaşılan verilerin yapısı farklılaşmaktadır. Bu nedenle farklılaşan veri yapısına uygun farklı ölçüler tanımlamak da zorunlu hale gelmiştir.

Karşılaşılan bu zorunluluktan ötürü; sürekli, kesikli ve kategorik veriler için ayrı ayrı çok sayıda ölçü öne sürülmüş, birçok çalışma da bu ölçülerin etkinliklerini ölçme üzerine yapılmıştır. Bu çalışmalarda, özel olarak kategorik veriler, kendi içerisinde ikili (binary) ve çok seçenekli olarak iki sınıfa ayrılmış, yine bu ayrıma dayalı olarak tanımlamalar ve kullanımlar olmuştur.

Uzaklık ve benzerlik ölçülerinin doğru seçilmesi kadar bu ölçülere dayalı olarak uygulanacak yöntemler de önem arz etmektedir. Bu noktada ölçüyü kullanarak veriden bilgiye geçişi sağlamayı mümkün kılan, yoğun olarak da kullanılmakta olan istatistiksel yöntemlere başvurulur. Bu yöntemler genel olarak iki amaçla uygulanır: Birincisi, veriyi oluşturan birimlerin benzerliklerine göre gruplanması, ikincisi ise söz konusu büyük bir veri ile ilişki yapılarına bağlı olarak daha küçük boyutlu veriye ulaşılmasıdır. Bahsedilen amaçlarla uygulanan ve yoğun olarak kullanılan yöntemlere en belirgin örnek kümeleme analizidir.

Kümeleme doğal sınıflamaları hakkında açık bilgi bulunmayan durumlarda üzerinde çalışılan gözlemler topluluğunu, kendi içinde türdeş fakat diğerlerinden farklı olacak biçimde kümelere ayırma olanağı tanıyarak ana kitleye ilişkin tahminlerin yapılmasında yararlanılan çok değişkenli istatistiksel bir yöntem olarak tanımlanmaktadır (Sharma,1996).

Bahsedilen bu çok değişkenli istatistiksel yöntemin kullanım alanları oldukça geniş olup, her bir alan aynı zamanda uzaklık ve benzerlik ölçülerinin doğal kullanım alanlarıdır. Örnek olarak aşağıdaki alanlar verilebilir:

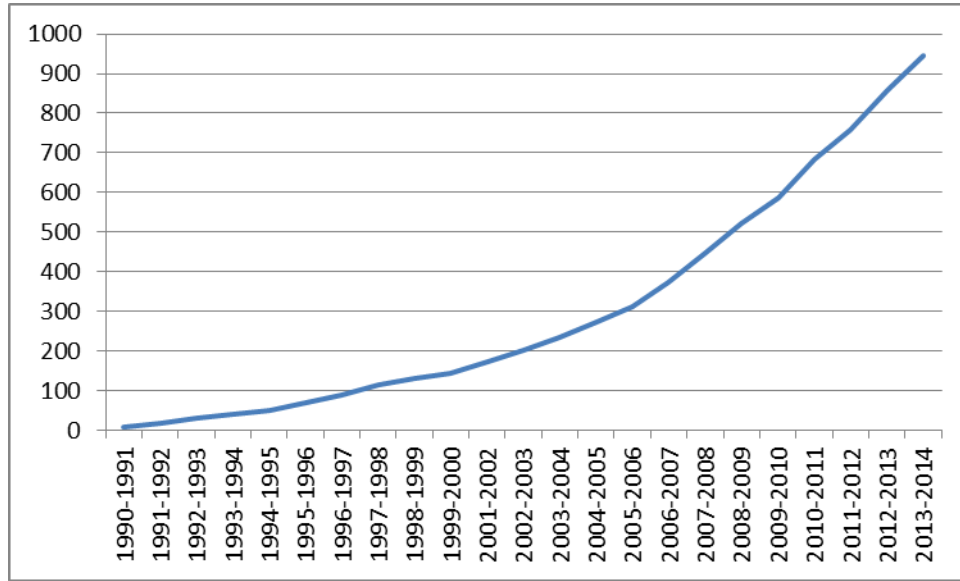
- Bilgisayar bilimleri (web-veri madenciliği),
- Yaşam ve tıp bilimleri (genetik, biyoloji, mikrobiyoloji, paleontoloji, psikiyatri, klinik, patoloji, kimya),
- Mühendislik (yapay zekâ, makine öğrenimi, görüntü tanıma, makine mühendisliği, elektrik mühendisliği)
- Astronomi ve yer bilimleri (coğrafya, jeoloji, uzaktan algılama)
- Sosyal bilimler (sosyoloji, psikoloji, arkeoloji, antropoloji, eğitim)
- Ekonomi (pazarlama, işletme, ekonometri, mekânsal veri analizi)
- İstatistik

Bu çalışmada ikili uzaklık ve benzerlik ölçülerinin etkilerini araştıracağız. İkili yapıdan kastedilen, söz konusu niteliğin iki değerden birini alabileceği yani evet-hayır, var-yok şeklinde muhtemel iki sonucun olduğu durumlardır. İki ya da daha fazla sayıda birimin çok sayıda değişken için aldıkları ikili sonuçlar arasındaki benzerlik, ikili uzaklık ve benzerlik ölçüleri yardımıyla hesaplanabilir. Bu basit anlamda ikili veri yapısında olan vektörlerin karşılaştırılmasıdır.

İkili veri yapısına dayalı olarak hesaplanan bu ölçülerin ilk örneği 1901 yılında Alpler’de bulunan bitki türlerinin kümelenmesi amacıyla Paul Jaccard tarafından ortaya atılmıştır. Jaccard, ölçüsünü herhangi iki bitki türünün tanımladığı değişkenlere göre taşıdıkları ortaklığın (uyumun), uyumsuzlıklara oranı olarak ifade etmiştir. Hala özellikle ekoloji başta olmak üzere birçok alanda yoğun olarak kullanılan ölçüdür.

Thomson Reuters tarafından hazırlanan SCI, SCI-Expanded, SSCI ve AHCI indekslerince taranan çalışmalara ulaşılabilen ISI Web of Knowledge SM (<http://apps.isiknowledge.com>) arama sayfasında arama kutucuğuna “binary similarity and dissimilarity measures ” terimi yazılıp yıllara göre arama yapılarak 1990 ile 2014 yılları arasında yıllara göre içerisinde binary similarity and

dissimilarity measures kavramı geçen çalışma sayıları Şekil 1.1’de verilmiştir. Grafikten ölçülere ilişkin çalışma sayılarının sürekli olarak arttığı görülmektedir. 1990-1991 yılları arasında çalışma sayısı 19 iken 2013-2014 yılları arasında bu sayı 1000’e yaklaşmıştır.



Şekil 1.1. Yıllara göre ikili benzerlik ve uzaklık ölçüleri kavramını içeren çalışma sayıları (ISI Web of Science)

Yapılan çalışmalarda öncelikli amaç veriye en uygun ölçüyü belirlemek olduğundan farklı veri yapılarına uygun birçok ölçü önerilmiş ve önerilmektedir. İkili yapıda olan ölçülerin kendi içinde farklı sınıflandırmaları bulunmaktadır. Bu ölçüler genel olarak üç başlık altında gruplandırılabilirler: Birincisi, nitelik temelli olarak adlandırılan, basit olarak frekanslara dayalı olarak oluşturulan ölçülerdir. İkincisi, korelasyon temelli olarak nitelendirilip, genellikle frekans sayılarının çarpımsal formlarını içeren ölçülerdir. Son olarak uzaklık ölçüleri olarak adlandırılan grupta ise çoğunlukla uyumsuzluğu temsil eden değerlerden oluşan ölçüler yer alır.

Ölçülerin yapısal farklılıkları olmasına rağmen, temelde amaç birim, gözlem ya da nesnelerin benzerliklerini veya farklılıklarını ortaya koymaktır. Hedeflenen benzerlikleri araştırmak ise kullanılan benzerlik ölçüsünün değerinin artması, benzerliğin arttığı anlamını taşıırken, tersi durumda farklılaşma derecesi artmaktadır şeklinde yorum yapılır. Amaç farklılıkları görmek ise uzaklık ölçüleri kullanılacak ve

burada küçük uzaklık ölçü değeri birbirine benzer birimleri, yüksek değer ise birbirinden farklı birimleri ifade edecektir.

Kümeleme analizinin teorisi uzaklık ve benzerlik ölçüleri kullanılarak birim veya değişkenlere ilişkin elde edilen ölçü değerlerinin oluşturmuş olduğu matrisler üzerine kuruludur. Bu nedenle ölçüler önemli bir yere sahiptir. Oldukça çok sayıda ve farklı veri tipine uygun ölçüler tanımlanmış olmasına rağmen, bahsedilen analizleri uygulayabileceğimiz yazılımlarda tanımlı olarak çok daha az sayıda ölçü bulunur. Tanımlı olan bu ölçüler ise en sık kullanılan ölçülerdir.

Bu tez çalışması kapsamında, ikili veri kümeleri için tanımlanmış uzaklık ve benzerlik ölçülerinin kümeleme analizi üzerindeki etkileri incelenmiştir. Bu amaçla kümeleme analiz yöntemlerinden aşamalı kümeleme analizi kullanılmıştır.

Tez çalışması altı bölümden oluşmaktadır. Çalışılan konunun uygulama alanları ve önemini içeren bu bölümün ardından, ikinci bölümde konu ile ilgili önceki çalışmalara yer verilmiştir.

Çalışmanın üçüncü bölümünde kümeleme analiz yöntemleri ile ilgili tanımlar, kavramlar incelenip, aşamalı kümeleme analiz yöntemleri irdelenmiş, bu yöntemlerde kullanılan kümeleme yapısı geçerliliğine ilişkin bazı indeksler verilmiştir. Dördüncü bölümde, ikili yapıdaki veriler için kullanılan uzaklık ve benzerlik ölçüleri tanımlanmış, özellikleri anlatılmış, devamında alanyazında tanımlı hemen hemen bütün ölçüler kısaca incelenmiş, geniş bir listesi verilmiştir. Yine bu bölümde teorik bir takım koşullara göre ölçülerin özellikleri araştırılmıştır. Çalışmanın beşinci bölümünde, dördüncü bölümde verilen ölçülerin kümeleme analizi üzerindeki etkileri incelenmiştir. Bu amaçla, biri gerçek dördü farklı özellikteki beş farklı veri kümesi kullanılarak ölçülere bağlı elde edilen uzaklık matrislerine aşamalı kümeleme analizi uygulanmıştır. Kullanılan uzaklık ve benzerlik ölçülerinin etkinliklerini ölçmek amacıyla küme yapı geçerlik indekslerinden kofenetik korelasyon katsayısı kullanılmıştır. Bu çerçevede elde edilen sonuçlara dayalı yorumlara yer verilmiştir. Altıncı ve son bölümde ise sonuçlar ve öneriler belirtilmiştir.

2. ÖNCEKİ ÇALIŞMALAR

Sınıflama amaçlı yapılan çalışmalarda birimler arası benzerliğin belirlenmesi oldukça önemlidir. Benzerliğin belirlenmesinde kullanılan ölçünün seçimi kilit rol oynar. Bu amaçla, ikili uzaklık ve benzerlik ölçülerinin seçimi ile ilgili alanyazında yer alan çalışmalardan bazıları aşağıda verilmiştir:

İkili uzaklık ve benzerlik ölçülerine ilişkin ilk çalışma, 1901 yılında Paul Jaccard tarafından yapılmıştır. Jaccard bu çalışmasında Alpler'i dört bölgeye ayırmış ve burada bulunan bitki türlerinin sayısını belirlemiştir. Sonrasında bitki türlerinin aynı bölgede görülme sayılarını incelemiş ve buradan hareketle hala ekolojide yoğun olarak kullanılan Jaccard benzerlik ölçüsünü tanımlamıştır. Daha sonra yoğun olarak bu alanda çalışmalar yürütülmüştür.

Forbes (1907, 1925), ekolojik olarak ilişkili türlerin kümelenmesi amacıyla bir ölçü önermiştir. 1907 yılındaki makalesinde Illinois bölgesindeki balık türlerinin dağılımını araştırmış ve incelediği 13 türün ayrı ayrı görüldükleri bölgeleri haritalandırmıştır.

Russell ve Rao (1940), yine balık türüne ait larvaların özelliklerini incelemiş ve sınıflandırmışlardır. Fakat tanımladıkları bu ölçüye göre, birimlerin kendine benzerliği 1'den küçük çıkabilmektedir.

Simpson (1943), memeli hayvanların farklı kıtalardaki dağılımını ve yapısını ortaya koyma amacıyla 8 ayrı bölgede görülen hayvanların aile, cins ve tür sayılarını tespit etmiş, benzerliklerini ortaya koymuştur.

Dice (1945), "tesadüf (çakışım) indeksi" olarak tanımladığı ölçüsünde varsayımsal olarak A ve B gibi iki türün ayrı ayrı tek başlarına ve birlikte görülme sıklıkları üzerinden, birlikte ortaya çıkma miktarlarına iki kat ağırlık vermiştir. Daha sonra, elde ettiği bu ölçünün istatistiksel olarak geçerliliğini ki-kare testi ile sınamıştır.

Tanimoto (1957), biyolojik sınıflama ve parmak izi tespiti gibi alanlarda geniş olarak kullanılan benzerlik ölçüsünü ortaya atmıştır.

Sokal ve Sneath (1963), yaptıkları çalışmada 29 organizma türünü benzerliklerine göre sınıflamış ve buna ilişkin elde ettiği dendrogram grafiğini yorumlamışlardır.

Janson ve Vegelius (1981), altı kritere göre 20 ikili benzerlik ölçüsünü teorik olarak incelemiş ve gruplamıştır. Elde edilen sonuçlara göre Ochiai, Dice ve Jaccard ölçüleri tatmin edici sonuçlar ortaya koymuş ve önerilen ölçüler arasında yer almıştır.

Hubalek (1982), 43 benzerlik ölçüsünü bir araya getirmiş, kabul edilebilirlik koşulları şeklinde nitelendirdiği noktalarda ölçüleri incelemiş, bunlardan 20'sini mantar verisi üzerinde kümeleme analizinde kullanmış ve beş küme elde etmiştir. Jaccard, Dice, Kulczynski-1, Driver&Kroeber, Pearson ve Baroni-Urbani&Buser-1 ölçüleri kabul edilebilir ölçüler olarak nitelendirilmiştir.

Baulieu (1989), 21 ikili uzaklık ölçüsünü çeşitli uzaklık aksiyomlarına göre incelemiş, her bir ölçünün hangi koşullarda bu özellikleri sağladığını göstermeye çalışmıştır. Çalışmada Eşleştirme, Sokal&Sneath-2, Rogers&Tanimoto, Sokal&Sneath-3 ölçülerinin incelenen aksiyomların tamamını sağladıkları görüldüğünden, bu ölçülerin çalışmalarda öncelikli olarak düşünülebilecek ölçüler olduğu söylenebilir.

Tubbs (1989), şablon eşleme sorununa ilişkin yaygın yedi ölçüyü incelemiştir.

Jackson ve arkadaşları (1989), 25 balık türünü içeren ekolojik veri kümesinde en iyi benzerlik ölçüsünü bulabilmek için 8 ölçüyü karşılaştırmıştır.

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

Zhang ve Srihari (2003), Tubbs'ın incelemiş olduğu yedi ölçüyü el yazısını tanıma yeterlilikleri açısından karşılaştırmıştır.

Willett (2003), ikili parmak izi verisi için 13 benzerlik ölçüsünün değerlerini hesaplamıştır.

Cha ve ark. (2006), karşılaştırmalı çalışmalarda performansa dayalı sınıflamayı geliştirmek için ağırlıklandırılmış ikili ölçümü ortaya atmıştır.

Erişoğlu (2011), uzaklık fonksiyonlarının kümeleme analiz yöntemleri üzerindeki etkilerini incelemiştir. Kofenetik korelasyon katsayısı ve Wilks'in lambda test istatistiği değerlerine göre örneklem hacminin düşük olduğu ($n < 75$) durumda tüm aşamalı kümeleme yöntemlerinde Mahalanobis uzaklık ölçüsünü en başarılı ölçü olduğunu ortaya koymuştur.

İkili benzerlik ve uzaklık ölçüleri, özellikle kümeleme analizinde yoğun olarak kullanılmıştır. Bu çalışmada da aşamalı kümeleme analizi yöntemi kullanıldığından, bu alanda ortaya konan ve yararlanılan çalışmalar aşağıda verilmiştir:

Veri kümeleme analizi kavram olarak ilk kez Tryon (1939) tarafından kullanılmıştır. Clements (1954), antropolojik verinin kümelenmesi çalışmasında veri kümeleme analizi kavramını kullanmıştır. Cox (1957) ve Fisher (1958), kümeleme analizi kavramını gruplama adını verdikleri çalışmalarında kullanmışlardır. Sokal ve Sneath (1963), sayısal sınıflandırmanın prensipleri adını verdikleri kitaplarında kümeleme analizini anlatmışlardır (Erişoğlu, 2011).

Sorensen (1948), "oran benzerliği" olarak isimlendirdiği ikili benzerlik ölçüsünü tanımladığı çalışmasında, benzer birimlerin birleştirilmesi ile oluşturulan yeni kümeye olan uzaklıkların hesaplanmasında maksimum uzaklığın kullanıldığı tam bağlantı kümeleme yöntemini önermiştir.

Sneath (1957) tarafından birimlerin sınıflanması için tek bağlantı kümeleme yöntemi önerilmiştir.

Sokal ve Michener (1958) basit benzerlik ölçüsü olarak adlandırılan, negatif ve pozitif tam uyumların, toplam frekansa oranlanması olarak ifade edilecek kendi adlarıyla anılan ölçüyü tanımlamış aynı zamanda ortalama bağlantı ve merkez bağlantı kümeleme yöntemlerini önermişlerdir.

Sokal ve Rohlf (1962), aşamalı kümeleme yöntemlerinin geçerliliği için kofenetik korelasyon katsayısını önermişlerdir. Kofenetik korelasyon katsayısı, başlangıçta hesaplanan benzerlik matrisi D ile kümeleme sonrası bağlantı uzaklıklarını içeren uzaklıklar matrisinin R'nin karşılıklı elemanları arasındaki

ilişkiyi ölçer. Kofenetik korelasyon katsayısı 1'e ne kadar yakınsa kümelemenin geçerliliği o kadar yüksektir.

Gower (1966), aşamalı kümeleme analizi yöntemleri içerisinde yer alan medyan bağlantı yöntemini önermiştir.

Lance ve Williams (1967), kümelemede birimlerin yeni oluşturulan kümeye olan uzaklıklarının hesaplanmasındaki farklılıklara göre oluşturulan yöntemler arasındaki bağıntıyı α , β ve γ parametrelerini kullanarak açıklamışlardır.

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

Hartigan ve Wong'un (1979) k – ortalamalar yöntemi ile ilgili çalışması yöntemin gelişimine önemli katkılar sağlamıştır. Günümüzden yaklaşık 50-55 yıl öncesi önerilen k – ortalamalar algoritması basitlik, uygulanabilirlik, deneysel çalışmalardaki başarısı ve etkinliği nedeniyle hala en çok tercih edilen kümeleme yöntemidir.

Hands ve Everitt (1987), küme üyeliklerini bilerek ürettikleri rassal veri kümelerini kullanarak beş farklı aşamalı kümeleme yöntemini karşılaştırmışlardır. Bu çalışmada tek bağlantı, tam bağlantı, ortalama, merkez ve Ward kümeleme yöntemleri iki farklı kritere göre karşılaştırılmıştır. Çalışma sonucunda bu beş yöntemin çoğu kez benzer kümeleme performansı gösterdiğini, buna karşın bu yöntemler içerisinde tek bağlantı kümeleme yönteminin diğerlerine göre daha kötü performans sergilediğini, özellikle kümelerde yer alan birim sayılarının yaklaşık eşit olması durumunda Ward yöntemi ile daha iyi sonuçlar elde edildiğini belirtmişlerdir.

3. KÜMELEME ANALİZİ

Kümeleme analizi, öncelikli amacı nesneleri ya da birimleri taşıdıkları karakteristiklere- niteliklere göre gruplamayı amaçlayan çok değişkenli yöntemler bütünüdür. Kümelemede her bir küme içerisindeki nesnelerin ya da birimlerin benzer, kümelerin ise birbirinden farklı olması amaçlanır. Yani oluşturulan kümelere ilişkin, kendi içlerinde olabildiğince homojen, kümeler arası olabildiğince heterojen olan en uygun yapı bulunmaya çalışılır. Böylece, eğer elde edilen yapı (sınıflandırma) başarılı ise her bir küme içindeki nesnelerin birbirlerine geometrik olarak oldukça yakın, farklı kümelerden oldukça uzak (ayrı) olması beklenir.

Kümeleme analizi sayısal sınıflama ve yüzey tespiti ile ilişkilidir. Kümeleme analizi yöntemleri antropoloji, arkeoloji, biyoloji, coğrafya, ekonomi, kriminoloji, mühendislik, pazar araştırmaları, psikiyatri, sosyoloji, tıp ve uzaktan algılama gibi birçok alanda yoğun olarak uygulanmaktadır. Çok çeşitli alanlarda, farklı amaçlarla kullanılmasından ötürü kümeleme analizi için tek bir şekilde tanım yapmak zordur. Benzer şekilde küme kavramı da farklı şekillerde ifade edilebilir, fakat en genel haliyle;

- Küme, içerisindeki herhangi iki nokta arası uzaklığın, yine küme içerisindeki herhangi bir noktanın, küme dışındaki herhangi bir noktaya olan uzaklığından küçük olduğu bir test uzayındaki noktaların bütünüdür

şeklinde tanımlanabilir (Xu ve Wunsch, 2009). Kümeleme analizine ilişkin bazı tanımlar ise şöyledir:

- Kümeleme analizi, her bir kümedeki birim ya da gözlemler benzer fakat kümeler birbirinden farklı olacak şekilde, araştırmacı için de anlamlı en iyi gruplamayı bulmayı amaçlayan bir yöntemdir (Rencher, 2002).

- Kümeleme analizi, öncelikli amacı birimleri sahip oldukları niteliklere dayanarak gruplamak olan çok değişkenli yöntemler bütünüdür (Hair ve Black, 2009).
- Kümeleme analizi, aynı gruptaki gözlemler benzer, farklı gruplarda yer alan gözlemler farklı olacak şekilde, bilinmeyen sayıda grup veya küme yapısına sahip belirli sayıda gözlem grubunu parçalamayı – sınıflamayı amaçlayan bir yöntemdir (Timm, 2002).
- Kümeleme analizi, birim veya değişkenlerin doğal grup yapılarını keşfetmeyi amaçlayan bir yöntemdir (Johnson ve Wichern, 2014).
- Alpar (2011) kümeleme analizini ham (gruplanmamış) X veri matrisindeki birimleri (bazen de değişkenleri) sahip oldukları özellikler çerçevesinde kümelemek (homojen alt gruplara ayırmak) amacıyla geliştirilmiş yöntemler topluluğu olarak tanımlamaktadır.
- Kümeleme analizi küme yapısı hakkında bir ön bilgiye sahip olunmayan birimlerin p değişken bakımından ölçülmesi ile oluşan $n \times p$ boyutlu X veri matrisindeki farklı birimleri farklı kümelerde, benzer birimleri aynı kümede bir araya getirerek veri seti içerisindeki gizli kümeleri ortaya çıkaran ve bu kümelerin özelliklerini belirleyerek kitle hakkında tahminlerin yapılmasına olanak sağlayan yöntemlerden oluşan çok değişkenli istatistiksel bir analizdir (Erişoğlu, 2011).

Kümeleme analizinin amaçlarına ilişkin ise farklı sınıflamalar yapılabilir. Aldenderfer ve Blashfield (1984), kümeleme analizinin dört temel hedefe ulaşmada yararlı olduğunu belirtmektedirler:

1. Bir sınıflama geliştirmek,
2. Birimleri gruplama amacıyla kavramsal sınıflama cetvelleri geliştirmek,
3. Veri keşfi yoluyla hipotez üretmek,
4. Hipotez test etmek veya bir veri setinde başka yöntemlerle yapılmış olan gruplamaların varlığını belirlemek.

Özdamar (2013)'a göre ise kümeleme analizinin amaçları şöyledir:

1. n sayıda bireyi p değişkene ilişkin özelliklerine göre, kendi içinde olabildiğince homojen ve kendi aralarında da heterojen alt kümelere ayırmak,
2. p sayıda değişkeni, n sayıda bireyde saptanan değerlere göre ortak özellikleri açıkladığı varsayılan alt kümelere ayırmak ve ortak faktör yapılarını ortaya koymak,
3. Hem bireyleri, hem de değişkenleri birlikte ele alarak n bireyi p değişkene göre ortak özellikli alt kümelere ayırmak,
4. p değişkene göre saptanan değerler bakımından bireylerin biyolojik ve taksonomik sınıflamasını ortaya koymak.

Everitt ve arkadaşları (2011) ise kümeleme analizinin genel amaçlarını,

1. Gerçek tipleri (cinsleri-ırkları) belirlemek,
2. Gruplar için ön tahmin yapmak,
3. Hipotezlerin testi,
4. Veri yapısını netleştirmek – belirginleştirmek,
5. Veri indirgemek (daha küçük boyutlu kümelerin değerlendirilmesi),
6. Aykırı değerlerin tespiti,

şeklinde sıralamıştır.

Bir veri kümesindeki özellikle büyük veri kümelerinde, muhtemel tüm kümelere ulaşmak genellikle mümkün değildir. Birim sayısı n olan bir veri kümesini g alt kümeye ayırmak için mümkün parçalanmaların sayısı,

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

ile hesaplanabilir (Duran ve Odell, 1974; Rencher, 2002).

Örnek olarak 12 sayı değerini ve bunlara bağlı elde edilebilecek mümkün küme durumlarını inceleyelim. Karşılaştırılması muhtemel kümelerden bazıları Şekil 3.1’de verilmiştir:

1	2	3
4	5	6
7	8	9
10	11	12

1	2
3	4
5	6
7	8
9	10
11	12

2	1
3	4
5	6
7	8
11	9
	10
	12

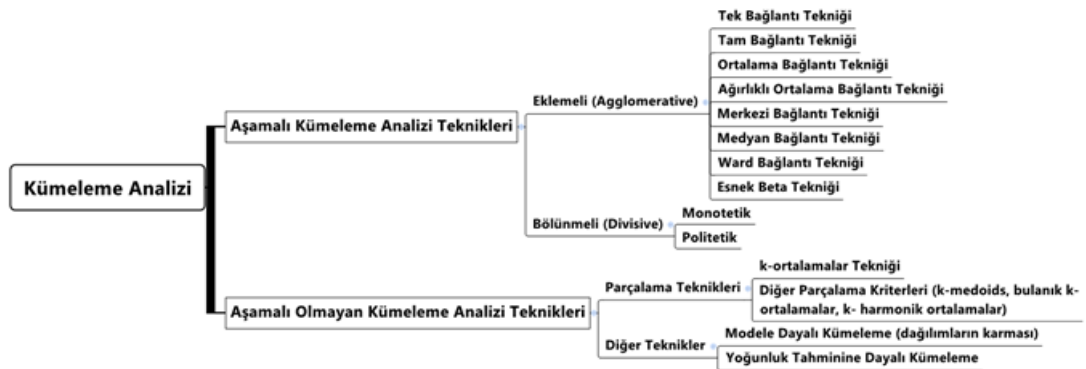
(a) Ardışık sayılar

(b) Tek ve Çift Sayılar

(c) Asal ve Asal Olmayan Sayılar

Şekil 3.1. 12 sayının kümelenmesinden elde edilebilecek farklı bazı kümeler

Örnekteki 12 sayıyı tek bir kümede toplamının 1 yolu varken, iki kümeye ayırmak için 2047 farklı yol vardır. 12 sayıyı 3 kümeye ayırmak için ise 86526 farklı yol vardır. Sadece 12 birimi 3 kümeye parçalamada bile olası tüm yapıları oluşturup, en iyisini seçmek oldukça zor olduğundan en iyi kümelenme yapısını elde edebilmek için algoritmalara ihtiyaç vardır. Bu amaçla alanyazında kümeleme analizinde kullanılan çok sayıda algoritma öne sürülmüştür. Bu algoritmaları genel olarak iki temel başlık altında toplamak mümkündür. Bunlardan biri “aşamalı (hierarchical) kümeleme yöntemleri”, diğeri ise “aşamalı olmayan (nonhierarchical) kümeleme yöntemleridir” (Everitt ve ark, 2011; Büyüköztürk ve ark., 2012).



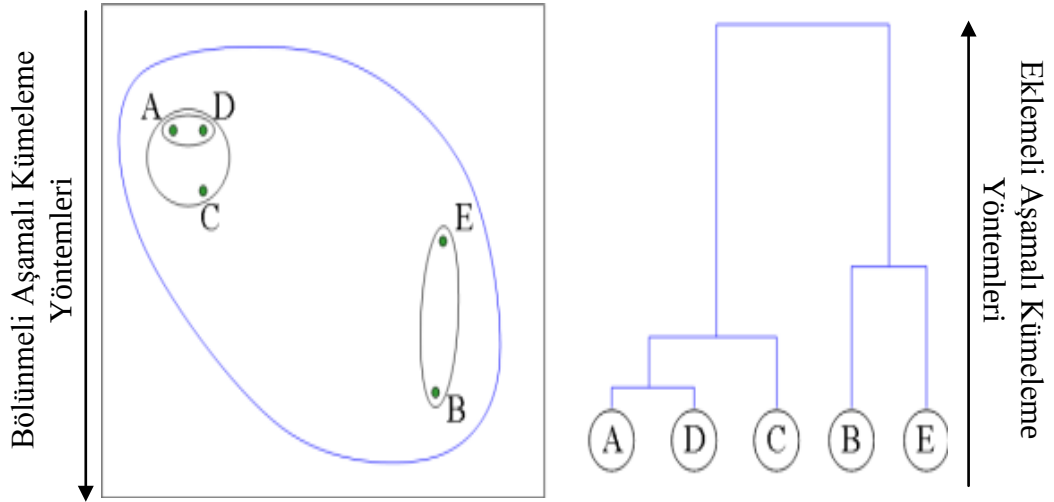
Şekil 3.2. Kümeleme analizi yöntemlerinin sınıflandırılması

3.1. Aşamalı Kümeleme Analizi Yöntemleri

Aşamalı kümeleme yöntemlerinde, veri tek bir adımda belirli sayıda sınıfa ya da kümeye parçalanmaz, bir dizi süreçten oluşur. Bu süreç birim sayısı n iken $n-1$ kararı içerir. Tüm birimleri içeren tek bir kümeyle süreç yürütülebileceği gibi, her biri birer birim içeren n kümeyle de süreç başlayabilir. Bu işleyiş farklılığından hareketle, aşamalı kümeleme analiz yöntemleri kendi içerisinde eklemeli aşamalı kümeleme analizi ve bölünmeli aşamalı kümeleme analizi olarak ikiye ayrılır. Eklemeli (agglomerative) kümeleme analizinde uygulanan süreç şöyledir;

Başlangıçta, verideki her bir birim birer küme olarak düşünülür. Kullanılan uzaklık ölçüsüne ve bağlantı yöntemine bağlı olarak en benzer iki birim birleştirilir. Her bir aşamada uzaklık matrisi yenilerek işlem tekrarlanır. Tüm birimler tek bir küme oluşturduğunda süreç sonlandırılır (Hair ve Black, 2009).

Bölünmeli (divisive) aşamalı kümeleme analizi yönteminde ise, eklemeli kümeleme analiz yönteminin tersi bir yaklaşım vardır. Tüm gözlemler tek bir küme olarak ele alınıp, ilk aşamada birbirinden farklı iki alt kümeye parçalanır. Her bir adımda k birim sayısı olmak üzere $2^{k-1} - 1$ mümkün parçalanma vardır (Everitt ve ark., 2011). Daha sonra bu parçalanmalar kendi içinde birbirinden olabildiğince farklı alt kümelere ayrılır. Süreç her bir birim birer kümeyi temsil edinceye kadar devam eder (Johnson ve Wichern, 2014). Ayrıca bu yöntem parçalanma sırasında dikkate alınan değişken sayısı bakımından iki alt başlıkta incelenebilir. Bunlar; bir kümenin iki alt kümeye yalnız tek bir değişkenin dikkate alınmasıyla bölündüğü monotetik bölünmeli kümeleme yöntemi ve birden fazla değişkenin göz önüne alınarak uygulanan politetik bölünmeli kümeleme analiz yöntemidir (Everitt ve ark. , 2011).



Şekil 3.3. Aşamalı kümeleme analiz yöntemlerinin ağaç diyagramı ve Venn şeması ile gösterimi (Seo, 2005)

Eklemeli aşamalı kümeleme analizi yedi farklı bağlantı yöntemine göre uygulanmaktadır. Bu yöntemler kümelerin oluşturulması sırasında bağlantı uzaklıklarının seçimi noktasında farklılık göstermektedir. Bu yöntemlerin pratikte nasıl uygulandığını göstermek amacıyla Çizelge 3.1’de verilen örnek veri kümesi kullanılacaktır.

Kullanılacak veri kümesi ikili yapıda olup, yine bu bölümde anlatılan ve sürekli verilerde kullanılabilen merkezi, medyan ve Ward yöntemlerinin uygulamasına değinilmeyecektir. Söz konusu üç yöntemin uygulanabilmesi için elde edilen her bir değerin geometrik olan yoruma uygun olması beklenir. Dolayısıyla küme merkezlerine dayalı tanımlanan bu üç yöntem, veri kümesi ikili olduğunda her ne kadar birtakım değerler üretse de sonuçlar anlamlı ve doğru değildir.

Çizelge 3.2.’deki uzaklık değerleri Lance&Williams uzaklık ölçüsü kullanılarak hesaplanmıştır. Bir sonraki bölümde anlatılan bu ölçünün tanımı Çizelge 4.3.’den görülebilir.

Çizelge 3.1. Eklemeli aşamalı kümelemede kullanılan bağlantı yöntemlerinin gösterimi için oluşturulan örnek veri kümesi

Birim No / Değişken	x	y	z	t
1	1	0	1	0
2	0	1	0	1
3	1	1	1	0
4	1	0	0	1
5	0	0	1	0

Çizelge 3.2. Birimler arasındaki Lance&Williams uzaklığına göre elde edilen uzaklık matrisi

Birim No	1	2	3	4	5
1	0	1	0,2	0,5	0,3333
2	1	0	0,6	0,5	1
3	0,2	0,6	0	0,6	0,5
4	0,5	0,5	0,6	0	1
5	0,3333	1	0,5	1	0

3.1.1. Tek Bağlantı Kümeleme Yöntemi (Single Linkage Method)

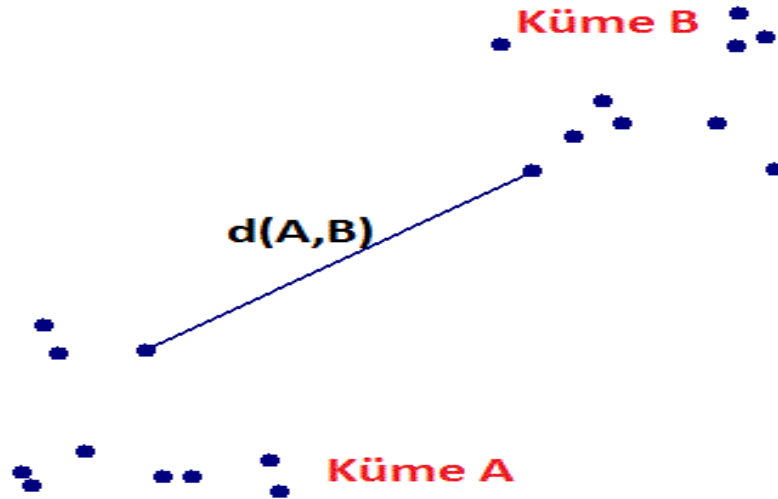
En basit aşamalı kümeleme yöntemidir. “En yakın komşuluk” olarak da bilinir. İlk olarak Florek ve arkadaşları (1951) tarafından ortaya konmuştur. Daha sonra Sneath(1957) ve Johnson (1967) ayrı olarak bu yöntemi çalışmalarında tanımlamışlardır (Everitt, 2011). Bu yöntemde kümeler birbirine en yakın komşuluğa sahip birimlerin birleştirilmesiyle şekillenir. Burada “en yakın komşuluk” uzaklık bakımından en yakın olmayı ifade eder. Tek bağlantı yöntemi uygulanırken, ilk aşamada birbirine en yakın iki birim birleştirilir ve bir küme elde edilir. Daha sonra bu kümenin diğer birim veya kümelerle uzaklığı hesaplanırken mümkün tüm uzaklıklardan en küçük değere sahip olan birim veya kümeler birleştirilir. Süreç bu şekilde devam ettirilir (Johnson ve Wichern, 2014).

Tek bağlantı kümeleme yöntemi göreceli olarak uzak olan iki küme arasında “orta” değerlerin olması durumunda var olan orta değerleri ayrı bir kümede toplamak yerine var olan bir küme içerisine katma eğilimi anlamına gelen “zincirleme etkisine” sahip olduğu için uzamış kümeleri ortaya çıkarma eğilimindedir (Everitt ve ark, 2011; Alpar, 2011). Tamamen farklı özelliklere sahip iki küme bu sorunun varlığı sebebiyle birleştirilebilir. Bu nedenle kümeler birbirinden uzak ayrılmış ise

tek bağlantı kümeleme yöntemi iyi sonuç vermektedir (Xu ve Wunsch, 2009). Tek bağlantı kümeleme yönteminde c_1 kümesinin c_i ve c_j kümelerinin birleşimini ifade eden (c_i, c_j) kümesine olan uzaklığı,

$$d(c_1, (c_i, c_j)) = \text{enküçük}(d(c_1, c_i), d(c_1, c_j)) \quad (3.2.)$$

eşitliği ile hesaplanır (Xu ve Wunsch, 2009). Şekil 3.4’de kümeler arasındaki uzaklıkların tek bağlantı kümeleme yöntemine göre belirlenmesi grafiksel olarak gösterilmiştir.



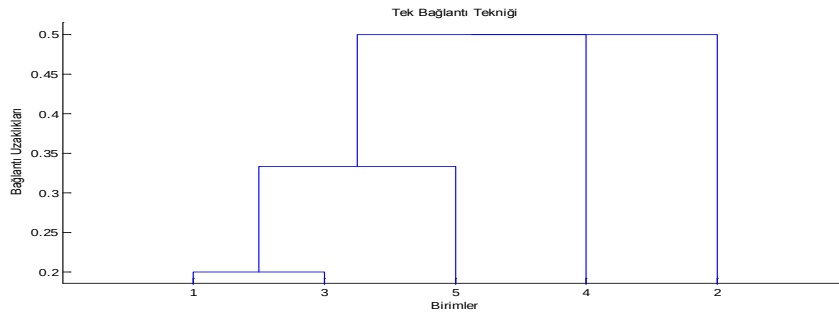
Şekil 3.4. Tek bağlantı kümeleme yöntemine göre iki küme arasındaki uzaklığın belirlenmesi

Çizelge 3.1’de verilen örnek veri kümesindeki birimlerin tek bağlantı kümeleme yöntemine göre bağlanma aşamalarında elde edilen uzaklık matrisleri Çizelge 3.3’de verilmiştir.

Çizelge 3.3. Tek bağlantı kümeleme yöntemine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri

1.Adım					
	1	2	3	4	5
1	0	1	0,2	0,5	0,3333
2	1	0	0,6	0,5	1
3	0,2	0,6	0	0,6	0,5
4	0,5	0,5	0,6	0	1
5	0,3333	1	0,5	1	0
2.Adım					
	{1,3}	2	4	5	
{1,3}	0	0,6	0,5	0,3333	
2	0,6	0	0,5	1	
4	0,5	0,5	0	1	
5	0,3333	1	1	0	
3.Adım					
	{1,3,5}	2	4		
{1,3,5}	0	0,6	0,5		
2	0,6	0	0,5		
4	0,5	0,5	0		
4.Adım					
	{1,3,5,4}	2			
{1,3,5,4}	0	0,5			
2	0,5	0			

Çizelge 3.3 incelendiğinde ilk adımda en kısa uzaklığa sahip 1. ve 3. birimler $d=0.2$ uzaklık değerinde bir küme oluşturmuştur. Benzer şekilde 3. adımda $\{1,3,5\}$ kümesi ile 4. birim arasındaki uzaklık olarak $d(\{1,3\},4)=0.5$ ile $d(5,4)=1$ uzaklıklarından küçük olan 0.5 değeri alınmıştır ve 4. birim ile $\{1,3,5\}$ kümesi birleştirilerek yeni küme oluşturulmuştur. Elde edilen kümeler ve bağlantı uzaklıkları Şekil 3.5'te verilmiştir.



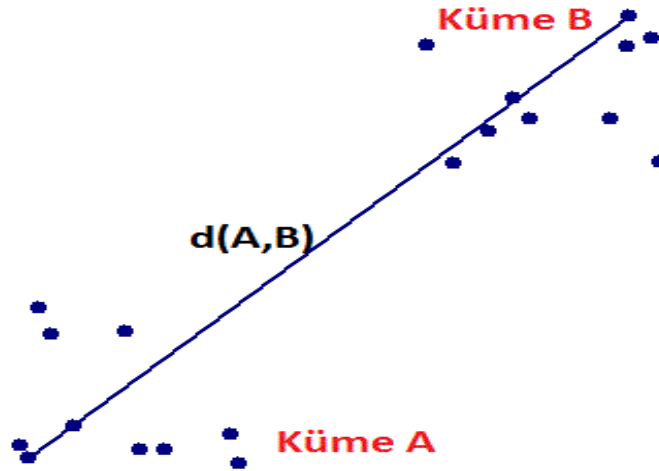
Şekil 3.5. Tek bağlantı kümeleme yöntemi ile birimlerin bağlanmasının ağaç grafiği ile gösterimi

3.1.2. Tam Bağlantı Kümeleme Yöntemi (Complete Linkage Method)

Tam bağlantı kümeleme yöntemi, her bir aşamada her biri ayrı kümelerde bulunan en uzak iki birim arası mesafeyi iki küme arası mesafe olarak almasının dışında tek bağlantı yöntemine benzer şekilde uygulanır. Tam bağlantı kümeleme yönteminde c_1 kümesinin c_i ve c_j kümelerinin birleşimini ifade eden (c_i, c_j) kümesine olan uzaklığı,

$$d(c_1, (c_i, c_j)) = \text{enbüyük}(d(c_1, c_i), d(c_1, c_j)) \quad (3.3.)$$

eşitliği ile hesaplanır. Şekil 3.6'da kümeler arasındaki uzaklıkların tam bağlantı kümeleme yöntemine göre belirlenmesi grafiksel olarak gösterilmiştir.



Şekil 3.6. Tam bağlantı kümeleme yöntemine göre iki küme arasındaki uzaklığın belirlenmesi

Tam bağlantı yöntemi, tek bağlantı yöntemindeki zincirleme etkisini ortadan kaldırır, daha küçük ve kompakt kümeler oluşturmada etkilidir (Xu ve Wunsch, 2009). Çizelge 3.1'deki örnek veri kümesindeki birimlerin tam bağlantı kümeleme yöntemine göre bağlanma aşamalarında elde edilen uzaklık matrisleri Çizelge 3.4'de verilmiştir.

Çizelge 3.4. Tam bağlantı kümeleme yöntemine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri

1.Adım					
	1	2	3	4	5
1	0	1	0,2	0,5	0,3333
2	1	0	0,6	0,5	1
3	0,2	0,6	0	0,6	0,5
4	0,5	0,5	0,6	0	1
5	0,3333	1	0,5	1	0
2.Adım					
	{1,3}	2	4	5	
{1,3}	0	1	0,6	0,5	
2	1	0	0,5	1	
4	0,6	0,5	0	1	
5	0,5	1	1	0	
3.Adım					
	{1,3,5}	2	4		
{1,3,5}	0	1	1		
2	1	0	0,5		
4	1	0,5	0		
4.Adım					
	{1,3,5}	{2,4}			
{1,3,5}	0	1			
{2,4}	1	0			

Tam bağlantı yönteminde 2.adımda {1,3} kümesi ile 5. birim arasındaki uzaklık değeri bulunurken $d(1,5)$ ile $d(3,5)$ değerleri arasından büyük olan 0,5 değeri seçilmiştir ve bu birimler birleştirilerek yeni küme elde edilmiştir. Bu yöntemin uygulanması sırasında oluşan kümeler ve birleşme değerleri Şekil 3.7’de verilmiştir.



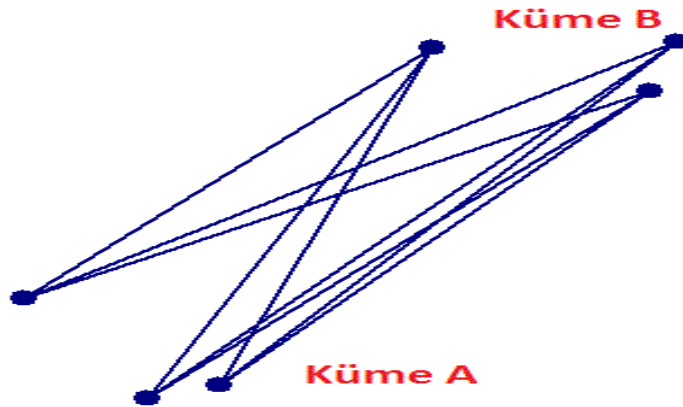
Şekil 3.7. Tam bağlantı kümeleme yöntemi ile birimlerin bağlanmasının ağaç grafiği ile gösterimi

3.1.3. Ortalama Bağlantı Kümeleme Yöntemi (Average Linkage Method)

Ortalama bağlantı kümeleme yönteminde iki küme arası uzaklık olarak, her biri ayrı kümelerde bulunan mümkün tüm ikililer arasındaki uzaklıkların ortalaması alınır (Johnson, 2014). Bu yöntem “ağırlıksız grup çiftleri yöntemi” olarak da bilinir (Everitt ve ark., 2011). Tek bağlantı yönteminde birbirine en yakın, tam bağlantı yönteminde birbirine en uzak komşulardan başlanarak kümeleme yapıldığından, bu iki yöntem uç değerlerin etkisine oldukça açıktır. Ancak ortalama bağlantı yöntemi, uçlarda yer alan gözlem çiftleri yerine tüm olası çiftler arasından uzaklıkları dikkate aldığından uç (aşırı) değerlerden daha az etkilenir (Büyüköztürk ve ark., 2012; Alpar, 2011). Ortalama bağlantı kümeleme yönteminde c_1 kümesinin c_i ve c_j kümelerinin birleşimini ifade eden (c_i, c_j) kümesine olan uzaklığı,

$$d(c_1, (c_i, c_j)) = \frac{1}{2} (d(c_1, c_i) + d(c_1, c_j)) \quad (3.4.)$$

eşitliği ile tanımlanır. Bu yöntem küçük küme içi değişkenliğe sahip kümeleri birleştirme eğilimindedir (Alpar, 2011). Şekil 3.8’de kümeler arasındaki uzaklıkların ortalama bağlantı kümeleme yöntemine göre belirlenmesi grafiksel olarak gösterilmiştir.

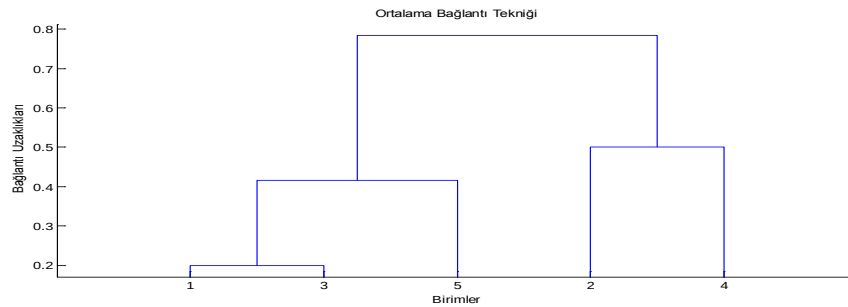


Şekil 3.8. Ortalama bağlantı kümeleme yöntemine göre iki küme arasındaki uzaklığın belirlenmesi

Çizelge 3.5. Ortalama bağlantı kümeleme yöntemine göre birimlerin bağlanmaları ve her aşamada elde edilen uzaklık matrisleri

1.Adım					
	1	2	3	4	5
1	0	1	0,2	0,5	0,3333
2	1	0	0,6	0,5	1
3	0,2	0,6	0	0,6	0,5
4	0,5	0,5	0,6	0	1
5	0,3333	1	0,5	1	0
2.Adım					
	{1,3}	2	4	5	
{1,3}	0	0,8	0,55	0,4166	
2	0,8	0	0,5	1	
4	0,55	0,5	0	1	
5	0,4166	1	1	0	
3.Adım					
	{1,3,5}	2	4		
{1,3,5}	0	0,867	0,7		
2	0,867	0	0,5		
4	0,7	0,5	0		
4.Adım					
	{1,3,5}	{2,4}			
{1,3,5}	0	0,783			
{2,4}	0,783	0			

Çizelge 3.5 incelendiğinde 2.adımda {1,3} kümesi ile 5. birim arasındaki uzaklık değeri $[d(1,5)+d(3,5)]/(2 \times 1)$ olacak şekilde hesaplanarak 0,4166 bulunmuştur. 4.adımda ise {1,3,5} kümesi ile {2,4} kümesi $[d(1,3)+d(1,4)+d(3,2)+d(3,4)+d(5,2)+d(5,4)]/(3 \times 2)$ değeri yani 0,783 uzaklığında birleştirilerek tek bir küme elde edilmiş ve süreç sonlandırılmıştır. Bu adımlara ilişkin sonuçlar Şekil 3.9’da verilmiştir.



Şekil 3.9. Ortalama bağlantı kümeleme yöntemi ile birimlerin bağlanmasının ağaç grafiği ile gösterimi

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

Aritmetik ortalama ile ağırlıklı çift grup yöntemi olarak da bilinen bu yöntem ortalama bağlantı yöntemiyle benzer olmakla beraber, bu yöntem uygulanırken elde edilen her yeni kümelenme ile uzaklık hesaplanırken kümelerin sahip oldukları birim sayıları kullanılarak ağırlıklandırma yapılır. (Xu ve Wunsch, 2009; Jain ve Dubes, 1988). Dolayısıyla daha fazla birime sahip olan kümenin uzaklığa katkısı daha fazla olacaktır. Ağırlıklı ortalama bağlantı kümeleme yönteminde c_i kümesinin c_i ve c_j kümelerinin birleşimini ifade eden (c_i, c_j) kümesine olan uzaklığı,

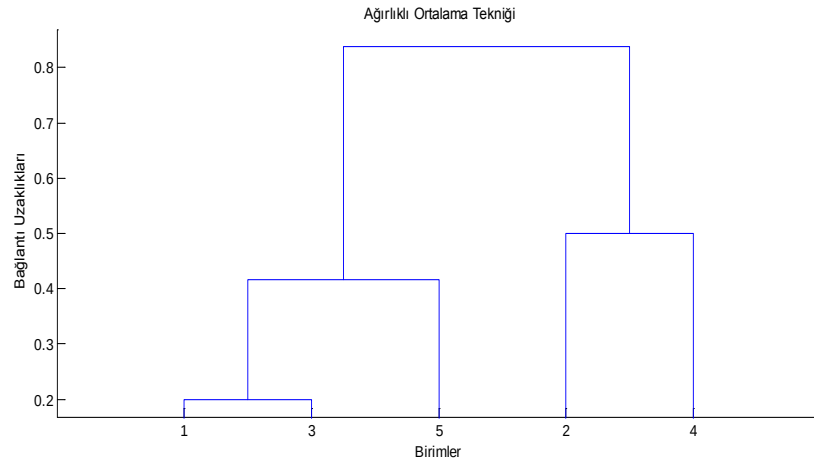
$$d(c_i, (c_i, c_j)) = \frac{n_i}{n_i + n_j} d(c_i, c_i) + \frac{n_j}{n_i + n_j} d(c_i, c_j) \quad (3.5.)$$

eşitliği ile hesaplanır (Xu ve Wunsch, 2009).

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

1.Adım					
	1	2	3	4	5
1	0	1	0,2	0,5	0,3333
2	1	0	0,6	0,5	1
3	0,2	0,6	0	0,6	0,5
4	0,5	0,5	0,6	0	1
5	0,3333	1	0,5	1	0
2.Adım					
	{1,3}	2	4	5	
{1,3}	0	0,8	0,55	0,4166	
2	0,8	0	0,5	1	
4	0,55	0,5	0	1	
5	0,4166	1	1	0	
3.Adım					
	{1,3,5}	2	4		
{1,3,5}	0	0,9	0,775		
2	0,9	0	0,5		
4	0,775	0,5	0		
4.Adım					
	{1,3,5}	{2,4}			
{1,3,5}	0	0,8375			
{2,4}	0,8375	0			

Ağırlıklı ortalama bağlantı yönteminin uygulama adımlarını gösteren Çizelge 3.6'ya göre 4.adımda $\{1,3,5\}$ kümesinin $\{2,4\}$ kümesine olan uzaklığı $[d(\{1,3,5\},2)+d(\{1,3,5\},4)]/2$ den 0,8375 olarak elde edilmiş ve bu uzaklık değerinde de kümeler birleştirilmiştir. Ağırlıklı ortalama bağlantı yönteminin uygulanması sonucu oluşan kümeler ve birleşme uzaklıkları Şekil 3.10'da verilmiştir.



Şekil 3.10. Ağırlıklı ortalama bağlantı kümeleme yöntemi ile birimlerin bağlanmasının ağaç grafiği ile gösterimi

3.1.5. Merkezi Bağlantı Kümeleme Yöntemi (Centroid Linkage Method)

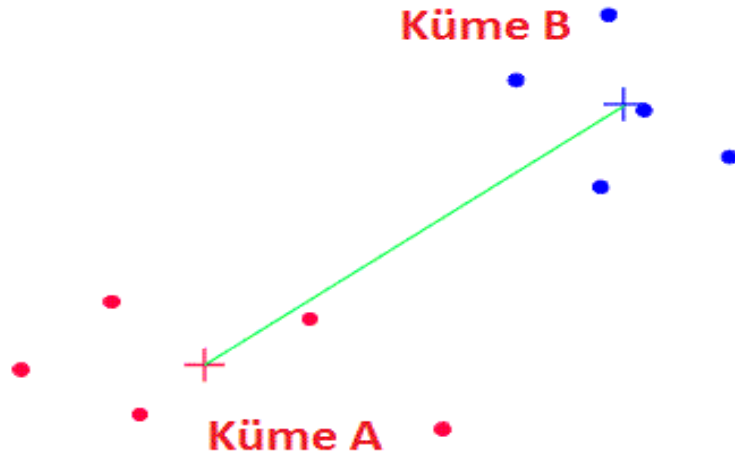
Aşırı değerlerden en az etkilenen aşamalı kümeleme yöntemi olan merkezi bağlantı kümeleme yönteminde iki küme arasındaki uzaklık, küme merkezlerine dayalı olarak hesaplanır (Johnson ve Wichern, 2014). Bu yöntemde, A ve B gibi iki küme arasındaki uzaklık, iki kümenin ortalama vektörleri arasındaki Öklid uzaklığı olarak tanımlanır:

$$d(A, B) = d(\bar{y}_A, \bar{y}_B) \quad (3.6.)$$

Değişken sayısı p , birim sayısı n iken, X veri matrisi, j ise tüm elemanları 1 olan bir vektör olmak üzere c_i kümesinin küme merkezi $\bar{y}_i$,

$$\bar{y}_i = \frac{1}{n} \sum_{j=1}^n x_{ij} \quad (3.7.)$$

eşitliği ile hesaplanır (Rencher, 2002). Bu yöntemde, her adımda kümeler yeni birimler eklendikçe kümelerin merkezleri de değişecektir. Ayrıca diğer bağlantı yöntemlerinin aksine bu yöntem de kümelerin birleşme noktaları monoton artan yapıda olmayabilir. Yani herhangi bir adımda kümeler önceki adımlarda daha küçük uzaklık değerlerinde birleşebilir (Hair ve Black, 2009). Bu durum grafiksel olarak kümelerin yorumlanmasını zorlaştırır. Bu yöntemin bir diğer özelliği de aşırı değerlerden daha az etkilenen yöntemlerden olmasıdır. Şekil 3.11’de kümeler arasındaki uzaklıkların merkezi bağlantı kümeleme yöntemine göre belirlenmesi grafiksel olarak gösterilmiştir.



Şekil 3.11. Merkezi bağlantı kümeleme yöntemine göre iki küme arasındaki uzaklığın belirlenmesi

3.1.6. Medyan Bağlantı Kümeleme Yöntemi (Median Linkage Method)

A ve B gibi iki küme birleştirildiğinde elde edilen yeni kümenin merkezi, birleştirilen kümelerden birim sayısı fazla olanın merkezine yakın olmaktadır. Küme birim sayısına olan bağlılığı ortadan kaldırmak için uygulanan medyan bağlantı yönteminde iki küme merkezi arasındaki orta nokta merkez olarak alınır (Rencher, 2002). A ve B gibi iki kümenin bağlanması ile oluşan yeni küme merkezi,

$$\bar{y}_{\{A,B\}} = \frac{1}{2}(\bar{y}_{\{A\}} + \bar{y}_{\{B\}}) \quad (3.8.)$$

ile elde edilebilir. Medyan bağlantı kümeleme yönteminde c_1 kümesinin c_i ve c_j kümelerinin birleşimini ifade eden (c_i, c_j) kümesine olan uzaklığı,

$$d(c_1, (c_i, c_j)) = \frac{1}{2}d(c_1, c_i) + \frac{1}{2}d(c_1, c_j) - \frac{1}{4}d(c_i, c_j) \quad (3.9.)$$

ile hesaplanır. Bu yöntemde de kümelerin bağlantı uzaklık değerleri monoton artan olmayabilir (Alpar, 2011).

3.1.7. Ward Bağlantı Kümeleme Yöntemi (Ward Linkage Method)

Ward bağlantı yöntemi, 1963 yılında Joe H. Ward tarafından önerilmiş olup, alanyazında “en küçük varyans yöntemi” olarak da bilinir (Everitt, 2011). Bu yöntemde, küme için ve kümeler arası uzaklıklar kullanılır (Ward, 1963; Hair ve Black, 2009). İki kümenin birleşmesi sonucu oluşan bilgi kaybını minimize etmeyi amaçlayan bu yöntemde, c_i ve c_j gibi iki kümenin birleşmesiyle oluşan bilgi kaybındaki artış,

$$I_{(c_i, c_j)} = \frac{n_i n_j}{n_i + n_j} (\bar{x}_{c_i} - \bar{x}_{c_j})' (\bar{x}_{c_i} - \bar{x}_{c_j}) \quad (3.10.)$$

eşitliği ile hesaplanır (Rencher, 2002). Ward bağlantı kümeleme yönteminde c_1 kümesinin c_i ve c_j kümelerinin birleşimini ifade eden (c_i, c_j) kümesine olan uzaklığı,

$$d(c_1, (c_i, c_j)) = \frac{(n_1 + n_i)d(c_1, c_i) + (n_1 + n_j)d(c_1, c_j) - n_1 d(c_i, c_j)}{n_1 + n_i + n_j} \quad (3.11.)$$

eşitliği kullanılarak hesaplanır (Xu ve Wunsch, 2009).

Ward bağlantı yöntemi, birim sayısı az olan kümeleri birleştirme eğilimindedir, çünkü kareler toplamına dayanır ve kareler toplamı da doğrudan kümedeki birim sayısına bağlıdır. Aynı zamanda eşit sayıda birime sahip kümeler oluşturma eğilimindedir. Dolayısıyla araştırmacının beklentisi yaklaşık veya eşit sayıda birime sahip kümeler elde etmekse bu yöntem uygundur (Hair ve Black, 2009).

3.1.8. Esnek Beta Yöntemi

Lance ve Williams'ın (1967), ortaya koydukları bu yöntem bir eşitliği ifade ediyor olup, herhangi iki küme arasındaki uzaklık değerinin bu eşitlik ile elde edilebilir. Bu yöntemde, c_i kümesinin c_i ve c_j kümelerinin birleşimini ifade eden (c_i, c_j) kümesine olan uzaklığı,

$$d(c_i, (c_i, c_j)) = \alpha_i d(c_i, c_i) + \alpha_j d(c_i, c_j) + \beta d(c_i, c_j) + \gamma |d(c_i, c_i) - d(c_i, c_j)| \quad (3.12.)$$

kullanılarak bulunabilir (Rencher, 2002). Yukarıdaki eşitlikteki parametreler için Lance ve Williams (1967) aşağıdaki kısıtlamaları önermişlerdir (Rencher, 2002);

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

$$\alpha_i = \alpha_j$$

$$\gamma = 0$$

$$\beta < 1$$

Kısıtlar incelendiğinde $\alpha_i = \alpha_j$ ve $\gamma = 0$ alındığında $\alpha_i = \alpha_j = \frac{1-\beta}{2}$ olarak

elde edilip, eşitliğin sadece β parametresinde bağlı olduğu görülebilir. Bu nedenle de esnek beta yöntemi olarak adlandırılır. Lance ve Williams (1967), $\beta = -0,25$ gibi küçük negatif bir β değerinin kullanılmasını önermişlerdir. Ayrıca veri kümesi aykırı

değerler içeriyorsa, bu değerlerin etkisini azaltmak için $\beta = -0,5$ gibi daha küçük β değerlerinin daha uygun olacağını belirtmişlerdir (Rencher, 2002).

Eşitlik (3.12)'deki α_i , α_j , β ve γ parametrelerinin kullanılan yönteme göre aldığı değerler (n_i , n_j ve n_l sırasıyla c_1 , c_i ve c_j kümelerindeki birim sayıları olmak üzere) Çizelge 3.7'de verilmiştir.

Çizelge 3.7. Lance ve Williams'ın önerdiği genel formüldeki parametrelerin bağlantı yöntemine göre değerleri

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

3.2. Kümelerin Geçerliliği

Kümeleme analizinde kullanılan her bir algoritma veriyi parçalayabilir fakat farklı algoritmalar veya giriş parametreleri, farklı kümeler oluşturabilir; farklı kümeleme yapılarını ortaya çıkarabilir. Bu nedenle, elde edilen kümelerin anlamlı olmasını anlamına gelen küme geçerliliği oldukça önemlidir (Xu ve Wunsch, 2009; Jain ve Dubes, 1988). Yapısal olarak geçerli kümelerin elde edilmesi de, sonuçların farklı yöntemlerde kullanılmasını mümkün kılar (Jain ve Dubes, 1988).

Bir küme geçerlilik indeksi uygulanan kümeleme analizi sonucu elde edilen yapının yeterliliğini nesnel olarak yorumlanmaya uygun olacak şekilde ölçer. Yeterli bir kümelenme yapısı, elde edilen sonucun verideki doğru bilgiyi sağlamada veya veride gizlenmiş nitelikleri ortaya koymada başarılı olduğu anlamını taşır. Küme yapılarının geçerliliğini ölçen indeksler üç başlık altında toplanabilir; dışsal (external) indeksler, içsel (internal) indeksler ve oransal (relative) indeksler (Jain ve Dubes, 1988). Dışsal indeksler, sahip olunan bir ön bilgi ile elde edilen küme yapısını eşleştirerek karşılaştırır ve performansını ölçer. İçsel indeksler sadece verinin kendisini kullanarak, veri ile elde edilen küme yapısı arasındaki uyumu değerlendirir. Oransal indeksler ise iki farklı küme yapısının hangisinin veriyi daha iyi temsil ettiğini belirlemeye yönelik indekslerdir (Jain ve Dubes, 1988).

3.2.1. Dışsal İndeksler

N birim içeren bir X veri matrisi kullanılarak elde edilmiş bilinen bir P parçalanması ve aynı veri kümesine uygulanan bir kümeleme algoritması sonrası elde edilen parçalanma C ile gösterilsin. Dışsal indeks değeri, bu iki parçalanma kullanılarak elde edilir. Örneğin, X veri matrisinin elemanları olan x_i ve x_j birim çifti ele alınsın. Bu birimlerin C ve P parçalanmaları için sahip olabildikleri (bulunabilecekleri) dört durum söz konusudur (Xu ve Wunsch, 2009).

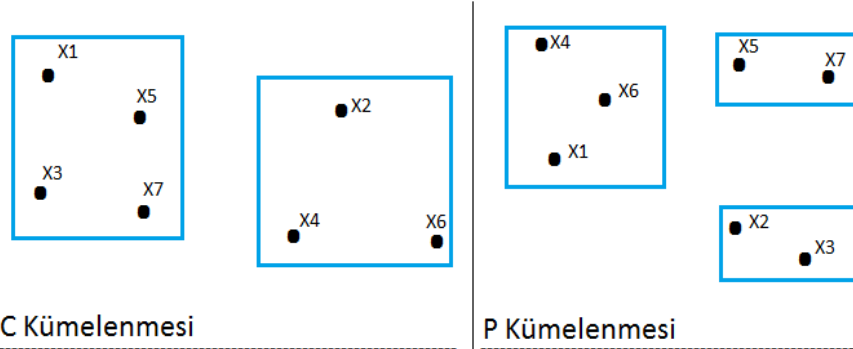
Durum 1: x_i ve x_j , C ve P parçalanmalarında aynı kümelerde yer almaktadır.

Durum 2: x_i ve x_j , C parçalanmasında aynı kümede yer alırken ve P parçalanmasında farklı kümelerde yer almaktadır.

Durum 3: x_i ve x_j , C parçalanmasında farklı kümelerde yer alırken ve P parçalanmasında aynı kümede yer almaktadır.

Durum 4: x_i ve x_j , C ve P parçalanmalarında farklı kümelerde yer almaktadır.

Muhtemel dört durumu göstermek için kullanılacak **C** ve **P** kümelenmesi Şekil 3.12'deki gibi olsun.



Şekil 3.12. **C** ve **P** kümelenmesi için birim çiftlerinin muhtemel dört durumu için örnek gösterim

Çizelge 3.8. Şekil 3.12'deki örnek **C** ve **P** kümelenmesi için hesaplanan muhtemel dört duruma karşılık gelen değerler ve açıklamaları

Durum	Veri kümesindeki birim çiftleri	Toplam
1	(x_4, x_6) ve (x_5, x_7)	2
2	$(x_1, x_5); (x_1, x_3); (x_1, x_7); (x_3, x_5); (x_2, x_4); (x_2, x_6); (x_3, x_7)$	7
3	$(x_1, x_4); (x_2, x_3)$ ve (x_1, x_6)	3
4	$(x_1, x_2); (x_3, x_4); (x_3, x_6); (x_5, x_2); (x_5, x_4); (x_5, x_6); (x_7, x_2); (x_7, x_4)$ ve (x_7, x_6)	9

Tanımlanan dışsal indeks değerleri sırasıyla Çizelge 3.8'de verilmiş olan dört duruma karşılık gelen a, b, c ve d parametrelerine bağlı tanımlanır. Tanımlanan bu değerlerden a ve d değerleri parçalanmaların benzerliklerini ifade ederken, diğer iki değer farklılığını temsil eder. Örneğin, her iki kümelenmede de aynı küme içerisinde yer alan birimler (x_4, x_6) ve (x_5, x_7) olduğundan $a=2$, C parçalanmasında aynı kümede yer alırken, P parçalanmasında farklı kümelerde bulunan $(x_1, x_5); (x_1, x_3); (x_1, x_7); (x_3, x_5); (x_2, x_4); (x_2, x_6); (x_3, x_7)$ ikililerinden dolayı $b=7$ olarak elde edilmiştir. Bu parametre değerlerine bağlı tanımlanan ve en yaygın kullanılan dört indeks değeri aşağıda tanımlanmıştır:

3.2.1.1. Rand İndeksi (Rand, 1971)

Rand indeksi, $[0,1]$ aralığında değerler alıp, bire ne kadar yakın ise söz konusu iki parçalanmanın birbirine o kadar benzer olduğu anlamı taşır. Bu indeks aşağıdaki gibi tanımlanır (Xu ve Wunsch, 2009):

$$R = \frac{a + d}{a + b + c + d} \quad (3.13)$$

3.2.1.2. Jaccard Katsayısı (Jaccard, 1901)

Jaccard katsayısı aşağıdaki gibi tanımlanır:

$$J = \frac{a}{a + b + c} \quad (3.14)$$

Bu katsayı da Rand indeksi gibi $[0,1]$ aralığında değerler alır. Ancak Rand indeksinden farkı “d” parametresini hesaba katmıyor olmasıdır. Bir değerine yakın değerler alması benzerliğin arttığı, uzaklaşması iki parçalanmanın birbirinden farklı olduğu şeklinde yorumlanır (Xu ve Wunsch, 2009).

3.2.1.3. Fowlkes ve Mallows İndeksi (Fowlkes ve Mallows, 1983)

Fowlkes ve Mallows tarafından önerilen bu indeks aşağıdaki gibi tanımlanmakta olup, diğer indeksler için yapılan yorumlara benzer olarak bir değerine yakın değer alması parçalanmalar arası benzerliğin arttığı şeklinde yorumlanır (Xu ve Wunsch, 2009) :

$$FM = \sqrt{\left(\frac{a}{a + b}\right) \cdot \left(\frac{a}{a + c}\right)} \quad (3.15)$$

3.2.1.4. Γ İstatistiği (Xu ve Wunsch, 2009)

Γ istatistiği,

$$\Gamma = \frac{Ma - m_1 m_2}{\sqrt{m_1 m_2 (M - m_1)(M - m_2)}} \quad (3.16)$$

şeklinde tanımlanır. Eşitlikteki $M = a + b + c + d$, $m_1 = a + b$ ve $m_2 = a + c$ olarak tanımlanır. Bu indeks değeri $[-1, 1]$ aralığında değerler alıyor olup bir değerine yakınlıştıkça parçalanmaların benzerliğin arttığı, tersi durumda parçalanmalar birbirinden farklılaştığı yorumu yapılır (Xu ve Wunsch, 2009).

Çizelge 3.9. Küme geçerliliğinin ölçülmesinde kullanılan dışsal kriterler ve değer aralıkları

Küme Geçerliliğinin Ölçülmesinde Kullanılan Kriterler		
Kriter	İfadesi	Aralığı
Rand İndeksi	$R = \frac{a + d}{a + b + c + d}$	[0,1]
Jaccard Katsayısı	$J = \frac{a}{a + b + c}$	[0,1]
Fowlkes ve Mallows İndeksi	$FM = \sqrt{\frac{a}{a + b} \frac{a}{a + c}}$	[0,1]
Γ İstatistiği	$\Gamma = \frac{Ma - m_1 m_2}{\sqrt{m_1 m_2 (M - m_1)(M - m_2)}}$	[-1, 1]

3.2.2. İçsel İndeksler

İçsel indeksler sadece verinin kendisini kullanarak, veri ile elde edilen küme yapısı arasındaki uyumu değerlendirir (Jain ve Dubes, 1988). Bu çalışmada kofenetik korelasyon katsayısı ve Wilk's' Lambda istatistiği gibi özellikle aşamalı kümeleme yöntemlerinde kullanılan indeksler açıklanacaktır.

3.2.2.1. Kofenetik (Cophenetic) Korelasyon Katsayısı

Kofenetik korelasyon katsayısı (Jain ve Dubes, 1988; Rohlf ve Fisher, 1968), küme yapılarının geçerliliğini ölçmek için kullanılan bir indekstir. Bir uzaklık matrisinde $d(i, j)$, i ve j birimleri arasındaki uzaklığı; $d_c(i, j)$ belirli bir kümeleme yönteminin uygulanması sonrası i ve j birimlerinin ilk kez aynı kümede birleşme uzaklığı iken kofenetik korelasyon katsayısı, başlangıçta elde edilen $n \times n$ boyutlu uzaklık matrisi ile sonda elde edilen $n \times n$ boyutlu uzaklık matrisinin değerleri arasındaki korelasyon katsayısı olup (3.17)'deki gibi tanımlanır. Burada ana köşegen üzerinde $M=n(n-1)/2$ değeri kullanılmaktadır çünkü söz konusu iki matris simetriktr.

$$CPCC = \frac{(1/M) \sum (d(i, j) d_c(i, j) - (m_D m_C))}{\left[(1/M) \sum (d^2(i, j) - m_D) \right]^{1/2} \left[(1/M) \sum (d_c^2(i, j) - m_C) \right]^{1/2}} \quad (3.17)$$

Bu eşitlikte, $m_D = (1/M) \sum d(i, j)$; $m_C = (1/M) \sum d_c(i, j)$ olarak tanımlanır.

Kofenetik korelasyon katsayısı (CPCC), $[-1,1]$ aralığında değer alır ve bir (1)'e yakın katsayı değeri daha iyi uyum ve verinin daha iyi temsil edildiği anlamına gelir. Fakat burada kastedilen yakınlığın ne ölçüde olması gerektiği konusunda tartışmalar vardır. Rohlf ve Fisher (1968) düzgün ve normal dağılımdan rastgele alınmış veriler üzerinde ortalama bağlantı yöntemini kullanarak bu katsayı değerinin değişimini incelemişlerdir ve çalışmalarında bu katsayı değerinin kullanılan uzaklık ölçüsüne karşı oldukça hassas olduğunu göstermişlerdir. Ayrıca kofenetik katsayı değerinin kullanılan bağlantı yöntemiyle de ilişkisi olduğunu belirtmiştir (Jain ve Dubes, 1988).

3.2.2.2. Wilk's Lambda Test İstatistiği

Çok değişkenli varyans analizinde gruplar arasındaki farklılığının testinde kullanılan Wilks'in lambda test istatistiği, Freidman ve Rubin (1967) tarafından

küme geçerliliğinin değerlendirilmesi için önerilmiştir. Wilk's Lambda test istatistiği,

$$\lambda = \frac{|W|}{|W + B|} = \frac{|W|}{|T|} \quad (3.18.)$$

$$B = \sum_{j=1}^k n_j (\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})' \quad (3.19.)$$

$$W = \sum_{j=1}^k \sum_{i=1}^{n_j} n_j (X_{ij} - \bar{X})(X_{ij} - \bar{X})' \quad (3.20.)$$

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

$\bar{X}_j$: j. kümenin ortalama vektörü

$\bar{X}$: Tüm birimlerin örneklem ortalama vektörü

n_j : j. kümedeki birim sayısı

eşitliği ile tanımlanır. Eşitlikte yer alan $|W|$, kümeler içi kareler ve çarpımlar toplamı matrisinin determinantını; $|B|$ kümeler içi kareler ve çarpımlar toplamı matrisinin determinantını ve $|T|$ ise genel kareler ve çarpımlar toplamı matrisinin determinantını göstermektedir (Tatlıldil, 1996). Wilk's Lambda test istatistiği aynı zamanda benzerlik oran kriteri olarak da adlandırılmaktadır (Dinçer, 1992; Erişoğlu, 2011). Wilk's Lambda test istatistiği yeterince küçük ise küme merkezlerine ait ortalama vektörlerinin eşit olduğunu öne süren yokluk (sıfır) hipotezi red edilir (Overall ve Klett, 1972; Erişoğlu, 2011). Elde edilen λ değeri ne kadar küçükse gerçekleştirilen kümeleme o kadar başarılıdır.

3.3.3. Oransal İndeksler

Dışsal ve içsel indeksler yoğun hesaplamalara dayanan istatistiksel testlerin kullanımını gerektirir. Oransal indekslerde bu gereklilik yoktur ve farklı algoritmalarla veya aynı algoritmanın farklı başlangıç değerleriyle elde edilmiş kümeleme sonuçlarının karşılaştırılmasına dayanır (Xu ve Wunsch, 2009).

Aşamalı kümeleme analizinde verideki küme sayısı ağaç diyagramında kişiye bağlı bir kesme noktası seçilerek belirlenir. Aşamalı olmayan yöntemlerde ise küme sayısının belirlenmesi çalışma öncesinde yapılmaktadır. Bu nedenle “k” küme sayısının belirlenmesi oldukça önemli bir konudur. Bazı durumlarda küme sayısı araştırmacının deneyimi veya farklı uygulamalara dayanan ön bilgiyle tahmin edilebilse de, hala yoğun olarak verinin kendisi kullanılarak çıkarım yapma yoluna gidilir. Var olandan fazla ya da eksik tahmin elde edilen küme sayısı, kümelerin kalitesini düşürecektir. Çok fazla kümeyi içeren bir parçalanma gerçek küme yapısını karmaşık hale getirebileceği gibi, elde edilen sonuçları analiz etmek ve yorumlamak da zorlaşır. Diğer taraftan çok az küme içeren bir parçalanma bilgi kaybına neden olmanın yanında yanlış karar almaya neden olabilir (Xu ve Wunsch, 2009). Oransal indeksler de işte bu temel probleme odaklanarak tanımlanmış indekslerdir. İndeksler hesaplanırken sınırlı sayıda k değeri için hesaplama yapılabilirken, belirli aralıkta çok sayıda k değeri içinden en uygun değerin belirlenmesi yoluna da gidilebilir. Alanyazında bu amaçla birçok indeks önerilmiştir. Silhouette indeksi, Calinski ve Harabasz indeksi, Kaufman ve Rousseeuw indeksi, Davies-Bouldin indeksi, Marriott indeksi, Dunn indeksi, Mojena indeksi, Krzanowski ve Lai İndeksi ile Gap istatistiği en önemli ve yaygın olarak kullanılan indekslere örnek olarak verilebilir (Xu ve Wunsch, 2009). Bu çalışmada sadece aşamalı kümeleme analizi kullanılacağından aşamalı olmayan kümeleme yöntemlerinde kullanılan bu indekslerin tanım ve hesaplamalarına değinilmeyecektir.

4. UZAKLIK VE BENZERLİK ÖLÇÜLERİ

Benzerlik iki birim veya nesnenin ne ölçüde benzer olduklarının sayısal derecesi olarak tanımlanabilir. Ortaklık, yakınlık ve ilişki kavramlarıyla bağlantılıdır (Sneath ve Sokal, 1973). Benzerlik ölçüleri ise iki birimin benzerliğinin sayısal tahminini elde etmek amacıyla ortaya atılmıştır. Benzerlik iki nitelik veya birim arasındaki ilişkinin miktarını yansıtır ve genellikle ya -1 ile +1 arasında ya da normalleştirilerek 0 ile 1 arasında değer alır (Teknomo, 2006).

Uzaklık ise farklılıklara dayanır. Yani iki birimin birtakım niteliklere sahip olup olmamasına dayanan ve bu birimlerin ne ölçüde farklı olduklarının sayısal derecesini ifade eder. Zıtlık veya uyumsuzlukların bir ölçüsüdür (Teknomo, 2006).

Benzerlik ölçümleri bir gözlemin diğerlerinden ayırt edilmesini sağlar. Bu sayede gözlemler benzerlik veya farklılıklarına dayalı olarak gruplara ayrılabilir. Gözlemler alt gruplara veya kümelere bir kere atandıktan sonra her grubun karakteristikleri anlaşılabilir ve kümelerin özellikleri açıklanabilir. Gruplama ile bilginin düzenlenmesi ve sonuçların elde edilmesi daha etkin sağlanır. Yeni gözlemler bu alt grup veya kümelere sınıflandırılabilir ve özellikleri tahmin edilebilir. Veri bu sayede basitleştirilerek daha etkin değişkenler üzerinde çalışılabilir. Veri kümesi içerisindeki belirgin olmayan gizli yapı bulunabilir. Verinin içerisindeki kümelenme yapısına ve bu kümelerin parametre tahminine dayalı olarak plan ve kararlar alınabilir (Servi, 2009).

Bu bölümde 44 ikili benzerlik ölçüsü ve 10 ikili uzaklık ölçüsü verilecektir. Ölçüler arasındaki ilişkileri veya birliktelikleri gözlemlemek için ölçülerin genel özellikleri üzerinden gruplama çalışması yürütülmüştür. Sözü edilen özellikler benzerlik ölçüsüne negatif uyumların dâhil olup olmaması, ölçünün pay ve paydasındaki terimler arasındaki ilişki, ölçünün niteliklerin görülme sıklığına mı yoksa ilişkiye mi dayalı olduğu gibi özelliklerdir. Benzerlik ve farklılık kavramları ile bazı temel tanım ve özellikler aşağıda verilmiştir.

4.1. Benzerlik ve Farklılık Ölçülerine İlişkin Temel Kavramlar

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

$$\text{i. } d(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \quad (\text{Pozitiflik}) \quad (4.1.)$$

$$\text{ii. } d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i) \quad (\text{Simetri}) \quad (4.2.)$$

$$\text{iii. } d(\mathbf{x}_i, \mathbf{x}_i) = 0 \quad (\text{Özdeşlik}) \quad (4.3.)$$

koşulları sağlanıyorsa d , X üzerinde bir uzaklık fonksiyonu olarak adlandırılır (Deza ve Deza, 2006).

Tanım 4.1.1. : (Uzaklık Uzayı) Bir d uzaklığı ile ilişkilendirilmiş X 'in (X, d) kümesine bir uzaklık uzayı denir.

Tanım 4.1.2. : (Benzerlik) X bir küme olmak üzere negatif olmayan, simetrik $s : X \times X \rightarrow \mathcal{R}$ fonksiyonu tüm $x, y \in X$ için $s(x, y) \leq s(x, x)$ koşulunu sağlıyorsa ve eşitlik ancak ve ancak $x = y$ olduğunda gerçekleşiyorsa s fonksiyonu benzerlik (yakınlık) olarak adlandırılır.

Bir $n \times p$ boyutlu X veri matrisinin her biri $1 \times p$ boyutlu $x_1, x_2, \dots, x_n$ gözlem vektörleri için $i = 1, 2, \dots, n$ ve $j = 1, 2, \dots, n$ olmak üzere $n \times n$ boyutlu $D = (d_{ij})$ uzaklıklar matrisi hesaplanabilir. Burada $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ 'dir. D uzaklıklar matrisi köşegen elemanları sıfır olan simetrik bir matristir.

4.2. İki Sonuçlu (Binary) Değişkenler İçin Kullanılan Benzerlik Ölçüleri

İkili benzerlik ve uzaklık ölçüleri sınıflandırma, kümeleme analizi ve tespit-teşhis çalışmalarında kritik öneme sahiptir. Çünkü doğru ölçü seçimi daha doğru ve güvenilir sonuç ortaya koyacaktır. Yüzyılı aşkın süredir, birçok araştırmacı kendi alanlarında en anlamlı, güvenilir ikili benzerlik ölçüsünü bulmak için yoğun çaba harcamaktadır. Jaccard benzerlik ölçüsü-ikili yapıda ki ölçülerin ilk uygulaması-

özellikle ekolojik sınıflandırmalarda sıklıkla kullanılmış ve hala yoğun olarak kullanılmaktadır. Daha sonra sırasıyla biyoloji, kökenbilim, sınıflandırma, görüntü işleme, jeoloji ve kimya alanlarında kullanılmıştır. Ek olarak, özellikle biyometri alanında göz taramaları, parmak izi tespiti ve el yazısı teşhisi gibi alanlarda aktif olarak kullanılmaktadır. Bahsedilen bu alanlarda kullanılmak üzere birçok ölçü ortaya atılmış, özellikleri incelenmiştir.

Ortaya konan bu ölçüler, verideki tüm değişkenler sadece iki sonuçlu değer alıyorsa bu verideki gözlemler arasındaki yakınlıkları ölçmek için tanımlanmışlardır. İkili benzerlik ve uzaklık ölçüleri kendi içerisinde simetrik ve simetrik olmayan ölçüler olarak iki gruba ayrılabilir. Bu ayrımın sebebi temelde, iki sonuçlu değişkenlerin aldığı iki değerın eşit öneme sahip olup olmamasıdır (Kaufman ve Rousseeuw, 1990). Bu konu hala tartışılmakta olan bir konudur. Bazı durumlarda 0-0 eşleştirmesi, 1-1 eşleştirmesi ile eşdeğer olabilir. Örneğin cinsiyetin erkek için 1 değeri ve bayan için 0 değeri ile kodlanması veya tam tersi şeklinde kodlanması arasında bir fark yoktur. Bazı durumlarda ise 0-0 eşleştirmesi bir gözlem üzerinde herhangi bir özelliğin bulunup bulunmadığı ile ilgili olabilir. Örneğin bir böceğin incelenen özelliği bir kanada sahip olup olması gibi (Sokal ve Sneath, 1963; Servi, 2009). Ancak çoğunlukla pozitif uyum, negatif uyumdan daha önemli ve anlamlı bulunmuştur. Bunun yanında birçok çalışmada negatif uyumun benzerlik hesaplanmasında kullanılıp kullanılamayacağı da tartışılmıştır (Finely, 1884; Gilbert, 1884; Goodman ve Kruskal, 1954; Goodman ve Kruskal, 1959; Sokal ve Sneath, 1963; Goodman ve Kruskal, 1963; Goodman ve Kruskal, 1972; Sneath ve Sokal, 1973; Choi, 2008). Birimlerin belirli bir niteliği taşımamaları onların benzer olduğunu göstermez. Çünkü iki birimde var olmayan olası sonsuz nitelik söz konusu olabilir (Sokal ve Sneath, 1963).

İkili Veri Yapılarının Gösterimi

Operational Taxonomic Units (OTUs), belirttiğimiz 54 ölçünün ifade edilmesinde kolaylık sağlamaktadır. Verinin 2x2 frekans tablosu olarak verilmesidir.

i ve j birimlerine ilişkin olarak aşağıdaki gibi gösterilir:

Çizelge 4.1. İkili Yapıdaki Verilere İlişkin Frekans Tablosu

		i		
		1 – Var	0 - Yok	Toplam
j	1 –Var	a	b	a+b
	0 - Yok	c	d	c+d
	Toplam	a+c	b+d	a+b+c+d=n

Tabloda;

- a: İncelenen niteliğin i ve j birimlerinin her ikisinde de bulunma sayısını (Pozitif Tam Uyum)
- b: İncelenen niteliğin j de bulunup, i de bulunmama sayısı (Uyuşmazlık)
- c: İncelenen niteliğin i de bulunup, j de bulunmama sayısı (Uyuşmazlık)
- d: İncelenen niteliğin birimlerin ikisinde de bulunmama sayısı (Negatif Tam Uyum)

Burada,

a+d: Tam uyum sayısını,

b+c: Uyuşmazlıkların sayısını ifade eder.

Frekans tablosu incelendiğinde, herhangi iki birim arasındaki benzerlik katsayısının en açık örneği,

$$r = \frac{a}{a+b} \text{ veya } r = \frac{a}{a+c} \quad (4.4.)$$

olarak tanımlanabilir.

Fakat bu tanımlama, örneğin iki birimi iki tür olarak düşünersek, bu türler aynı sayıda bölgede bir arada gözlemlenmedikleri sürece simetrik değildir. Bu nedenle, yapılan bu iki tanımın bir şekilde ortalamasının alınması oldukça kabul edilebilirdir. Bu düşünceyle farklı yaklaşımlarla farklı tanımlamalar yapılabilir. Doğrudan aritmetik ortalaması alınarak,

$$r = \frac{1}{2} \left[\left(\frac{a}{a+c} \right) + \left(\frac{a}{a+b} \right) \right] \quad (4.5.)$$

şeklinde bir ölçü önerilmiştir.

Benzer şekilde,

$$r = \frac{a}{\sqrt{(a+b)}\sqrt{(a+c)}} \quad (4.6.)$$

ölçüsü verilen iki oranın geometrik ortalaması alınarak ortaya konulmuştur.

Bu iki oranın harmonik ortalamasının alınmasıyla aşağıdaki ölçü elde edilir:

$$r = \frac{2a}{2a + b + c} \quad (4.7.)$$

Ayrıca, $d/(b+d)$ ve $d/(c+d)$ oranları da hesaba katılarak ölçülerin sayısı genişletilmiştir. Verilen ilk iki oran ile son iki oran toplamının aritmetik ortalaması alınarak aşağıdaki ölçü ortaya atılmıştır:

$$r = \frac{1}{4} \left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d} \right] \quad (4.8.)$$

Benzer şekilde verilen bu temel dört oranın geometrik ortalamasının karesi de bir ölçü olarak önerilmiştir.

$$r = \frac{ad}{\sqrt{((a+b)(a+c)(b+d)(c+d))}} \quad (4.9.)$$

Uzaklık ve benzerlik ölçüleri temel olarak bahsettiğimiz şekildeki yaklaşımlarla ortaya konulmaktadır. Bu bölümün devamında tez kapsamında derlenmiş ve alanyazında tanımlanmış ölçülerin tanım ve ifadeleri verilmiştir.

Jaccard ölçüsü (Jaccard, 1901) özellikle biyolojik birliktelik çalışmaları (Cheetham ve Hazel, 1969), ekolojide türlerin kümelenmesi (Magurran, 2003; Pielou, 1969) gibi çeşitli amaçlarla, çok fazla alanda geniş çapta kullanım alanına

sahiptir (Sokal ve Sneath, 1963). Bu ölçü negatif uyumları (0-0) içermez. Bu nedenle pozitif uyum sayısının, toplam değişken sayısı ile negatif uyum sayısı farkına oranı olarak tanımlanır.

$$S_{\text{Jaccard}} = \frac{a}{a + b + c} \quad (4.10.)$$

Tanimoto katsayısı (Tanimoto, 1957) , çoğunlukla Cosine katsayısı olarak da bilinir, Jaccard ölçüsünün genişletilmiş bir halidir.

$$S_{\text{Tan}} = \frac{a}{(a + b) + (a + c) - a} \quad (4.11.)$$

olarak tanımlanır.

Tanimoto benzerlik ölçüsü, biyolojik sınıflandırma ve bilgi çıkarımı gibi alanlarda sıklıkla karşılaşılan normalleştirilmiş iç çarpımın bir diğer versiyonudur ve ikili parmak izi (Willett, 2003) tespiti amacıyla yoğunlukla kullanılır.

Dice ve Sorensen katsayısı (Dice, 1945; Sorensen, 1948) pozitif uyumlara iki kat ağırlık verir. Buradan, bu ölçü iki kat ağırlıklandırılmış pozitif uyumların, pozitif uyum sayısı ile uyumsuzlukların sayısının toplamına oranı olarak tanımlanır.

$$S_{\text{DS}} = \frac{2a}{2a + b + c} \quad (4.12.)$$

Dice (1945) bu ölçüyü “uyum indeksi” olarak tanımlarken, Sorensen (1948) “oran benzerliği” olarak isimlendirmiştir. Dice ve Sorensen katsayısı ikili veriler için Czekanowski (Gordon, 1999) katsayısına denktir ve Jaccard katsayısı ile monoton olarak ilişkilidir.

Dice ve Sorensen katsayısının aksine, Sokal&Sneath 1 katsayısı (Sokal ve Sneath, 1963) negatif uyumsuzluklara iki kat ağırlık verir. Böylece, bu ölçü pozitif

uyumların, pozitif uyumlara iki kat ağırlıklandırılmış uyuşmazlıkların eklenmesiyle oluşan değere oranı olarak ifade edilir.

$$S_{SS1} = \frac{a}{a + 2b + 2c} \quad (4.13.)$$

İkili değişkenler için Nei ve Li'nin genetik katsayısı (Nei ve Li, 1979),

$$S_{NL} = \frac{2a}{(a+b) + (a+c)} \quad (4.14.)$$

şeklinde tanımlanır ve bu katsayının Dice ve Sorensen katsayısına denk olduğu görülür.

Kulczynski 1 katsayısı (Kulczynski, 1927) paydasında uyuşmazlıkları hesaba katar ve şöyle tanımlanır;

$$S_{KULCI} = \frac{a}{b+c} \quad (4.15.)$$

Intersection benzerlik ölçüsü pozitif uyumların toplam sayısına eşittir ve bundan dolayı,

$$S_{INTER} = a \quad (4.16.)$$

olarak tanımlanır.

Sokal ve Michener katsayısı (Sokal ve Michener, 1958), çoğunlukla “basit uyum ölçüsü veya eşleştirme katsayısı” olarak bilinir; toplam uyumun, toplam nitelik sayısına oranı olarak tanımlanır.

$$S_{SM} = \frac{a+d}{a+b+c+d} \quad (4.17.)$$

Sokal ve Michener bu katsayıyı ikili niteliklerden oluşmuş iki birim arasındaki benzerliğin derecesinin bir ölçüsü olarak tanımlamıştır. Sokal'ın varsayımı, uyum katsayısının dağılımının, aynı ortalamalı binom dağılımıyla yaklaşık olarak aynı olabileceğidir (Sokal ve Michener, 1958). Basit uyum katsayısı, negatif uyum olmadığı durumda Jaccard katsayısına denktir.

Rogers ve Tanimoto katsayısı (Rogers ve Tanimoto, 1960), basit benzerliğin uzantısıdır. Paydasında uyumsuzlıklara iki kat ağırlık verir ve

$$S_{RT} = \frac{a + d}{a + 2(b + c) + d} \quad (4.18.)$$

şeklinde tanımlanır.

Faith katsayısı, pay kısmında negatif uyuma yarım ağırlık vererek,

$$S_{FAITH} = \frac{a + 0.5d}{a + b + c + d} \quad (4.19.)$$

şeklinde tanımlanmıştır (Faith ve ark., 1987).

Gower ve Legendre katsayısı (Gower ve Legendre, 1986), uyumsuzlıklara payda kısmında yarım ağırlık vererek,

$$S_{GL} = \frac{a + d}{a + 0.5(b + c) + d} \quad (4.20.)$$

gibi tanımlanır.

Hamann katsayısı (Hamann, 1961) , toplam uyum sayısından uyumsuzlıkların farkının toplam nitelik sayısına oranı olarak aşağıdaki gibi tanımlanır;

$$S_{HAMANN} = \frac{(a + d) - (b + c)}{a + b + c + d} \quad (4.21.)$$

Bu ölçü -1 ile 1 arasında değer alabilir ve monoton olarak Basit Uyum, Sokal&Sneath-2 ve Rogers&Tanimoto katsayılarıyla ilişkilidir.

Dispersion katsayısı (Gordon, 1999), Hamann katsayısına benzerdir ve

$$S_{DISP} = \frac{ad - bc}{(a + b + c + d)^2} \quad (4.22.)$$

olarak tanımlanmıştır.

Sneath ve Sokal, 14 benzerlik ölçüsünü ele almışlardır (Sneath ve Sokal, 1973). Bu ölçülerden, Sokal&Sneath-2 ve Sokal&Sneath-3 katsayıları aşağıdaki gibi tanımlanır:

$$S_{SS2} = \frac{2(a + d)}{2a + b + c + 2d} \quad (4.23.)$$

$$S_{SS3} = \frac{a + d}{b + c} \quad (4.24.)$$

Sokal&Sneath-3 katsayısı pozitif ve negatif uyumları paydasında içermezken, Sokal&Sneath-2 katsayısı iki kat ağırlık verir.

Russell&Rao benzerlik ölçüsü (Russell ve Rao, 1940) pozitif uyumların, toplam nitelik sayısına oranı olarak verilir. Çoğunlukla nokta iç çarpım katsayısı olarak bilinir ve şöyle tanımlanır:

$$S_{RR} = \frac{a}{a + b + c + d} \quad (4.25.)$$

Özdeş iki birimin bu katsayı ile farklı ölçülebilmesi ilginçtir. Bu ölçüde, uyumsuzluk olmasa dahi negatif uyumlar arttıkça benzerlik değeri düşmektedir.

Baroni-Urbani&Buser-2 katsayısı (Baroni-Urbani ve Buser, 1976),

$$S_{BUB2} = \frac{\sqrt{ad} + a - (b+c)}{\sqrt{ad} + a + b + c} \quad (4.26.)$$

şeklinde tanımlanmıştır.

En geniş kullanım alanına sahip Pearson r (Pearson, 1905) korelasyon katsayısı, doğrusal veya çarpım moment korelasyonu olarak da bilinir. Ziraat çalışmasında ilişkisel metotlar geliştiren Karl Pearson'dan sonra bu isimle anılmıştır. Bu katsayı, skorların ortalamadan sapmalarının çarpımlarının toplamı alınarak hesaplanır.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4.27.)$$

Pearson çarpım moment katsayısının ikili değişkenler için hesaplanan karşılığı, Phi katsayısıdır (Yule, 1912). Pearson bu katsayıyı “Fourfold Point Korelasyon Benzerliği” olarak tanımlamaktadır (Pearson ve ark., 1913). Phi korelasyon katsayısı eşitlik (4.23) deki gibi tanımlanır:

$$S_{PHI} = \Phi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (4.28.)$$

İkili yapıda değerlere sahip iki birimin, eğer pozitif ve negatif uyumların toplamı, uyumsuzlıkların toplamından fazla ise pozitif yönde ilişkili oldukları düşünülür. Benzer şekilde, eğer pozitif ve negatif uyumların toplamı, uyumsuzlıkların toplamından az ise negatif yönde ilişkili oldukları düşünülür.

Phi korelasyon katsayısının pay kısmında uyumsuzlıkların çarpımını (bc) içermeyen diğer varyasyonları, Sokal&Sneath-IV (Sokal ve Sneath, 1963), Sokal&Sneath-V (Sokal ve Sneath, 1963), Ochiai (Ochiai, 1957) ve Gower (Wishart, 2001) katsayılarıdır. Bu ölçüler ise sırasıyla şöyle tanımlanır:

$$S_{SS4} = \frac{\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{b+d}}{4} \quad (4.29.)$$

$$S_{SS5} = \frac{ad}{(a+b)(a+c)(b+d)(c+d)^{0.5}} \quad (4.30.)$$

$$S_{Och2} = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (4.31.)$$

$$S_{GOWER} = \frac{a+d}{\sqrt{(a+b).(a+c).(b+d).(c+d)}} \quad (4.32.)$$

Yule-Q katsayısı (Yule, 1900), çoğunlukla ilişki katsayısı olarak da bilinir ve Yule-W katsayısı (Yule, 1912), ortaklık katsayısı olarak da bilinir, çapraz orana dayanır. Bu ölçüler, Goodman&Kruskal's Gamma (Goodman ve Kruskal, 1972) ordinal benzerlik ölçüsünün ikili versiyonlarıdır.

$$S_{YULEQ} = \frac{ad - bc}{ad + bc} \quad (4.33.)$$

$$S_{YULEw} = \frac{\sqrt{ad - bc}}{\sqrt{ad + bc}} \quad (4.34.)$$

olarak tanımlanmışlardır.

Çapraz orana dayalı olarak kurulan diğer ölçüler Pearson&Heron-2 katsayısı (Pearson ve Heron, 1913) ve Michael katsayısıdır (Michael, 1920).

Pearson&Heron-2 katsayısı,

$$S_{PHI} = \cos\left(\frac{\pi\sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}\right) \quad (4.35.)$$

ve Michael katsayısı ise,

$$S_{MICHAEL} = \frac{4(ad - bc)}{(a + d)^2 + (b + c)^2} \quad (4.36.)$$

şeklinde tanımlanır.

Goodman&Kruskal katsayısı (Goodman ve Kruskal, 1954, 1959, 1963, 1972) ve Anderberg katsayısı (Anderberg, 1973) olasılıksal katsayılardır. İkili veriler için Goodman & Kruskal katsayısı, Lambda olarak adlandırılır.

$$S_{Good\&Kruskal} = \frac{\sigma - \sigma'}{2n - \sigma'}, \sigma = \max(a, b) + \max(c, d) + \max(a, c) + \max(b, d) \quad (4.37.)$$

$$\sigma' = \max(a + c, b + d) + \max(a + b, c + d)$$

Anderberg katsayısı, sıklıkla Anderberg' D olarak da bilinir, Lambda katsayısının bir varyasyonudur ve

$$S_{Anderberg} = \frac{\sigma - \sigma'}{2n}, \sigma = \max(a, b) + \max(c, d) + \max(a, c) + \max(b, d) \quad (4.38.)$$

$$\sigma' = \max(a + c, b + d) + \max(a + b, c + d)$$

şeklinde tanımlanır.

Goodman&Kruskal ve Anderberg katsayıları ikili değerler alan birimlerin oluşumlarının tahmininde hatalı uyumları azaltmaya çalışır.

Peirce katsayısı (Peirce, 1884) ise şöyle tanımlanır:

$$S_{PEIRCE} = \frac{ab + bc}{ab + 2bc + cd} \quad (4.39.)$$

Diğer taraftan, Tarantula katsayısı (Abreu ve ark, 2006; Johnson, 1967) ve Ample katsayısı (Dallmeier ve ark, 2005) pay kısmında negatif uyumları içerirken, payda kısmında hariç tutar.

$$S_{\text{Ample}} = \frac{a(c+d)}{c(a+b)} \quad (4.40.)$$

$$S_{\text{Tarant}} = \frac{a(c+d)}{c(a+b)} \quad (4.41.)$$

olarak tanımlanmıştır.

Bazı benzerlik ölçüleri ise negatif uyumları içermez. Bu ölçülere, Ochiai-1, Sorgenfrei, McConnaughey ve Kulczynski-2 ölçüleri örnek verilebilir.

Ochiai-1 (Ochiai, 1957) ve Otsuka katsayıları birbirine denktir ve aşağıdaki gibi tanımlanır:

$$S_{\text{Och1}} = \frac{a}{\sqrt{(a+b)(a+c)}} \quad (4.42.)$$

Ochiai-1 ve Otsuka katsayıları, Cosine benzerlik ölçüsünün ikili versiyonlarıdır.

$$\cos \theta = \frac{\sum_i (X_{ij})(X_{ik})}{\sqrt{\sum_i (X_{ij})^2 \sum_i (X_{ik})^2}} \quad (4.43.)$$

Burada X_{ij} , j.birimin i. özelliğini göstermektedir ve sadece 0 veya 1 ikili değerlerini alır. $(X_{ij})(X_{ik})$ çarpımı pozitif uyumlara karşılık 1, uyumsuzluklar ve negatif uyumlara karşılık 0 elde edileceğinden, pay kısmı pozitif uyumların toplamına, yani a'ya eşittir. $(X_{ij})^2$ ve $(X_{ik})^2$ terimlerinden her biri, her bir birimdeki var oluşumların veya 1 değerlerinin toplamına eşittir. Sırasıyla a+b ve a+c ye eşittir. Bu nedenle payda kısmı $\sqrt{(a+b)(a+c)}$ olarak bulunur.

Sorgenfrei katsayısı (Sorgenfrei, 1959), korelasyon oranı olarak da bilinir, Ochiai-1 katsayısının karesine eşittir.

$$S_{SORG} = \frac{a^2}{(a+b)(a+c)} \quad (4.44.)$$

Ayrıca, Cosine katsayısının birçok varyasyonu aşağıdaki gibi tanımlanmıştır: McConnaughey katsayısı (McConnaughey, 1964) şöyle tanımlanır:

$$S_{McConn} = \frac{a^2 - bc}{(a+b)(a+c)} \quad (4.45.)$$

Forbes-1 katsayısı (Forbes, 1907),

$$S_{FORBES1} = \frac{na}{(a+b)(a+c)} \quad (4.46.)$$

şeklinde tanımlanır.

Forbes-2 katsayısı (Forbes, 1925),

$$S_{FORBES2} = \frac{na - (a+b)(a+c)}{n \min\{(a+b), (a+c)\} - (a+b)(a+c)} \quad (4.47.)$$

şeklinde tanımlanmıştır.

Kulczynski-2 katsayısı (Kulczynski, 1927) şöyle tanımlanmaktadır:

$$S_{KULC2} = \frac{\frac{a}{2}(2a+b+c)}{(a+b)(a+c)} \quad (4.48.)$$

Driver&Kroeber benzerlik ölçüsü (Driver ve Kroeber, 1932), Kulczynski-2 katsayısına denktir ve

$$S_{D\&K} = \frac{a}{2} \left(\frac{1}{a+b} + \frac{1}{a+c} \right) \quad (4.49.)$$

şeklinde tanımlanır.

Johnson katsayısı (Johnson, 1967), Kulczynski-2 katsayısının başka bir versiyonudur ve şöyle tanımlanır:

$$S_{JOHNSON} = \frac{a}{a+b} + \frac{a}{a+c} \quad (4.50.)$$

Tarwid katsayısı (Tarwid, 1960) şöyle tanımlanmaktadır:

$$S_{TARWID} = \frac{na - (a+b)(a+c)}{na + (a+b)(a+c)} \quad (4.51.)$$

Simpson katsayısı (Simpson, 1943), payda kısmında birimlerin toplam uyum sayılarından minimum olanı olarak tanımlanırken, pay kısmı ortak uyumları alır.

$$S_{SIMPSON} = \frac{a}{\min\{(a+b), (a+c)\}} \quad (4.52.)$$

Braun-Blanquet katsayısı (Braun - Blanquet, 1932) ise şöyle tanımlanır:

$$S_{\text{Braun\&Blanquet}} = \frac{a}{\max\{(a+b), (a+c)\}} \quad (4.53.)$$

Mountford katsayısı (Mountford, 1962) ise aşağıdaki gibi tanımlanır:

$$S_{MF} = \frac{a}{0,5(ab + ac) + bc} \quad (4.54.)$$

En yaygın bilinen uzaklık ölçüleri Öklid ile City Block uzaklık ölçüleridir (Krause, 1986). Öklid iki nokta arasındaki en kısa uzaklığın bir doğru ifade ettiğini belirtmiş ve bu uzaklığı aşağıdaki gibi tanımlamıştır:

$$D_{\text{Öklid}} = \sqrt{(b+c)} \quad (4.55.)$$

Bu uzaklıktan hareketle kare Öklid uzaklığı da,

$$D_{\text{KareÖklid}} = b + c \quad (4.56.)$$

City Block ölçüsü Manhattan ölçüsü olarak da bilinir. Aynı zamanda ikili veri kümelerinde bu ölçüler Hamming ve Minkowski ölçülerine denktir.

Ortalama Manhattan ölçüsü ise aşağıdaki gibi tanımlanır:

$$D_{\text{Ort.Manhat.}} = \frac{b+c}{a+b+c+d} \quad (4.57.)$$

Size Difference uzaklık ölçüsü Ortalama Manhattan ölçüsünün karesine denktir ve

$$D_{\text{SizeD}} = \frac{(b+c)^2}{(a+b+c+d)^2} \quad (4.58.)$$

şeklinde tanımlanmaktadır (Choi, 2008).

Pattern Difference ve Variance uzaklık ölçüleri Ortalama Manhattan ölçüsünün diğer varyasyonlarıdır ve şöyle tanımlanmaktadırlar:

$$D_{\text{Pattern}} = \frac{b+c}{(a+b+c+d)^2} \quad (4.59.)$$

$$D_{\text{Var}} = \frac{b+c}{4(a+b+c+d)} \quad (4.60.)$$

Lance&Williams uzaklık ölçüsü (Lance ve Williams, 1967) Dice benzerlik ölçüsünün uzaklık formuna denk olup,

$$D_{L\&W} = \frac{b+c}{2a+b+c} \quad (4.61.)$$

şeklinde tanımlanır. Bu ölçü ikili yapıdaki veriler için Bray-Curtis uzaklık ölçüsüne denktir (Bray ve Curtis, 1957).

Hellinger (Bhattacharyya, 1943) ve Chord (Legendre ve Legendre, 1998) katsayılarının ikili versiyonları cosine benzerlik ölçüsünün dönüştürülmesiyle elde edilen uzaklık ölçüleridir ve

$$D_{\text{Chord}} = 2 \left(1 - \frac{a}{\sqrt{(a+b)(a+c)}} \right) \quad (4.62.)$$

$$D_{\text{Hellinger}} = 2 \sqrt{1 - \frac{a}{\sqrt{(a+b)(a+c)}}} \quad (4.63.)$$

şeklinde tanımlanmıştır.

Çizelge 4.2. İkili Benzerlik Ölçüleri ve Aralık Değerleri

Ölçü	İfadesi	Aralık
S1	$s_{\text{Jaccard}} = \frac{a}{a+b+c}$ (Tanimoto)	[0,1]
S2	$s_{\text{Jaccard-3W}} = \frac{3a}{3a+b+c}$	[0,1]
S3	$s_{\text{Dice}} = \frac{2a}{2a+b+c}$ (Czekanowski, Sorensen)	[0,1]
S4	$s_{\text{Sokal\&Sneath-I}} = \frac{a}{a+2b+2c}$	[0,1]
S5	$s_{\text{Nei\&Lei}} = \frac{2a}{(a+b)+(a+c)}$	[0,1]
S6	$s_{\text{Kulczynski-I}} = \frac{a}{b+c}$	[0,∞)
S7	$s_{\text{Intersection}} = a$	[0,∞)
S8	$s_{\text{Inner Product}} = a+d$	[0,∞)
S9	$s_{\text{Sokal\&Michener}} = \frac{a+d}{a+b+c+d}$ (Simple Matching, Eşleştirme)	[0,1]
S10	$s_{\text{Roger\&Tanimoto}} = \frac{a+d}{a+2(b+c)+d}$	[0,1]

Çizelge 4.2. (Devamı) İkili Benzerlik Ölçüleri ve Aralık Değerleri

Ölçü	İfadesi	Aralık
S11	$s_{Faith} = \frac{a + 0.5d}{a + b + c + d}$	[0,1]
S12	$s_{Gower\&Legendre} = \frac{a + d}{a + 0.5(b + c) + d}$	[0,1]
S13	$s_{Hamann} = \frac{(a + d) - (b + c)}{a + b + c + d}$	[-1,1]
S14	$s_{Dispersion} = \frac{ad - bc}{(a + b + c + d)^2}$	[-1,1]
S15	$s_{Sokal\&Sneath-II} = \frac{2(a + d)}{b + c + 2(a + d)}$	[0,1]
S16	$s_{Sokal\&Sneath-III} = \frac{a + d}{b + c}$	[0, ∞)
S17	$s_{Russell\&Rao} = \frac{a}{a + b + c + d}$	[0,1]
S18	$s_{Baroni\&Urbani\&Buser-I} = \frac{\sqrt{ad} + a}{\sqrt{ad} + a + b + c}$	[0,1]
S19	$s_{Baroni\&Urbani\&Buser-II} = \frac{\sqrt{ad} + a - (b + c)}{\sqrt{ad} + a + b + c}$	[-1,1]
S20	$s_{Phi} = \frac{ad - bc}{\sqrt{abcd}}$	[-1,1]
S21	$s_{Sokal\&Sneath-IV} = \frac{\frac{a}{(a+b)} + \frac{a}{(a+c)} + \frac{d}{(c+d)} + \frac{d}{(b+d)}}{4}$	[0,1]
S22	$s_{Sokal\&Sneath-V} = \frac{ad}{(a+b)(a+c)(b+d)(c+d)^{0.5}}$	[0,1]
S23	$s_{Ochiai-II} = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	[0,1]
S24	$s_{Gower} = \frac{a + d}{\sqrt{(a+b).(a+c).(b+d).(c+d)}}$	[0, ∞)
S25	$s_{Yule-Q} = \frac{ad - bc}{ad + bc}$	[-1,1]
S26	$s_{Yule-w} = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$	[-1,1]
S27	$s_{Pearson\&Heron-II} = \cos(\frac{\pi\sqrt{bc}}{\sqrt{ad} + \sqrt{bc}})$	[-1,1]
S28	$s_{Michael} = \frac{4(ad - bc)}{(a + d)^2 + (b + c)^2}$	[-1,1]

Çizelge 4.2. (Devamı) İkili Benzerlik Ölçüleri ve Aralık Değerleri

Ölçü	İfadesi	Aralık
S29	$S_{\text{Goodman\&Kruskal}} = \frac{\sigma - \sigma'}{2n - \sigma'}$ $\sigma = \max(a, b) + \max(c, d) + \max(a, c) + \max(b, d)$ $\sigma' = \max(a + c, b + d) + \max(a + b, c + d)$	[0,1]
S30	$S_{\text{Anderberg}} = \frac{\sigma - \sigma'}{2n}; \sigma = \max(a, b) + \max(c, d) + \max(a, c) + \max(b, d)$ $\sigma' = \max(a + c, b + d) + \max(a + b, c + d)$	[0,1]
S31	$S_{\text{Peirce}} = \frac{ab + bc}{ab + 2bc + cd}$	[0,1]
S32	$S_{\text{Ample}} = \frac{a(c + d)}{c(a + b)} \text{ (Tarantula)}$	[0, ∞)
S33	$S_{\text{Cosine}} = \frac{a}{\sqrt{(a + b)(a + c)}} \text{ (Ochiai-1, Otsuka)}$	[0,1]
S34	$S_{\text{Sorgenfrei}} = \frac{a^2}{(a + b)(a + c)}$	[0,1]
S35	$S_{\text{McConnaughey}} = \frac{a^2 - bc}{(a + b)(a + c)}$	[-1,1]
S36	$S_{\text{Forbes-I}} = \frac{na}{(a + b)(a + c)}$	[0, ∞)
S37	$S_{\text{Forbes-II}} = \frac{na - (a + b)(a + c)}{n \min\{(a + b), (a + c)\} - (a + b)(a + c)}$	[-1,1]
S38	$S_{\text{Driver\&Kroeber}} = \frac{a}{2} \left(\frac{1}{(a + b)} + \frac{1}{(a + c)} \right) \text{ (Kulczynski-II)}$	[0,1]
S39	$S_{\text{Johnson}} = \frac{a}{a + b} + \frac{a}{a + c}$	[0,2]
S40	$S_{\text{Tarwid}} = \frac{na - (a + b)(a + c)}{na + (a + b)(a + c)}$	[-1,1]
S41	$S_{\text{Simpson}} = \frac{a}{\min\{(a + b), (a + c)\}}$	[0,1]
S42	$S_{\text{Braun\&Banquet}} = \frac{a}{\max\{(a + b), (a + c)\}}$	[0,1]
S43	$S_{\text{Mountford}} = \frac{a}{0.5(ab + ac) + bc}$	[0,1]
S44	$S_{\text{Pearson\&Heron-I}} = \frac{ad - bc}{\sqrt{(a + b).(a + c).(b + d).(c + d)}}$	[-1,1]

Çizelge 4.3. İkili Uzaklık Ölçüleri

Ölçü	İfadesi	Aralık
1. Öklid Uzaklığı	$d = \sqrt{b+c}$	$[0, \infty)$
2. Canberra (Cityblock, Hamming, Kare Öklid, Manhattan, Minkowski)	$d = (b+c)$	$[0, \infty)$
3. Ortalama Manhattan	$d = \frac{b+c}{a+b+c+d}$	$[0,1]$
4. Size Difference Uzaklığı	$d = \frac{(b-c)^2}{(a+b+c+d)^2}$	$[0,1]$
5. Pattern Difference Uzaklığı	$d = \frac{bc}{(a+b+c+d)^2}$	$[0,1]$
6. Variance Uzaklığı	$d = \frac{(b+c)}{4(a+b+c+d)}$	$[0,1]$
7. Lance ve Williams (Bray ve Curtis)	$d = \frac{b+c}{2a+b+c}$	$[0,1]$
8. Chord	$d = \sqrt{2(1 - \frac{a}{\sqrt{(a+b)(a+c)}})}$	$[0,1]$
9. Hellinger	$d = 2\sqrt{1 - \frac{a}{\sqrt{(a+b)(a+c)}}}$	$[0,1]$
10. Yule-Q Uzaklık Ölçüsü	$d = \frac{bc}{ad+bc}$	$[0,1]$

4.2.1. İkili Uzaklık ve Benzerlik Ölçülerinin Teorik Olarak İncelenmesi

İkili uzaklık ve benzerlik ölçüleri sınıflama, kümeleme gibi çalışma alanlarında çok önemli bir yere sahiptir. Bu nedenle ölçünün seçimi sırasında bir takım teorik aksiyomları sağlaması istenebilir. Bu bölümde ölçülerin hem birer benzerlik ölçüsü olarak hem de uzaklık ölçüsü olarak teorik yapıları incelenmiştir. Bu inceleme sadece benzerlik ölçüleri üzerinde yapılmıştır. Her bir benzerlik ölçüsüne karşılık gelen uzaklık ölçüleri elde edilip ayrı ayrı incelenmiştir. Bu amaçla ölçüleri farklı koşullarla her iki açıdan inceleyebilmek için her bir benzerlik ölçüsünü uzaklık ölçüsüne çevirmek amacıyla bir takım dönüşümler yapılmıştır. Bu dönüşümler benzerlik ölçüsünün aralığı,

$$[0,1] \text{ ise } d = 1 - s, \quad (4.64.)$$

$$[-1,1] \text{ ise } d = (1 - s)/2, \quad (4.65.)$$

$$[0, \infty) \text{ ise } d = 1/(s+1), \quad (4.66.)$$

olacak şekilde uygulanmıştır (Baulieu, 1989).

4.2.2. Ölçü Seçimine İlişkin Kabul Edilebilirlik Koşulları

Ölçü seçimine ilişkin birçok koşul veya kriter ortaya atılmıştır. Vegelius ve Janson (1981), Hubalek (1982), Baulieu (1989), Jackson ve ark. (1989), Dekang Lin(1998), Santini ve Jain (1999) ölçüler üzerinde teorik incelemelerde bulunup, çalışmalar yürütmüşlerdir. Bu çalışmalara dayanarak benzerlik ölçülerinin incelenmesi amaçlı 15 koşul derlenmiştir. Uzaklık ölçülerinin incelenmesi amaçlı ise Baulieu (1989) tarafından önerilen 12 koşula bağlı olarak inceleme yapılmıştır.

4.2.2.1. Benzerlik Ölçülerinin Teorik Olarak İncelenmesi

Aşağıda verilen koşullardan 1-6 arası koşullar Janson ve Vegelius (1981), koşul 1 ve 7-15 arası koşullar ise Hubalek (1982) tarafından önerilen koşullardır.

Koşul 1: Simetri

$s(A,B)$ değeri $s(B,A)$ değerine eşit olmalıdır. Diğer bir deyişle, benzerlik ölçüsü b ve c parametrelerine göre simetrik yapıda olmalıdır.

Koşul 2: Homojenlik

Eğer a,b,c ve d parametreleri aynı sabitle çarpılırsa, bu işlem benzerlik ölçüsünün değerini etkilememelidir. Yani, ölçü değeri, oranlar sabit kalmak koşulu ile birim sayısından bağımsızdır.

Koşul 3: 'd' Parametresinden Bağımsız Olma

Genel kanı güçlü ölçülerin negatif uyumu ifade eden 'd' parametresinden bağımsız olması gerektiğidir. Çünkü ölçü değeri bir arada çok nadir bulunan türler arasında yüksek bir birliktelik (benzerlik) ifade eden yüksek değer alma eğilimi gösterebilecektir.

Koşul 4: 'a = 0' olduğunda benzerlik değeri 0'a eşit olmalıdır.

Eğer koşul 3 makul olarak düşünülürse, bunun geçerli sonucu olarak, benzerlik katsayısı ancak ve ancak pozitif uyum (a) sıfır olduğunda minimum değerini alacaktır. Bu da bize, örneğin iki türün asla birlikte bulunmadığı, hiçbir ortaklıklarının olmadığını ifade edecektir.

Koşul 5: 'b = c = 0' olduğunda benzerlik değeri 1'e eşit olmalıdır.

Maksimum birliktelik iki tür daima birlikte görülüyorsa oluşmalıdır. Diğer bir deyişle b ve c parametreleri sıfır (0) olmalıdır.

Koşul 6: E-Coefficient özelliğine sahip olmalıdır.

Eğer bir benzerlik (ya da birliktelik) ölçüsü, Öklid uzayında normalleştirilmiş vektörler arasındaki skaler çarpımın gereksinimlerini sağlıyorsa, e-coefficient olarak tanımlanır (Janson ve Vegelius, 1981). Bu özelliğe sahip bir 's' benzerlik ölçüsü, A, B ve C birimleri için aşağıdaki özellikleri sağlar:

$$a) s_i(A, B) = s_i(B, A)$$

$$b) s_i(A, A) = 1$$

$$c) s_i(A, B) \leq 1$$

$$d) \text{Eğer } s_i(A, C) = s_i(B, C) \text{ ise } s_i(A, B) = 1 \text{ 'dir.}$$

e) Bu katsayıya bağlı elde edilen korelasyon matrisi pozitif tanımlı bir matristir.

Koşul 7: Eş birimler arasındaki benzerlik değeri 1 olmalıdır.

$$s_i(A, A) = 1$$

Koşul 8: Bir birimin kendisinden farklı herhangi bir birimle arasındaki benzerlik değeri, kendisiyle olan benzerlik değerinden daha küçük olmalıdır.

$$s_i(A, B) \leq s_i(A, A) = \text{maksimum}(s_i)$$

Koşul 9: Koşul 5'e benzer olarak benzerlik ölçüsü maksimum değerini $b=c=0$ iken almalıdır.

$$b = c = 0 \{a \neq 0, d \neq 0\} \text{ iken } s_i(A, B) = \text{Maksimum}$$

Koşul 10: Koşul 4'e benzer olarak, benzerlik ölçüsü minimum değerini $a=d=0$ iken almalıdır.

$$a = d = 0 \{b \neq 0, c \neq 0\} \text{ iken } s_i(A, B) = \text{Minimum}$$

Koşul 11: Bu koşul pozitif ve negatif birliktelik arasındaki ayrıma dayanır. Koşula göre, $a' = \frac{(a+b)(a+c)}{n}$ iken $s_i(a > a') > s_i(a < a')$ eşitsizliği geçerlidir (Kendall ve Stuart, 1967; Pielou, 1969).

Aşağıda verilmiş diğer koşullar, yukarıda verilen koşullar kadar önemli olmasa da, en iyi ölçünün seçiminde faydalı olabilecek koşullardır.

Koşul 12: Uyum açısından, benzerlik değeri üç aralıktan birine sahip olmalıdır:

$$-1 \leq s_i \leq +1, 0 \leq s_i \leq +1, 0 \leq s_i < \infty$$

Koşul 13: s_i ölçüsü,

Eğer $\max s_i = s_i(b = c = 0) > s_i(b \text{ ya } c = 0)$ ise kesin birliktelik katsayısı,

Eğer $\max s_i = s_i(b = c = 0) = s_i(b \text{ ya } c = 0)$ ise tam birliktelik katsayısı

olarak adlandırılır.

Koşul 14:

Eğer,

$$s_i(a+1, b-1, c-1, d+1) - s_i(a, b, c, d) = s_i(a+2, b-2, c-2, d+2) - s_i(a+1, b-1, c-1, d+1)$$

ise lineer yapıdadır denilebilir. Diğer bir deyişle eğer,

$$s_i(a+1) - s_i(a) \neq s_i(a+2) - s_i(a+1) \text{ ise lineer değildir.}$$

Koşul 15: Ölçünün $a=0$ durumuna karşı hassasiyeti gözlenmelidir.

$s_i(a = 0, b, c, d)$ durumunda ölçü değerinde önemli bir değişim (0.15 ve üzeri) varsa ölçü $a=0$ değerine karşı hassastır (tanımsız olduğu durumlar da gözlemlenebilir).

Çizelge 4.4. Benzerlik Ölçüleri İçin Elde Edilen Sonuçlar

Ölçü/Koşul	(K1)	(K2)	(K3)	(K4)	(K5)	(K6)	(K7)	(K8)	(K9)	(K10)	(K11)	(K12)	(K13)	(K14)	(K15)
S1	E	E	E	E	E	E	E	E	E	E	E	E	Mutlak	E	E
S2	E	E	E	E	E	E	E	E	E	E	E	E	Mutlak	E	E
S3	E	E	E	E	E	E	E	E	E	E	E	E	Mutlak	E	E
S4	E	E	E	E	E	E	E	E	E	E	E	E	Mutlak	H	E
S5	E	E	E	E	E	E	E	E	E	E	E	E	Mutlak	E	E
S6	E	E	E	E	H	H	H	E	E	E	H	E	Mutlak	H	E
S7	E	H	E	E	H	H	H	E	E	E	H	E	Tam	E	E
S8	E	H	H	H	H	H	H	E	E	E	H	E	Tam	E	H
S9	E	E	H	H	E	E	E	E	E	E	E	E	Mutlak	E	E
S10	E	E	H	H	E	E	E	E	E	E	E	E	Mutlak	H	E
S11	E	E	H	H	H	H	H	H	E	E	E	E	Mutlak	H	E
S12	E	E	H	H	H	H	E	E	E	E	E	E	Mutlak	E	E
S13	E	E	H	H	E	E	E	E	E	E	E	E	Mutlak	E	E
S14	E	E	H	H	H	H	H	H	H	H	E	E	Mutlak	H	E
S15	E	E	H	H	E	H	E	E	E	E	E	E	Mutlak	H	E
S16	E	E	H	H	H	H	H	E	E	E	H	E	Mutlak	H	E
S17	E	E	H	E	H	H	H	E	E	E	E	E	Mutlak	E	E
S18	E	E	H	E	E	H	E	E	E	E	E	E	Mutlak	H	E
S19	E	E	H	H	E	H	E	E	E	E	E	E	Mutlak	H	E
S20	E	E	H	H	H	H	H	H	H	H	E	E	Tam	H	H
S21	E	E	H	H	E	H	E	E	E	E	E	E	Mutlak	E	E
S22	E	E	H	E	E	E	E	E	E	E	E	E	Mutlak	H	E
S23	E	E	H	E	E	H	E	E	E	E	E	E	Mutlak	H	E
S24	E	H	H	H	H	H	H	E	E	E	H	E	Mutlak	E	E
S25	E	E	H	H	E	H	E	E	E	E	E	E	Mutlak	H	E
S26	E	E	H	H	E	H	E	E	E	E	E	E	Mutlak	H	E
S27	E	E	H	H	H	H	E	E	E	E	E	E	Tam	H	E
S28	E	E	H	H	H	H	H	E	H	H	E	E	Mutlak	H	E
S29	E	E	H	H	H	H	E	E	E	H	E	E	Mutlak	E	E
S30	E	E	H	H	H	H	E	E	E	H	E	E	Mutlak	E	E
S31	H	E	H	H	H	H	H	H	H	H	E	E	Mutlak	E	E
S32	H	E	H	E	H	H	H	H	E	E	H	E	Tam	H	E
S33	E	E	E	E	E	E	E	E	E	E	E	E	Mutlak	E	E
S34	E	E	E	E	E	H	E	E	E	E	E	E	Mutlak	H	E
S35	E	E	E	H	E	H	E	E	E	E	E	E	Mutlak	E	E
S36	E	E	H	E	H	H	H	H	E	E	H	E	Mutlak	E	E
S37	E	E	H	H	E	H	E	E	E	H	E	E	Mutlak	E	E
S38	E	E	E	E	E	H	E	E	E	E	E	E	Tam	E	E
S39	E	E	E	E	H	H	H	E	E	E	H	E	Mutlak	E	E
S40	E	E	H	H	H	H	H	H	E	E	E	E	Mutlak	H	E
S41	E	E	E	E	E	H	E	E	E	E	E	E	Mutlak	E	E
S42	E	E	E	E	E	E	E	E	E	E	E	E	Mutlak	E	E
S43	E	H	E	E	H	H	H	E	E	E	E	E	Mutlak	H	E
S44	E	E	H	H	E	E	E	E	E	E	E	E	Mutlak	E	E

Çizelge 4.4 incelendiğinde elde edilebilecek sonuçlar şöyle özetlenebilir;

Braun-Blanquet, Cosine, Dice, Jaccard, Jaccard-3W ve Nei&Li ölçüleri tüm koşulları sağlamaktadır.

Sokal&Sneath-1, Driver&Kroeber, Forbes-2, Hamann, McConnaughey, Ochiai-2, Roger&Tanimoto, Sokal&Michener, Simpson ve Sokal&Sneath-5, Sorgenfrei ölçüleri ise yukarıda verilen ölçülere birer alternatif olabilecek kadar iyi ölçüler olarak karşımıza çıkmaktadır. Çünkü bu ölçüler de hemen hemen tüm koşulları sağlamaktadır. Diğer taraftan, Ample, Faith, Michael, Sokal&Sneath-3, Dispersion, Peirce ve Phi benzerlik ölçüleri ise sadece birkaç koşulu sağlamaktadır.

4.2.2.2. Uzaklık Ölçülerinin Teorik Olarak İncelenmesi

Baileu (1989) tarafından yürütülen çalışmada 12 koşul ortaya atılmış ve bu ölçüler doğrultusunda bazı ölçüleri incelenmiştir.

Çoğu katsayı, bu a, b, c ve d değerlerine bağlı olarak tanımlanır. Bu nedenle, aşağıdaki bir katsayı uzayı tanımlanabilir:

$$D : (Z^+)^4 - \{(0,0,0,0)\} \longrightarrow (Z^+)$$

Tüm $(a, b, c, d) \in Z^+ - \{(0,0,0,0)\}$ için aşağıdaki özellikler sağlanır:

Koşul1 : $D, (a, b, c, d)$ haricinde tüm negatif olmayan (a, b, c, d) dörtlüleri için tanımlıdır.

Koşul2 : $0 \leq D(a, b, c, d) \leq 1$, tüm (a, b, c, d) için

Koşul3 : $D(a, 0, 0, 0) = 0$, tüm $a > 0$ değerleri için

Koşul4 : $D(a, b, c, d) = 1$, bazı (a, b, c, d) için

Koşul5 : $D(a+1, b, c, d) \leq D(a, b, c, d)$

Koşul6 : $D(a, b+1, c, d) \geq D(a, b, c, d)$

Koşul7 : $D(a, b, c, d+1) \leq D(a, b, c, d)$

Koşul8 : $D(a, b, c, d) = D(a, c, b, d)$

Koşul9 : D , pay ve paydası lineer olan rasyonel bir fonksiyondur. Bu nedenle aşağıdaki formda yazılabilir:

$$D(a, b, c, d) = \frac{\alpha'a + \beta'b + \lambda'c + \delta'd + \varepsilon'}{\alpha a + \beta b + \lambda c + \delta d + \varepsilon}; \text{ sabit } \alpha', \beta', \lambda', \delta', \varepsilon' \in \mathbb{R}$$

Koşul 10 : $D(0, b, c, 0) = 1; \forall b, c \in Z^+$

Koşul 11 : $D(a+1, b, c, d) < D(a, b+1, c, d); \forall a, b, c, d \in Z^+$

Koşul 12 : $D(a, 0, 0, d) = 0; \forall a, d \in Z^+$

Verilen 44 benzerlik ölçüsü tanım aralıklarına göre (4.64.), (4.65.) ve (4.66.) yardımıyla uzaklık ölçülerine dönüştürülmüştür. Bu şekilde elde edilen uzaklık ölçülerinin yukarıda verilen 12 koşulu sağlayıp sağlamadıklarına ilişkin sonuçlar Çizelge 4.5’te verilmiştir.

Çizelge 4.5. Uzaklık Ölçüleri İçin Elde Edilen Sonuçlar

Ölçü/Koşul	(K1)	(K2)	(K3)	(K4)	(K5)	(K6)	(K7)	(K8)	(K9)	(K10)	(K11)	(K12)
S1	H	E	E	E	E	E	E	E	E	E	E	E
S2	H	E	E	E	E	E	E	E	E	E	E	E
S3	H	E	E	E	E	E	E	E	E	E	E	E
S4	H	E	E	E	E	E	E	E	E	E	E	E
S5	H	E	E	E	E	E	E	E	E	E	E	E
S6	H	E	E	E	E	E	E	E	E	E	E	E
S7	H	E	H	E	E	E	E	E	E	E	E	H
S8	H	E	H	E	E	E	E	E	E	E	E	H
S9	E	E	E	E	E	E	E	E	E	E	E	E
S10	E	E	E	E	E	E	E	E	E	E	E	E
S11	E	E	E	E	E	E	E	E	E	E	E	H
S12	E	E	E	E	E	E	E	E	E	E	E	H
S13	E	E	E	E	E	E	E	E	E	E	E	E
S14	E	E	H	E	E	H	E	E	H	H	E	H
S15	E	E	E	E	E	E	E	E	E	E	E	E
S16	E	E	E	E	E	E	E	E	E	E	E	E
S17	E	E	E	E	E	E	H	E	E	E	E	H
S18	H	E	E	E	E	E	E	E	H	E	E	E
S19	H	E	E	E	E	E	E	E	H	H	E	E
S20	H	E	H	E	E	H	E	E	H	H	E	H
S21	H	E	H	E	E	E	E	E	H	E	E	E
S22	H	E	H	E	H	E	H	E	H	E	E	E
S23	H	E	H	E	H	E	E	E	H	E	E	E
S24	H	E	H	E	E	E	E	E	H	H	E	H
S25	H	E	H	E	H	H	E	E	H	E	E	E
S26	H	E	H	E	H	H	E	E	H	E	E	E
S27	H	E	H	E	H	H	H	E	H	E	H	E
S28	E	E	H	E	E	H	E	E	H	H	E	H
S29	E	E	E	E	E	H	H	E	E	H	H	H
S30	E	E	E	E	E	H	H	E	E	H	H	H
S31	H	E	H	E	E	E	H	H	H	H	E	H
S32	H	E	H	E	H	E	E	H	H	E	E	E
S33	H	E	E	E	E	E	E	E	H	E	H	E
S34	H	E	E	E	E	E	E	E	H	E	E	E
S35	H	E	E	E	E	E	E	E	H	E	E	E
S36	H	E	H	E	E	E	E	E	H	E	E	H
S37	H	E	H	E	E	H	H	E	H	H	H	H
S38	H	E	E	E	E	E	E	E	H	E	E	E
S39	H	E	E	E	E	E	E	E	H	H	E	E
S40	H	E	H	E	E	H	E	E	H	E	E	H
S41	H	E	E	E	E	H	E	E	E	E	H	E
S42	H	E	E	E	E	H	E	E	E	E	H	E
S43	H	E	H	E	E	E	E	E	H	E	E	H
S44	H	E	H	E	E	E	E	E	H	E	E	E

Çizelge 4.5 incelendiğinde;

Hamann, Roger&Tanimoto, Sokal&Michener, Sokal&Sneath-2 ve Sokal&Sneath-3 uzaklık ölçülerinin tüm koşulları sağladığı görülür. Dice, Faith, Gower&Legendre, Jaccard, Jaccard3W, Kulczynski-1, Nei&Li, Sokal&Sneath-1 ve Sokal&Sneath-2 ölçüleri de yukarıda verilen ölçülerden sonra hemen hemen tüm koşulları sağlamaları nedeniyle tercih edilebilecek ölçüler olarak karşımıza çıkmaktadır. Bu koşullar doğrultusunda, en kötü sonuçlar Goodman&Kruskal, Pearson&Heron-2, Simpson, Phi ve Forbes-2 ölçülerinde elde edilmiştir.

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ

Bu bölümde, ikili uzaklık ve benzerlik ölçülerinin aşamalı kümeleme analizi üzerindeki etkisi 37 ikili benzerlik ve 10 ikili uzaklık ölçüsü kullanılarak araştırılacaktır. İkili uzaklık ve benzerlik ölçüleri kullanılarak oluşturulan uzaklık matrislerine aşamalı kümeleme analizi uygulanacaktır. Elde edilen kümelerin geçerliliği için kriter olarak kofenetik korelasyon katsayısı kullanılacaktır. Ölçülerin etkisi gerçek veri ve farklı yapıda türetilen yapay dört veri kümesi üzerinde incelenecektir.

5.1. Aşamalı Kümeleme Analiz Yöntemlerinde İkili Uzaklık ve Benzerlik Ölçülerinin Etkisi

Uzaklık ölçüsünün aşamalı kümeleme yöntemleri üzerindeki etkisi öncelikle gerçek veri kümesi üzerinde incelenecektir. Bu amaçla, önceki bölümlerde verilen bağlantı yöntemlerinden tek bağlantı, tam bağlantı, ortalama bağlantı ve ağırlıklı ortalama bağlantı yöntemleri ayrı ayrı kullanılacaktır. Bağlantı yöntemleri uygulanırken kullanılacak uzaklık ve benzerlik ölçüleri Çizelge 5.1 ve 5.2 de verilmiştir.

Çizelge 5.1. Aşamalı Kümeleme Yöntemlerinde Etkisi İncelenen Benzerlik Ölçüleri

1) $s_{Anderberg} = \frac{\sigma - \sigma'}{2n}$, $\sigma = \max(a,b) + \max(c,d) + \max(a,c) + \max(b,d)$, $\sigma' = \max(a+c, b+d) + \max(a+b, c+d)$	6) $s_{Michael} = \frac{4(ad-bc)}{(a+d)^2 + (b+c)^2}$
2) $s_{Baroni\&Urbani\&Buser-I} = \frac{\sqrt{ad} + a}{\sqrt{ad} + a + b + c}$	7) $s_{Nei\&Lei} = \frac{2a}{(a+b) + (a+c)}$
3) $s_{Baroni\&Urbani\&Buser-II} = \frac{\sqrt{ad} + a - (b+c)}{\sqrt{ad} + a + b + c}$	8) $s_{Ochiai-II} = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
4) $s_{Braun\&Banquet} = \frac{a}{\max\{(a+b), (a+c)\}}$	9) $s_{Cosine} = \frac{a}{\sqrt{(a+b)(a+c)}} (Ochiai - 1, Otsuka)$
5) $s_{McConnaughey} = \frac{a^2 - bc}{(a+b)(a+c)}$	10) $s_{Dice} = \frac{2a}{2a + b + c} (Czekanowski)$

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

Çizelge 5.1. (Devamı) Aşamalı Kümeleme Yöntemlerinde Etkisi İncelenen Benzerlik Ölçüleri

11) $s_{Dispersion} = \frac{ad - bc}{(a+b+c+d)^2}$	25) $s_{Pearson\&Heron-II} = \cos(\frac{\pi\sqrt{bc}}{\sqrt{ad} + \sqrt{bc}})$
12) $s_{Faith} = \frac{a+0.5d}{a+b+c+d}$	26) $s_{Roger\&Tanimoto} = \frac{a+d}{a+2(b+c)+d}$
13) $s_{Forbes-I} = \frac{na}{(a+b)(a+c)}$	27) $s_{Russell\&Rao} = \frac{a}{a+b+c+d}$
14) $s_{Forbes-II} = \frac{na - (a+b)(a+c)}{n\min\{(a+b), (a+c)\} - (a+b)(a+c)}$	28) $s_{Simpson} = \frac{a}{\min\{(a+b), (a+c)\}}$
15) $s_{Goodman\&Kruskal} = \frac{\sigma - \sigma'}{2n - \sigma'}$, $\sigma = \max(a,b) + \max(c,d) + \max(a,c) + \max(b,c)$, $\sigma' = \max(a+c, b+d) + \max(a+b, c+d)$	29) $s_{Sokal\&Michener} = \frac{a+d}{a+b+c+d}$ (Simple Matching)
16) $s_{Gower} = \frac{a+d}{\sqrt{(a+b).(a+c).(b+d).(c+d)}}$	30) $s_{Sokal\&Sneath-I} = \frac{a}{a+2b+2c}$
17) $s_{Gower\&Legendre} = \frac{a+d}{a+0.5(b+c)+d}$	31) $s_{Sokal\&Sneath-II} = \frac{2(a+d)}{b+c+2(a+d)}$
18) $s_{Hamann} = \frac{(a+d) - (b+c)}{a+b+c+d}$	32) $s_{Sokal\&Sneath-IV} = \frac{\frac{a}{(a+b)} + \frac{a}{(a+c)} + \frac{d}{(c+d)} + \frac{d}{(b+d)}}{4}$
19) $s_{InnerProduct} = a+d$	33) $s_{Sokal\&sneath-V} = \frac{ad}{[(a+b).(a+c).(b+d).(c+d)]^{0.5}}$
20) $s_{Jaccard} = \frac{a}{a+b+c}$ (Tanimoto)	34) $s_{Sorgenfrei} = \frac{a^2}{(a+b)(a+c)}$
21) $s_{Jaccard-3W} = \frac{3a}{3a+b+c}$	35) $s_{Tarwid} = \frac{na - (a+b)(a+c)}{na + (a+b)(a+c)}$
22) $s_{Johnson} = \frac{a}{a+b} + \frac{a}{a+c}$	36) $s_{Yule-Y} = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$
23) $s_{kulczynski-II} = a/2(\frac{1}{(a+b)} + \frac{1}{(a+c)})$ (Driver&Kroeber)	37) $s_{Yule-Q} = \frac{ad - bc}{ad + bc}$
24) $s_{Pearson\&Heron-I} = \frac{ad - bc}{\sqrt{(a+b).(a+c).(b+d).(c+d)}}$	

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

Çizelge 5.2. Aşamalı Kümeleme Yöntemlerinde Etkisi İncelenen Uzaklık Ölçüleri

1) $d_{Canberra} = (b+c)$ (Kare Öklid, Manhattan, Cityblock, Minkowski, Hamming)	6) $d_{PatternDifference} = \frac{4bc}{(a+b+c+d)^2}$
2) $d_{Chord} = \sqrt{2(1 - \frac{a}{\sqrt{(a+b)(a+c)}})}$	7) $d_{ShapeDifference} = \frac{n(b+c) - (b-c)^2}{(a+b+c+d)^2}$
3) $d_{Hellinger} = 2\sqrt{1 - \frac{a}{\sqrt{(a+b)(a+c)}}}$	8) $d_{SizeDifference} = \frac{(b+c)^2}{(a+b+c+d)^2}$
4) $d_{Lance\&Williams} = \frac{b+c}{2a+b+c}$	9) $d_{Varyans} = \frac{(b+c)}{4(a+b+c+d)}$
5) $d_{Oklid} = \sqrt{b+c}$	10) $d_{OrtalamaManhattan} = \frac{b+c}{a+b+c+d}$

Çalışmada kullanılacak veri kümesi biyolojik bir veri kümesi olup üç farklı toprak türünde yetişebilen (kurak, tuzcul ve kumul) 361 bitki türünü içermektedir. Bu bitki türlerinin belirlenen 330 bölge içinde bulunup-bulunmama durumları araştırılmıştır. Elde edilen sonuçlar var-yok şeklinde ikili olarak kodlanmış ve bu değerlere göre bölgeler arası benzerlik araştırılmıştır.

Yapay olarak üretilen veri kümeleri ise tamamen rastgele ve sırasıyla %90, %50 ve %10 oranlarında ‘1’ değerini içermektedir. Örneğin, %10, her bir birimin araştırılan 100 değişkenden 10’unda söz konusu niteliği taşıdığı anlamına gelir. Her bir veri kümesi 100 değişkene bağlı 100 birim içermektedir. Çalışmada tekrar sayısı ise 1000 olarak belirlenmiştir. Dolayısıyla her bir veri 1000 adet türetilip, aşamalı kümeleme analizi bağlantı yöntemleri ayrı ayrı uygulanıp elde edilen kofenetik korelasyon katsayı değerlerinin ortalaması alınmıştır. Tanımlanan veri kümeleri aşağıdaki gibidir:

V1. Tamamen rastgele ikili veri kümesi

1	1	0	1	1	1	0	.	.	.	.	1	0	0	1	0	0	0	$\sum X_1 = 20, \sum X_0 = 80$
0	0	1	0	1	1	0	.	.	.	.	0	0	1	0	0	1	0	$\sum X_1 = 30, \sum X_0 = 70$
⋮																		⋮
1	0	0	0	0	0	1	.	.	.	.	1	0	1	0	1	0	1	⋮

Şekil 5.1. Tamamen rastgele veri kümesi

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

V2. %90 oranında 1, %10 oranında 0 içeren ikili veri kümesi

1	0	1	1	0	1	0	.	.	.	.	1	1	0	1	1	0	1		$\sum X_1 = 90, \sum X_0 = 10$
1	1	0	0	0	1	0	.	.	.	.	1	0	1	0	1	0	0		$\sum X_1 = 90, \sum X_0 = 10$
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1	0	1	0	1	0	1	.	.	.	.	1	0	1	1	0	1	1		.

Şekil 5.2. %90 oranında '1' içeren veri kümesi

V3. %10 oranında 1, %90 oranında 0 içeren ikili veri kümesi

1	0	0	1	0	0	0	.	.	.	.	0	1	0	0	0	1	0		$\sum X_1 = 10, \sum X_0 = 90$
0	0	1	0	0	1	0	.	.	.	.	1	0	1	0	1	0	1		$\sum X_1 = 10, \sum X_0 = 90$
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
0	1	0	1	0	0	0	.	.	.	.	0	0	0	1	0	0	0		.

Şekil 5.3. %10 oranında '1' içeren veri kümesi

V4. %50 oranında 1, %50 oranında 0 içeren ikili veri kümesi

1	0	0	1	1	0	1	.	.	.	.	0	1	0	1	0	1	0		$\sum X_1 = 50, \sum X_0 = 50$
0	1	0	1	1	0	0	.	.	.	.	1	0	1	0	1	0	1		$\sum X_1 = 50, \sum X_0 = 50$
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
0	1	0	0	1	0	0	.	.	.	.	0	0	0	1	0	1	0		.

Şekil 5.4. %50 oranında '1' içeren veri kümesi

Yöntem uygulanırken öncelikle birimlere ilişkin uzaklık matrisi oluşturulur. Bu amaçla Çizelge 5.1'de yer alan ölçülerden benzerlik ölçüleri, uzaklık ölçülerine çevrilir.

Benzerlik ölçüsünün aralığı,

$$[0,1] \text{ ise } d=1-s \quad (5.1.)$$

$$[-1,1] \text{ ise } d=(1-s)/2 \quad (5.2.)$$

$$[0,\infty) \text{ ise } d=1/(s+1) \quad (5.3.)$$

şeklinde uzaklık ölçüsüne dönüştürülmüştür (Bailieu, 1989).

Yukarıda verilen dönüşümler sonucu elde edilen uzaklık matrisleri üzerinde dağılım aralığı $[0,1]$ olacak şekilde ölçeklendirme yapılmıştır. Bu işlemler sonucu elde edilen uzaklık matrislerine aşamalı kümeleme analizi uygulanmıştır.

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

Aşamalı kümeleme analizi yöntemlerinde uzaklık ölçülerinin kümelemeye etkisi incelenirken küme geçerliğini incelemek amacıyla Bölüm 3’de verilen içsel kriterlerden kofenetik korelasyon katsayısı kullanılmıştır. Bu katsayı, kümelenme öncesi birimler arasındaki uzaklık değerleri ile kümeleme sonrası elde edilen uzaklık değerleri arasındaki ilişkinin ölçüsü olup, daha büyük değer alması daha geçerli ve iyi bir kümeleme anlamına gelecektir.

5.1.1. Tamamen Rastgele Üretilmiş Veri Kümesinin Aşamalı Kümeleme Analizi Üzerindeki Etkisinin Araştırılması

Tamamen rastgele üretilmiş 100 değişken ve 100 birimden oluşan veri kümesine ikili uzaklık ölçüleri kullanılarak aşamalı kümeleme analizi uygulanmıştır. Elde edilen kümelenme yapısının geçerliliğini ölçmek üzere kofenetik korelasyon katsayıları hesaplanmıştır. Farklı bağlantı yöntemlerinde elde edilen katsayı değerleri Çizelge 5.3’de verilmiştir.

Çizelge 5.3 incelendiğinde en yüksek kofenetik korelasyon katsayı değerleri tüm ölçüler için ortalama bağlantı yönteminde elde edilmiştir.

Çizelge 5.3. Tamamen Rastgele Veri Kümesi İçin Çalışmada Kullanılan Ölçülere Bağlı Elde Edilen Kofenetik Korelasyon Katsayıları

	Ölçü	Tek	Tam	Ortalama	Ağırlıklı
1	Anderberg	0,27597	0,3944	0,46947	0,4501
2	B.Urbani Buser-1	0,24201	0,25626	0,37392	0,34912
3	B.Urbani Buser-2	0,19559	0,27048	0,39709	0,36592
4	Braun-Blanquet	0,58533	0,3991	0,62992	0,59556
5	Canberra	0,1288	0,26584	0,35401	0,34529
6	Chord	0,40388	0,25697	0,47438	0,41273
7	Dice	0,42184	0,24917	0,48203	0,41683
8	Dispersion	0,13411	0,26718	0,35398	0,34533
9	Faith	0,32162	0,25849	0,41002	0,36885
10	Forbes-1	0,10144	0,20382	0,30899	0,30582
11	Forbes2	0,12968	0,25983	0,35537	0,34665

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

Çizelge 5.3. (Devamı) Tamamen Rastgele Veri Kümesi İçin Çalışmada Kullanılan Ölçülere Bağlı Elde Edilen Kofenetik Korelasyon Katsayıları

	Ölçü	Tek	Tam	Ortalama	Ağırlıklı
12	Goodman&Kruskal	0,25935	0,39578	0,46305	0,44612
13	Gower	0,10843	0,18209	0,28168	0,27692
14	Gower&Legendre	0,11869	0,24624	0,33782	0,32936
15	Hamann	0,1288	0,26584	0,35401	0,34518
16	Hellinger	0,40388	0,25697	0,47438	0,41273
17	Innerproduct	0,13359	0,24968	0,33641	0,33347
18	Jaccard	0,42571	0,26872	0,49218	0,43042
19	Jaccard3W	0,41942	0,23788	0,47613	0,40874
20	Johnson	0,37577	0,23759	0,44928	0,38587
21	Kulczynski-2	0,37577	0,2376	0,44926	0,38605
22	Lance&Williams	0,42184	0,24917	0,48201	0,41679
23	Mcconnaughey	0,37577	0,2376	0,44926	0,38596
24	Michael	0,12516	0,27309	0,34746	0,33877
25	Nei&Li	0,42184	0,24917	0,48203	0,41683
26	Cosine	0,40002	0,24265	0,46585	0,4036
27	Ochiai-2	0,083045	0,15749	0,25425	0,25068
28	Ort. Manhattan	0,1383	0,28126	0,36743	0,35864
29	Öklid	0,092684	0,18133	0,27798	0,2748
30	Pattern Difference	0,12282	0,23909	0,32945	0,32034
31	Pearson&Heron-1	0,13449	0,2672	0,35414	0,34543
32	Pearson&Heron-2	0,13054	0,26984	0,35132	0,34279
33	Roger&Tanimoto	0,14049	0,28663	0,37112	0,36238
34	Russell&Rao	0,66101	0,32229	0,70304	0,61308
35	Shape Difference	0,13957	0,26597	0,35365	0,3446
36	Simpson	0,15547	0,20927	0,32938	0,30129
37	Size Difference	0,11196	0,23849	0,32896	0,32037
38	Sokal&Michener	0,1288	0,26584	0,35413	0,34539
39	Sokal&Sneath-1	0,42893	0,28555	0,50115	0,4412
40	Sokal&Sneath-2	0,11869	0,24624	0,33782	0,32936
41	Sokal&Sneath-4	0,13449	0,26728	0,3541	0,34538
42	Sokal&Sneath-5	0,15257	0,29845	0,37777	0,36914
43	Sorgenfrei	0,40528	0,27683	0,48081	0,42179

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

Çizelge 5.3. (Devamı) Tamamen Rastgele Veri Kümesi İçin Çalışmada Kullanılan Ölçülere Bağlı Elde Edilen Kofenetik Korelasyon Katsayıları

	Ölçü	Tek	Tam	Ortalama	Ağırlıklı
44	Tarwid	0,089367	0,22221	0,33028	0,32083
45	Varyans	0,1288	0,26584	0,35392	0,3453
46	Yule-Q	0,12541	0,27294	0,34751	0,33891
47	Yule-Y	0,13474	0,2669	0,35435	0,3457

Her bir bağlantı yönteminde en iyi kümelenme sonuçlarının elde edildiği 5 ölçü Çizelge 5.4’de verilmiştir. Tek, Ortalama ve Ağırlıklı Ortalama bağlantı yöntemlerinde Russell&Rao en iyi yapıyı ortaya koyan ölçü iken Tam bağlantı yönteminde en yüksek kofenetik korelasyon katsayı değerleri Braun&Blanquet ölçüsü ile elde edilmiştir.

Çizelge 5.4. Tamamen Rastgele Üretilmiş Veri Kümesi İçin Bağlantı Yöntemlerine Göre En Yüksek Kofenetik Korelasyon Katsayı Değerine Sahip 5 Ölçü

Yöntem/Sıra	1	2	3	4	5
Tek	R&Rao	B-Blanquet	S-Sneath-1	Jaccard	Dice
Tam	B-Blanquet	G&Kruskal	Anderberg	R&Rao	S-Sneath-5
Ortalama	R&Rao	B-Blanquet	S-Sneath-1	Jaccard	Dice
A.Ortalama	R&Rao	B-Blanquet	Anderberg	G&Kruskal	S-Sneath-1

5.1.2. Yüzde 90 Oranında ‘1’ den Oluşan Veri Kümesinin Aşamalı Kümeleme Analizi Üzerindeki Etkisinin Araştırılması

Yüzde 90 oranında ‘1’ içeren 100 değişken ve 100 birimden oluşan veri kümesine ikili uzaklık ölçüleri kullanılarak aşamalı kümeleme analizi uygulanmıştır. Elde edilen kümelenme yapısının geçerliliğini ölçmek üzere kofenetik korelasyon katsayıları hesaplanmıştır. Farklı bağlantı yöntemlerinde elde edilen katsayı değerleri Çizelge 5.5’de verilmiştir.

Çizelge 5.5 incelendiğinde ölçülerin tamamında en yüksek kofenetik korelasyon katsayı değerini veren bağlantı yöntemi, ortalama bağlantı yöntemi olmuştur.

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

Çizelge 5.5. Yüzde 90 Oranında ‘1’ İçeren Veri Kümesi İçin Çalışmada Kullanılan Ölçülere Bağlı Elde Edilen Kofenetik Korelasyon Katsayıları

	Ölçü	Tek	Tam	Ortalama	Ağırlıklı
1	Anderberg	0,1483	0,1507	0,1788	0,1644
2	B.Urbani Buser-1	0,18941	0,31715	0,3672	0,35719
3	B.Urbani Buser-2	0,18941	0,31715	0,3672	0,35719
4	Braun-Blanquet	0,2261	0,36335	0,4148	0,40456
5	Canberra	0,2261	0,36335	0,416	0,4048
6	Chord	0,24206	0,37917	0,43181	0,42093
7	Dice	0,2261	0,36335	0,4148	0,40456
8	Dispersion	0,2261	0,36335	0,41597	0,40486
9	Faith	0,2261	0,36335	0,41592	0,40478
10	Forbes-1	0,000038504	0,12587	0,20063	0,18463
11	Forbes2	0,2261	0,36335	0,41614	0,40486
12	Goodman&Kruskal	0,1483	0,1507	0,1788	0,1644
13	Gower	0,000012963	0,1249	0,2001	0,18405
14	Gower&Legendre	0,22112	0,35819	0,40942	0,39964
15	Hamann	0,2261	0,36335	0,41605	0,4049
16	Hellinger	0,24206	0,37917	0,43181	0,42093
17	Innerproduct	0,000415	0,13002	0,20332	0,18702
18	Jaccard	0,23093	0,36831	0,42083	0,40973
19	Jaccard3W	0,22433	0,36153	0,4128	0,40302
20	Johnson	0,2261	0,36335	0,4148	0,40456
21	Kulczynski-2	0,2261	0,36335	0,4148	0,40456
22	Lance&Williams	0,2261	0,36335	0,41636	0,40479
23	Mcconnaughey	0,2261	0,36335	0,41569	0,40469
24	Michael	0,20904	0,34589	0,39715	0,38713
25	Nei&Li	0,2261	0,36335	0,4148	0,40456
26	Cosine	0,2261	0,36335	0,4148	0,40456
27	Ochiai-2	0,0986	0,12166	0,19801	0,18262
28	Ort. Manhattan	0,24206	0,37917	0,4318	0,42089
29	Öklid	0,1238	0,12442	0,19984	0,18403
30	Pattern Difference	0,1986	0,33569	0,38699	0,37682
31	Pearson&Heron-1	0,2261	0,36335	0,41614	0,40486

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

Çizelge 5.5. (Devamı) Yüzde 90 Oranında ‘1’ İçeren Veri Kümesi İçin Çalışmada Kullanılan Ölçülere Bağlı Elde Edilen Kofenetik Korelasyon Katsayıları

	Ölçü	Tek	Tam	Ortalama	Ağırlıklı
32	Pearson&Heron-2	0,12055	0,23798	0,28458	0,27423
33	Roger&Tanimoto	0,23435	0,37179	0,42439	0,41336
34	Russell&Rao	0,2261	0,36335	0,41568	0,40483
35	Shape Difference	0,2261	0,36335	0,41633	0,40483
36	Simpson	0,2261	0,36335	0,4148	0,40456
37	Size Difference	0,1986	0,33569	0,38699	0,37682
38	Sokal&Michener	0,2261	0,36335	0,41587	0,4048
39	Sokal&Sneath-1	0,23887	0,37638	0,42905	0,41806
40	Sokal&Sneath-2	0,22112	0,35819	0,40942	0,39964
41	Sokal&Sneath-4	0,2261	0,36335	0,41605	0,40484
42	Sokal&Sneath-5	0,2317	0,36915	0,42172	0,41057
43	Sorgenfrei	0,22896	0,3663	0,41882	0,40768
44	Tarwid	0,22327	0,36043	0,41171	0,40191
45	Varyans	0,2261	0,36335	0,41633	0,40483
46	Yule-Q	0,12204	0,24557	0,29325	0,28244
47	Yule-Y	0,1094	0,21784	0,26274	0,25241

Her bir bağlantı yönteminde en iyi kümelenme sonuçlarının elde edildiği 5 ölçü Çizelge 5.6’de verilmiştir. Tüm bağlantı yöntemlerinde en yüksek kofenetik korelasyon katsayı değeri Chord ölçüsü kullanılarak elde edilmiştir.

Çizelge 5.6. Yüzde 90 Oranında ‘1’ İçeren Veri Kümesi İçin Bağlantı Yöntemlerine Göre En Yüksek Kofenetik Korelasyon Katsayı Değerine Sahip 5 Ölçü

Yöntem/Sıra	1	2	3	4	5
Tek	Chord	Hellinger	Öklid	S-Sneath-1	R.Tanimoto
Tam	Chord	Hellinger	Öklid	S-Sneath-1	R.Tanimoto
Ortalama	Chord	Hellinger	Öklid	S-Sneath-1	R.Tanimoto
A.Ortalama	Chord	Hellinger	Öklid	S-Sneath-1	R.Tanimoto

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

5.1.3. Yüzde 10 Oranında ‘1’ den Oluşan Veri Kümesinin Aşamalı Kümeleme Analizi Üzerindeki Etkisinin Araştırılması

Yüzde 10 oranında ‘1’ içeren 100 değişken ve 100 birimden oluşan veri kümesine ikili uzaklık ölçüleri kullanılarak aşamalı kümeleme analizi uygulanmıştır. Elde edilen kümelenme yapısının geçerliliğini ölçmek üzere kofenetik korelasyon katsayıları hesaplanmıştır. Farklı bağlantı yöntemlerinde elde edilen katsayı değerleri Çizelge 5.7’de verilmiştir.

Çizelge 5.7 incelendiğinde ölçülerin tamamında en yüksek kofenetik korelasyon katsayı değerini veren bağlantı yöntemi, ortalama bağlantı yöntemi olmuştur.

Çizelge 5.7. Yüzde 10 Oranında ‘1’ İçeren Veri Kümesi İçin Çalışmada Kullanılan Ölçülere Bağlı Elde Edilen Kofenetik Korelasyon Katsayıları

	Ölçü	Tek	Tam	Ortalama	Ağırlıklı
1	Anderberg	0,13735	0,34332	0,38485	0,37834
2	B.Urbani Buser-1	0,12554	0,26272	0,35408	0,345
3	B.Urbani Buser-2	0,12554	0,26272	0,35408	0,345
4	Braun-Blanquet	0,12554	0,26272	0,35408	0,345
5	Canberra	0,12554	0,26272	0,35397	0,34531
6	Chord	0,13444	0,27824	0,3676	0,35841
7	Dice	0,12554	0,26272	0,35408	0,345
8	Dispersion	0,12554	0,26272	0,35409	0,34513
9	Faith	0,12554	0,26272	0,35404	0,34524
10	Forbes1	0,10222	0,20921	0,30396	0,30082
11	Forbes2	0,12554	0,26272	0,35386	0,34508
12	Goodman&Kruskal	0,13735	0,34332	0,38463	0,37818
13	Gower	0,086348	0,17947	0,27837	0,27503
14	Gower&Legendre	0,11599	0,24298	0,33754	0,32926
15	Hamann	0,12554	0,26272	0,35386	0,34508
16	Hellinger	0,13444	0,27824	0,3676	0,35841
17	Innerproduct	0,12347	0,24654	0,33583	0,33279
18	Jaccard	0,13652	0,28366	0,37129	0,36217
19	Jaccard3W	0,11967	0,25077	0,34364	0,33541
20	Johnson	0,12554	0,26272	0,35408	0,345
21	Kulczynski2	0,12554	0,26272	0,35408	0,345
22	Lance&Williams	0,12554	0,26272	0,35404	0,34519
23	Mcconnaughey	0,12554	0,26272	0,35386	0,34508
24	Michael	0,11765	0,26888	0,34676	0,33856

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

Çizelge 5.7. Yüzde 10 Oranında ‘1’ İçeren Veri Kümesi İçin Çalışmada Kullanılan Ölçülere Bağlı Elde Edilen Kofenetik Korelasyon Katsayıları

25	Nei&Li	0,12554	0,26272	0,35408	0,345
26	Cosine	0,12554	0,26272	0,35408	0,345
27	Ochiai2	0,070939	0,15413	0,25284	0,24966
28	Ort. Manhattan	0,13444	0,27824	0,36761	0,35844
29	Öklid	0,085776	0,17837	0,27734	0,27419
30	Pattern Difference	0,10969	0,23525	0,32867	0,32035
31	Pearson&Heron1	0,12554	0,26272	0,35386	0,34508
32	Pearson&Heron2	0,12199	0,26568	0,35033	0,34219
33	Roger&Tanimoto	0,13652	0,28366	0,37129	0,36217
34	Russell&Rao	0,12554	0,26272	0,35406	0,34511
35	Shape Difference	0,12554	0,26272	0,35404	0,34519
36	Simpson	0,12554	0,26272	0,35408	0,345
37	Size Difference	0,10969	0,23525	0,32867	0,32035
38	Sokal&Michener	0,12554	0,26272	0,35408	0,345
39	Sokal&Sneath1	0,14677	0,30175	0,38614	0,37722
40	Sokal&Sneath2	0,11599	0,24298	0,33754	0,32926
41	Sokal&Sneath4	0,12554	0,26272	0,35408	0,345
42	Sokal&Sneath5	0,14027	0,29636	0,37822	0,36917
43	Sorgenfrei	0,14027	0,29636	0,37822	0,36917
44	Tarwid	0,11161	0,23338	0,33	0,32169
45	Varyans	0,12554	0,26272	0,35404	0,34519
46	YuleQ	0,11765	0,26888	0,34676	0,33856
47	YuleY	0,12554	0,26272	0,35386	0,34508

Her bir bağlantı yönteminde en iyi kümelendirme sonuçlarının elde edildiği 5 ölçü Çizelge 5.8’de verilmiştir. Tek ve Ortalama bağlantı yöntemlerinde Sokal&Sneath-1 ölçüsü en iyi yapılanmayı ortaya koyan ölçü olmuştur. Tam ve Ağırlıklı ortalama bağlantı yöntemlerinde ise en yüksek kofenetik korelasyon katsayı değeri, Anderberg ölçüsü ile elde edilmiştir.

Çizelge 5.8. Yüzde 10 Oranında ‘1’ İçeren Veri Kümesi İçin Bağlantı Yöntemlerine Göre En Yüksek Kofenetik Korelasyon Katsayı Değerine Sahip 5 Ölçü

Yöntem/Sıra	1	2	3	4	5
Tek	S-Sneath-1	S-Sneath-5	Sorgenfrei	Anderberg	G&Kruskal
Tam	Anderberg	G&Kruskal	S-Sneath-1	S-Sneath-5	Sorgenfrei
Ortalama	S-Sneath-1	Anderberg	G&Kruskal	S-Sneath-5	Sorgenfrei
A.Ortalama	Anderberg	G&Kruskal	S-Sneath-1	S-Sneath-5	Sorgenfrei

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

5.1.4. Yüzde 50 Oranında ‘1’ den Oluşan Veri Kümesinin Aşamalı Kümeleme Analizi Üzerindeki Etkisinin Araştırılması

Yüzde 50 oranında ‘1’ içeren 100 değişken ve 100 birimden oluşan veri kümesine ikili uzaklık ölçüleri kullanılarak aşamalı kümeleme analizi uygulanmıştır. Elde edilen kümelenme yapısının geçerliliğini ölçmek üzere kofenetik korelasyon katsayıları hesaplanmıştır. Farklı bağlantı yöntemlerinde elde edilen katsayı değerleri Çizelge 5.9’da verilmiştir.

Çizelge 5.9 incelendiğinde ölçülerin tamamında en yüksek kofenetik korelasyon katsayı değerini veren bağlantı yöntemi, ortalama bağlantı yöntemi olmuştur.

Çizelge 5.9. Yüzde 50 Oranında ‘1’ İçeren Veri Kümesi İçin Çalışmada Kullanılan Ölçülere Bağlı Elde Edilen Kofenetik Korelasyon Katsayıları

	Ölçü	Tek	Tam	Ortalama	Ağırlıklı
1	Anderberg	0,13735	0,34332	0,38485	0,37834
2	B.Urbani Buser1	0,12554	0,26272	0,35408	0,345
3	B.Urbani Buser2	0,12554	0,26272	0,35408	0,345
4	Braun-Blanquet	0,12554	0,26272	0,35408	0,345
5	Canberra	0,12554	0,26272	0,35397	0,34531
6	Chord	0,13444	0,27824	0,3676	0,35841
7	Dice	0,12554	0,26272	0,35408	0,345
8	Dispersion	0,12554	0,26272	0,35409	0,34513
9	Faith	0,12554	0,26272	0,35404	0,34524
10	Forbes1	0,10222	0,20921	0,30396	0,30082
11	Forbes2	0,12554	0,26272	0,35386	0,34508
12	Goodman&Kruskal	0,13735	0,34332	0,38463	0,37818
13	Gower	0,086348	0,17947	0,27837	0,27503
14	Gower&Legendre	0,11599	0,24298	0,33754	0,32926
15	Hamann	0,12554	0,26272	0,35386	0,34508
16	Hellinger	0,13444	0,27824	0,3676	0,35841
17	Innerproduct	0,12347	0,24654	0,33583	0,33279
18	Jaccard	0,13652	0,28366	0,37129	0,36217
19	Jaccard3W	0,11967	0,25077	0,34364	0,33541
20	Johnson	0,12554	0,26272	0,35408	0,345
21	Kulczynski2	0,12554	0,26272	0,35408	0,345
22	Lance&Williams	0,12554	0,26272	0,35404	0,34519
23	Mcconnaughey	0,12554	0,26272	0,35386	0,34508
24	Michael	0,11765	0,26888	0,34676	0,33856

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

Çizelge 5.9. (Devamı) Yüzde 50 Oranında ‘1’ İçeren Veri Kümesi İçin Çalışmada Kullanılan Ölçülere Bağlı Elde Edilen Kofenetik Korelasyon Katsayıları

25	Nei&Li	0,12554	0,26272	0,35408	0,345
26	Cosine	0,12554	0,26272	0,35408	0,345
27	Ochiai2	0,070939	0,15413	0,25284	0,24966
28	Ort. Manhattan	0,13444	0,27824	0,36761	0,35844
29	Öklid	0,085776	0,17837	0,27734	0,27419
30	Pattern Difference	0,10969	0,23525	0,32867	0,32035
31	Pearson&Heron1	0,12554	0,26272	0,35386	0,34508
32	Pearson&Heron2	0,12199	0,26568	0,35033	0,34219
33	Roger&Tanimoto	0,13652	0,28366	0,37129	0,36217
34	Russell&Rao	0,12554	0,26272	0,35406	0,34511
35	Shape Difference	0,12554	0,26272	0,35404	0,34519
36	Simpson	0,12554	0,26272	0,35408	0,345
37	Size Difference	0,10969	0,23525	0,32867	0,32035
38	Sokal&Michener	0,12554	0,26272	0,35408	0,345
39	Sokal&Sneath1	0,14677	0,30175	0,38614	0,37722
40	Sokal&Sneath2	0,11599	0,24298	0,33754	0,32926
41	Sokal&Sneath4	0,12554	0,26272	0,35408	0,345
42	Sokal&Sneath5	0,14027	0,29636	0,37822	0,36917
43	Sorgenfrei	0,14027	0,29636	0,37822	0,36917
44	Tarwid	0,11161	0,23338	0,33	0,32169
45	Varyans	0,12554	0,26272	0,35404	0,34519
46	YuleQ	0,11765	0,26888	0,34676	0,33856
47	YuleY	0,12554	0,26272	0,35386	0,34508

Her bir bağlantı yönteminde en iyi kümelendirme sonuçlarının elde edildiği 5 ölçü Çizelge 5.10’da verilmiştir. Tek ve Ortalama bağlantı yöntemlerinde en yüksek kofenetik korelasyon katsayı değerleri Sokal&Sneath-1 ölçüsünde elde edilmişken, Tam ve Ağırlıklı ortalama bağlantı yöntemlerinde ise Anderberg ölçüsü en iyi yapılanmayı ortaya koyan ölçü olmuştur.

Çizelge 5.10. Yüzde 50 Oranında ‘1’ İçeren Veri Kümesi İçin Bağlantı Yöntemlerine Göre En Yüksek Kofenetik Korelasyon Katsayı Değerine Sahip 5 Ölçü

Yöntem/Sıra	1	2	3	4	5
Tek	S-Sneath-1	S-Sneath-5	Sorgenfrei	Anderberg	G&Kruskal
Tam	Anderberg	G&Kruskal	S-Sneath-1	S-Sneath-5	Sorgenfrei
Ortalama	S-Sneath-1	Anderberg	G&Kruskal	S-Sneath-5	Sorgenfrei
A.Ortalama	Anderberg	G&Kruskal	S-Sneath-1	S-Sneath-5	Sorgenfrei

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

5.1.5. Gerçek Veri Kümesinin Aşamalı Kümeleme Analizi Üzerindeki Etkisinin Araştırılması

Bitki verisinin Jaccard ve Dice ölçüsü kullanılarak elde edilmiş uzaklık matrisine tek bağlantı, tam bağlantı, ortalama bağlantı ve ağırlıklı ortalama bağlantı kümeleme yöntemlerinin uygulanmasıyla elde edilen kofenetik korelasyon katsayısı değerleri Çizelge 5.11’de verilmiştir.

Çizelge 5.11. Jaccard ve Dice Ölçüleriyle Elde Edilen Kofenetik Korelasyon Katsayıları

	Jaccard	Dice
Tek	0,6192	0,593
Tam	0,6043	0,5514
Ortalama	0,8046	0,7809
A.Ortalama	0,7377	0,7289

İncelenen 47 ölçünün farklı uzaklık ölçüleri kullanılarak uygulanan aşamalı kümeleme analizi sonucu elde edilen kofenetik korelasyon katsayısı değerleri Çizelge 5.12’de verilmiştir.

Kofenetik korelasyon katsayısı değerine göre, en iyi kümelenmenin elde edildiği ölçü uygulanan bağlantı yöntemine göre değişmektedir. Genel olarak inceleyecek olursak, en çok tercih edilen ölçülerden Jaccard ölçüsü için en iyi kümelenme ortalama bağlantı yönteminde elde edilmiştir.

Çizelge 5.12. Gerçek Veri Kümesi İçin Çalışmada Kullanılan Ölçülere Bağlı Elde Edilen Kofenetik Korelasyon Katsayıları

	Ölçü	Tek	Tam	Ortalama	Ağırlıklı
1	Anderberg	0,5265	0,7478	0,8499	0,8370
2	B.Urbani Buser1	0,3270	0,4181	0,7127	0,6387
3	B.Urbani Buser2	0,2007	0,3764	0,5653	0,5326
4	Braun-Blanquet	0,5769	0,5688	0,7984	0,7270
5	Canberra	0,8499	0,6315	0,8591	0,4364
6	Chord	0,6113	0,5784	0,7976	0,7094
7	Dice	0,5930	0,5514	0,7809	0,7289
8	Dispersion	0,3389	0,6105	0,7917	0,7099
9	Faith	0,7124	0,5223	0,8131	0,4729
10	Forbes1	0,1221	0,0823	0,1517	0,1021

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

Çizelge 5.12. (Devamı) Gerçek Veri Kümesi İçin Çalışmada Kullanılan Ölçülere Bağlı Elde Edilen Kofenetik Korelasyon Katsayıları

11	Forbes2	0,4506	0,5258	0,7671	0,7301
12	Goodman&Kruskal	0,4876	0,7308	0,8454	0,8304
13	Gower	0,8567	0,8521	0,8587	0,8546
14	Gower&Legendre	0,8487	0,6265	0,8581	0,4213
15	Hamann	0,8499	0,6315	0,8651	0,4309
16	Hellinger	0,6113	0,5784	0,7976	0,7094
17	Innerproduct	0,5530	0,1579	0,6922	0,6812
18	Jaccard	0,6192	0,6043	0,8046	0,7377
19	Jaccard3W	0,5776	0,5222	0,7716	0,7250
20	Johnson	0,3729	0,5179	0,7733	0,7348
21	Kulczynski2	0,3729	0,5179	0,7733	0,7348
22	Lance&Williams	0,5930	0,5514	0,7809	0,7289
23	Mcconnaughey	0,3729	0,5179	0,7733	0,7348
24	Michael	0,3017	0,6578	0,7872	0,7019
25	Nei&Li	0,5930	0,5514	0,7809	0,7289
26	Cosine	0,5848	0,5388	0,7799	0,6895
27	Ochiai2	0,0926	0,0433	0,1135	0,0850
28	Ort. Manhattan	0,8499	0,6315	0,8633	0,4388
29	Öklid	0,8499	0,6315	0,8591	0,4317
30	Pattern Difference	0,6635	0,3068	0,6715	0,3099
31	Pearson&Heron1	0,5931	0,6306	0,7885	0,6991
32	Pearson&Heron2	0,1904	0,3609	0,6777	0,5870
33	Roger&Tanimoto	0,8518	0,6395	0,8649	0,4514
34	Russell&Rao	0,4110	0,5453	0,7896	0,7275
35	Shape Difference	0,8409	0,4348	0,8458	0,4633
36	Simpson	0,4200	0,5078	0,7553	0,7287
37	Size Difference	0,8038	0,5570	0,8270	0,6234
38	Sokal&Michener	0,8499	0,6315	0,8633	0,4388
39	Sokal&Sneath1	0,6517	0,6640	0,8297	0,7987
40	Sokal&Sneath2	0,8487	0,6265	0,8581	0,4213
41	Sokal&Sneath4	0,4153	0,5670	0,7797	0,7442
42	Sokal&Sneath5	0,5839	0,5414	0,7812	0,7122
43	Sorgenfrei	0,5872	0,6411	0,8261	0,8059
44	Tarwid	0,0862	0,3786	0,6629	0,5542
45	Varyans	0,8499	0,6315	0,8591	0,4317
46	YuleQ	0,1549	0,3529	0,6720	0,6249
47	YuleY	0,2985	0,3699	0,6537	0,5655

Çizelge 5.12'nin incelendiğinde ortalama bağlantı yöntemi hariç diğer tüm yöntemlerde Gower ölçüsü en iyi yapıyı ortaya koyarken, ortalama bağlantı yönteminde Hamann ölçüsü en yüksek kofenetik korelasyon katsayısı değerine

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

ulaşmıştır. En iyi sonuçların tüm ölçülerde ortalama bağlantı yöntemi uygulandığında elde edilmiş olduğu görülmektedir. Bu doğrultuda incelendiğinde en iyi 5 ölçünün Çizelge 5.13.'deki gibi olduğu görülür.

Çizelge 5.13. Gerçek Veri Kümesi İçin Bağlantı Yöntemlerine Göre En Yüksek Kofenetik Korelasyon Katsayı Değerine Sahip 5 Ölçü

Yöntem/Sıra	1	2	3	4	5
Tek	Gower	R-Tanimoto	Canberra ¹	G.Legendre ²	Shape Dif.
Tam	Gower	Anderberg	G.Kruskal	S-Sneath1	Michael
Ortalama	Hamann	R-Tanimoto	S.Michener ³	Canberra ⁴	Gower
A.Ortalama	Gower	Anderberg	G&Kruskal	Sorgenfrei	S-Sneath1

Gerçek veri kümesi üzerinde en yüksek kofenetik korelasyon katsayı değerini üreten ortalama bağlantı yönteminde en iyi sonucu veren Hamann ölçüsü kullanılarak yapılan aşamalı kümeleme analizi sonrasında elde edilen ağaç diyagramı Şekil 5.5'de verilmiştir. Ağaç diyagramı kullanılarak sınıflama yapılırken seçilecek kesme noktasına ilişkin kabul görmüş nesnel bir ölçüt yoktur. Yine de birim sayısına bağlı küme sayısının tahminine ilişkin önerilen yaklaşımlardan biri n birim sayısı olmak üzere optimal küme sayısı, $k = \sqrt{n/2}$ değerine yakın olacak şekilde bir kesme noktası belirlemektir (Alpar, 2010). Ayrıca araştırmacının alan bilgisi de bu noktanın belirlenmesinde oldukça önemli bir role sahiptir. Bu çerçevede $r=0.90$ alınarak sınıflandırma yapılmıştır. Elde edilen sonuçlar Çizelge 5.14'de verilmiştir.

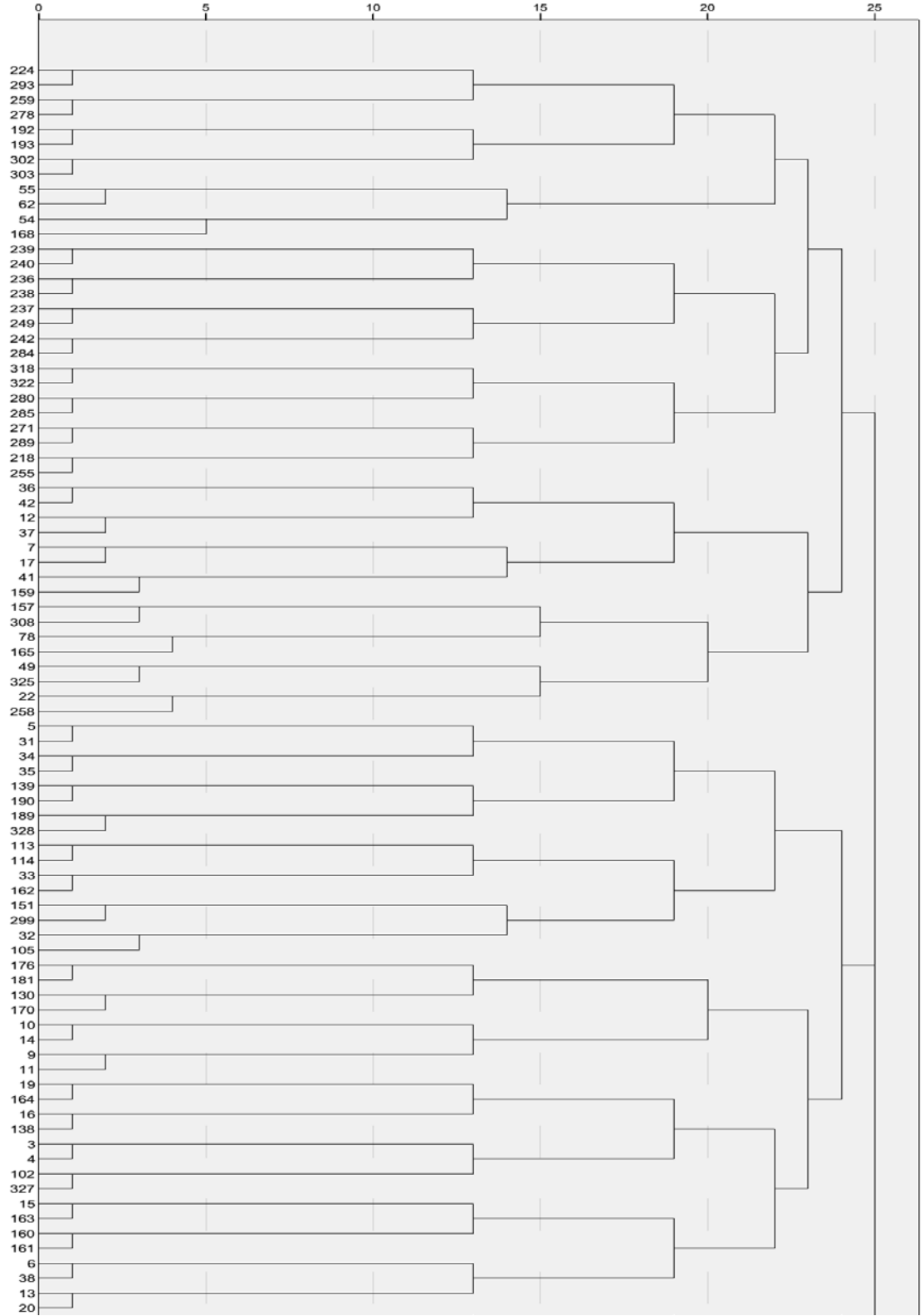
¹ Hamann, Ortalama Manhattan (City Block) , Öklid, Sokal&Michener ve Varyans ölçüleriyle aynı sıralamaya sahiptir.

² Sokal&Sneath-2 ölçüsü ile aynı sıralamaya sahiptir.

³ Ortalama Manhattan ölçüsü ile aynı sıralamaya sahiptir.

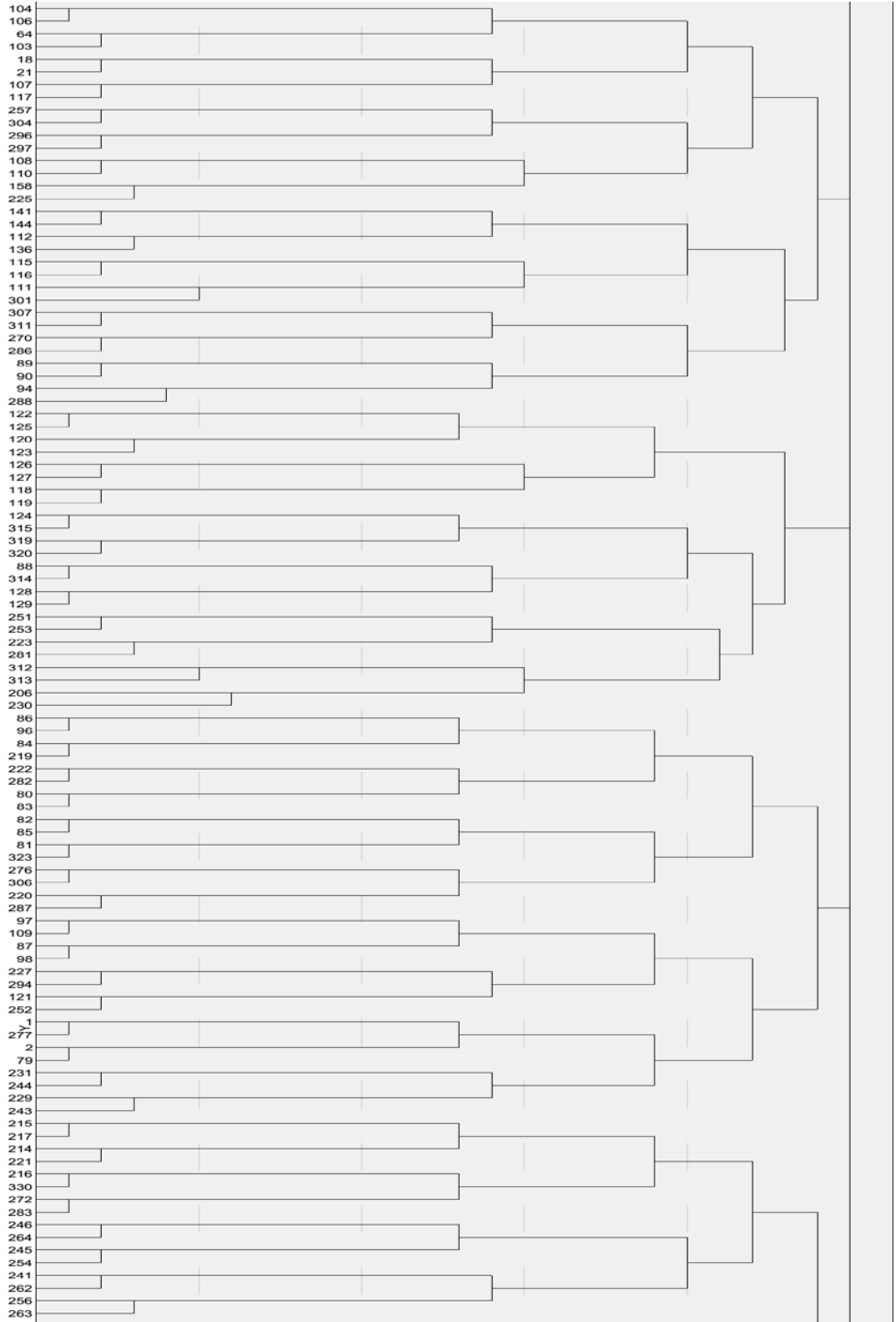
⁴ Öklid ve Varyans ölçüleriyle aynı sıralamaya sahiptir.

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM



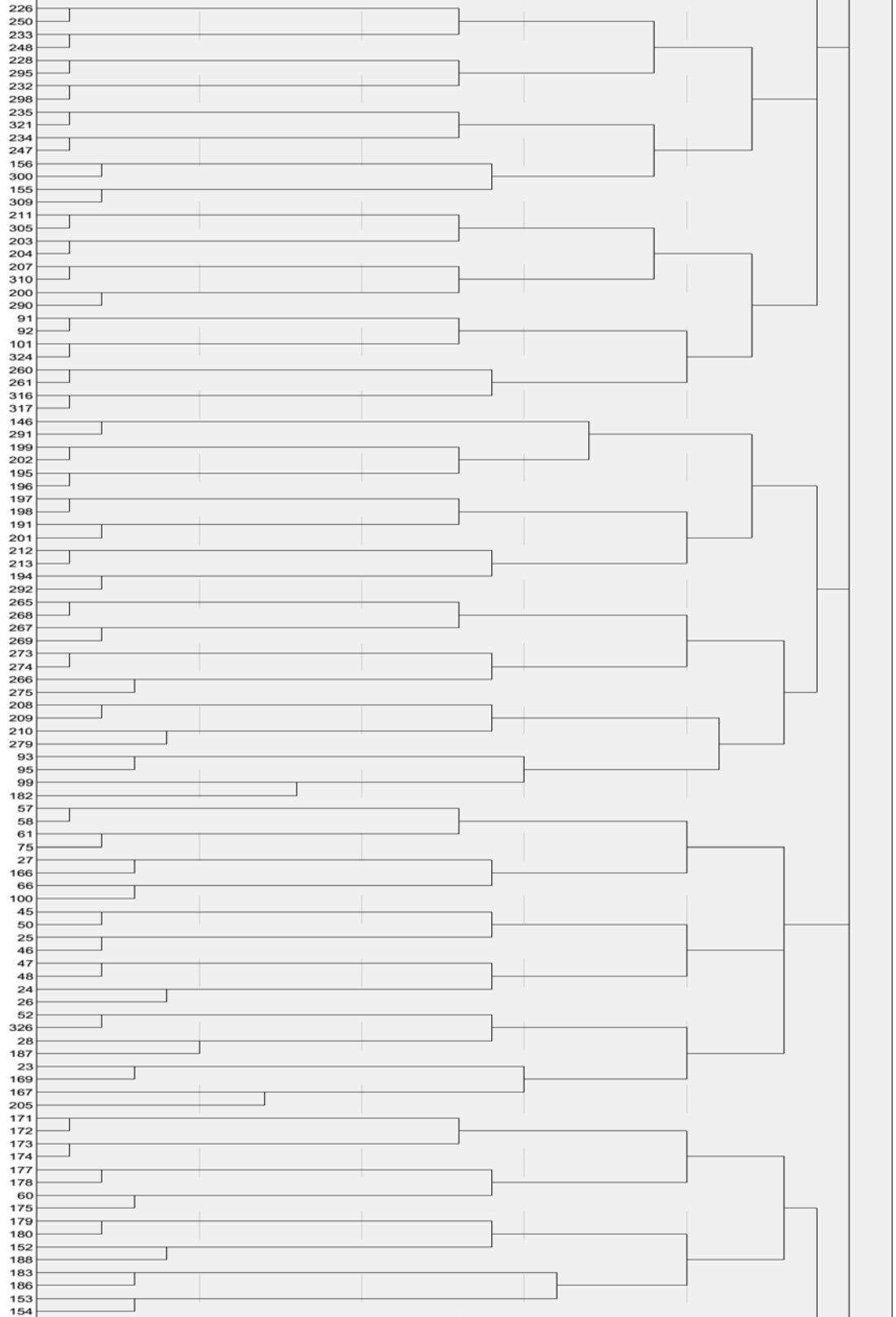
Şekil 5.5. Bitki verisi kullanılarak ortalama bağlantı yöntemi ile elde edilen ağaç diyagramı

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMEYE ETKİSİ Hasan YILDIRIM



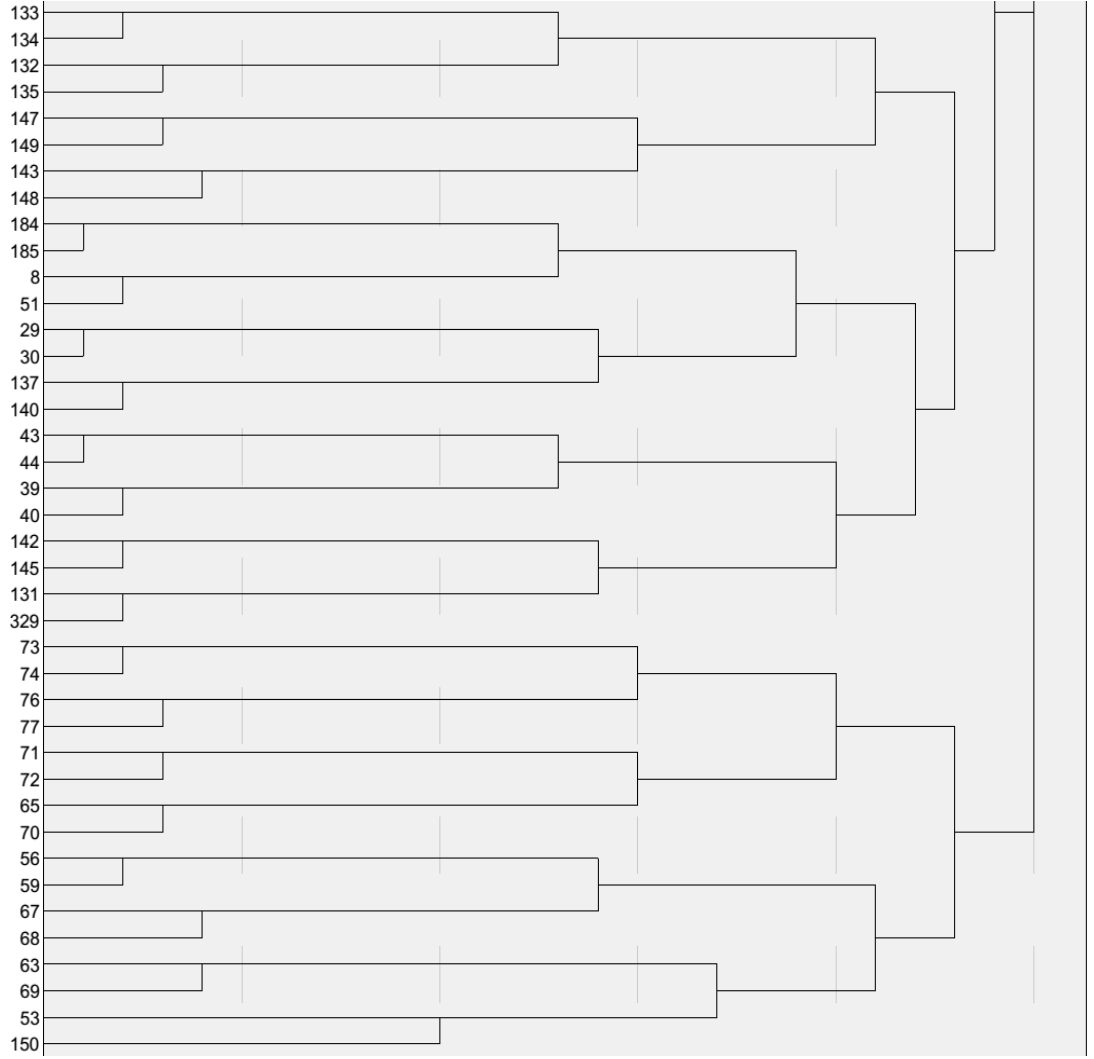
Şekil 5.5. (Devamı) Bitki verisi kullanılarak ortalama bağlantı yöntemi ile elde edilen ağaç diyagramı

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM



Şekil 5.5. (Devamı) Bitki verisi kullanılarak ortalama bağlantı yöntemi ile elde edilen ağaç diyagramı

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM



Şekil 5.5. (Devamı) Bitki verisi kullanılarak ortalama bağlantı yöntemi ile elde edilen ağaç diyagramı

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan YILDIRIM

Çizelge 5.14. Bitki Verisine Uygulanan Aşamalı Kümeleme Analizi Sonucu Elde Edilen Kümeler ve Birimleri

Küme No	Birimler
1	224, 293, 259, 278, 192, 193, 302, 303, 55, 62, 54, 168
2	239, 240, 236, 238, 237, 249, 242, 284, 318, 322, 280, 285, 271, 289, 218, 255
3	36, 42, 12, 37, 7, 17, 41, 159
4	157, 308, 78, 165, 49, 325, 22, 258
5	5, 31, 34, 35, 139, 190, 189, 328, 113, 114, 33, 162, 151, 299, 32, 105
6	176, 181, 130, 170, 10, 14, 9, 11
7	19, 164, 16, 138, 3, 4, 102, 327, 15, 163, 160, 161, 6, 38, 13, 20
8	104, 106, 64, 103, 18, 21, 107, 117, 257, 304, 296, 297, 108, 110, 158, 225
9	141, 144, 112, 136, 115, 116, 111, 301
10	307, 311, 270, 286, 89, 90, 94, 288
11	122, 125, 120, 123, 126, 127, 118, 119
12	124, 315, 319, 320, 88, 314, 128, 129, 251, 253, 223, 281, 312, 313, 206, 230
13	86, 96, 84, 219, 222, 282, 80, 83, 82, 85, 81, 323, 276, 306, 220, 287
14	97, 109, 87, 98, 227, 294, 121, 252, 1, 277, 2, 79, 231, 244, 229, 243
15	215, 217, 214, 221, 216, 330, 272, 283, 246, 264, 245, 254, 241, 262, 256, 263
16	226, 250, 233, 248, 228, 295, 232, 298, 235, 321, 234, 247, 156, 300, 155, 309
17	211, 305, 203, 204, 207, 310, 200, 290, 91, 92, 101, 324, 260, 261, 316, 317
18	146, 291, 199, 202, 195, 196, 197, 198, 191, 201, 212, 213, 194, 292
19	265, 268, 267, 269, 273, 274, 266, 275, 208, 209, 210, 279, 93, 95, 99, 182
20	57, 58, 61, 75, 27, 166, 66, 100, 45, 50, 25, 46, 47, 48, 24, 26, 52, 326, 28, 187, 23, 169, 167, 205
21	171, 172, 173, 174, 177, 178, 60, 175, 179, 180, 152, 188, 183, 186, 153, 154
22	133, 134, 132, 135, 147, 149, 143, 148, 184, 185, 8, 51, 29, 30, 137, 140, 43, 44, 39, 40, 142, 145, 131, 329
23	73, 74, 76, 77, 71, 72, 65, 70, 56, 59, 67, 68, 63, 69, 53, 150

5. UZAKLIK VE BENZERLİK ÖLÇÜLERİNİN KÜMELEMeye ETKİSİ Hasan
YILDIRIM

6. SONUÇLAR VE ÖNERİLER

Kümeleme doğal sınıflamaları hakkında açık bilgi bulunmayan durumlarda üzerinde çalışılan gözlemler topluluğunu, kendi içinde türdeş fakat diğerlerinden farklı olacak biçimde kümelere ayırma olanağı tanıyarak ana kitleye ilişkin tahminlerin yapılmasında yararlanılan çok değişkenli istatistiksel bir yöntem olarak tanımlanmaktadır (Sharma, 1996). Bu tanımda verilen kendi içinde türdeş ifadesi, mümkün olduğunca benzer birimlerden oluşan anlamına geldiğinden, birimler arası benzerliğin etkin bir şekilde ölçülmesi oldukça önem arz etmektedir. Bu çalışmada da alanyazında tanımlanmış hemen hemen tüm ikili uzaklık ve benzerlik ölçüleri derlenmiş, etkileri hem gerçek hem de farklı yapıda türetilen yapay veri kümeleri kullanılarak araştırılmıştır. Ayrıca bu ölçülerin genel sınıflamalarından bahsedilmiş ve bir takım koşullar çerçevesinde teorik yapıları incelenmiştir.

Bölüm 3’de kümeleme analizi ile ilgili temel tanım ve gösterimlere yer verilmiştir. Kümeleme analizi birçok alanda çeşitli amaçlarla yoğun olarak kullanılmaktadır. Genel olarak aşamalı ve aşamalı olmayan kümeleme yöntemleri olarak iki sınıfa ayrılmaktadır. Tez kapsamında sadece aşamalı kümeleme analiz yöntemi kullanıldığından bu bölümde bu yöntemi oluşturan iki alt gruptan – eklemeli ve bölünmeli- eklemeli kümeleme analizi ele alınmış ve tek bağlantı, tam bağlantı, ortalama bağlantı, ağırlıklı ortalama bağlantı, merkez bağlantı, medyan bağlantı, Ward bağlantı ve esnek beta yöntemleri incelenmiş ve bir uygulamayla yöntemin daha iyi anlaşılması amaçlanmıştır. Son olarak küme geçerliliği kavramından bahsedilmiştir. Küme geçerliliği için önerilen indeksler dışsal, içsel ve oransal indeksler olmak üzere üç başlık altında ele alınmıştır. Dışsal indekslerden Rand indeksi, Jaccard katsayısı, Fowlkes ve Mallows indeksi ve Γ istatistiği verilmiştir. İçsel indekslerden kofenetik korelasyon katsayısı ve Wilk’s lambda katsayısına değinilmiştir. Oransal indeksler daha çok aşamalı olmayan kümeleme analiz yöntemlerinde kullanıldığından yaygın olan indeksler verilmiş, detaylı değinilmemiştir.

Bölüm 4’de uzaklık ve benzerlik ölçüleriyle ilgili temel tanım ve kavramlara yer verilmiştir. Çeşitli veri yapılarına ilişkin tanımlanmış ölçüler ayrı ayrı

incelenmiş, ikili yapıdaki ölçülere ağırlık verilmiştir. Alanyazında tanımlı hemen hemen tüm ölçüler derlenmiş ve yapıları açıklanmaya çalışılmıştır. Ayrıca bir takım koşullar çerçevesinde bu ölçülerin teorik özellikleri incelenmiştir. İkili benzerlik ölçüleri 15 koşula (istenilen özelliğe) bağlı, uzaklık ölçüleri ise 12 koşula (istenilen özelliğe) bağlı olarak araştırılmıştır. Benzerlik ölçüleri incelendiğinde elde edilen sonuçlara göre, Braun-Blanquet, Cosine, Dice, Jaccard, Nei&Li ve Jaccard-3W ölçüleri tüm koşulları sağlamaktadır. Forbes-2, Kulczynski-2, McConnaughey, Simpson ve Sokal&Sneath-2 ölçüleri ise yukarıda verilen ölçülere birer alternatif olabilecek kadar iyi ölçüler olarak karşımıza çıkmaktadır. Çünkü bu ölçüler de çoğu koşulu sağlamaktadır. Diğer taraftan, Ample, Faith, Michael, Sokal&Sneath-3, Dispersion ve Peirce benzerlik ölçüleri ise sadece birkaç koşulu sağlamaktadır. Uzaklık ölçüleri incelendiğinde ise Faith, Hamann, Rogers&Tanimoto, Sokal&Michener ve Sokal&Sneath-3 uzaklık ölçülerinin tüm koşulları sağladığı görülmüştür. Dice, Gower&Legendre, Kulczynski-1, Sokal&Sneath-1, Sokal&Sneath-2 ve Jaccard ölçüleri de yukarıda verilen ölçülerden sonra çoğu koşulu sağlamaları nedeniyle tercih edilebilecek ölçüler olarak karşımıza çıkmaktadır. İncelenen koşullar doğrultusunda, en kötü sonuçlar Goodman&Kruskal, Pearson&Heron-2, Simpson, Ample, Anderberg, Phi ve Forbes-2 ölçülerinde elde edilmiştir.

Bölüm 5'te, Bölüm 4'te verilen ölçülerin kümeleme analizi üzerindeki etkileri araştırılmıştır. Bu amaçla biri gerçek, dördü farklı yapıda türetilmiş yapay veri kümeleri kullanılmıştır. Gerçek veri kümesi, 361 adet bitki türünün 330 bölge içerisinde gözlenip-gözlenmeme durumlarını içermektedir. Yapay olarak üretilen veri kümeleri ise tamamen rastgele ve sırasıyla %90, %50 ve %10 oranlarında '1' değerini içermektedir. Örneğin, %10, her bir birimin araştırılan 100 değişkenden 10'unda söz konusu niteliği taşıdığı anlamına gelir. Çalışmada tekrar sayısı 1000 olarak belirlenmiştir. Her bir veri kümesi 100 değişkene bağlı 100 birim içermektedir. Kümeleme analiz yöntemlerinden aşamalı kümeleme yöntemi uygulanmıştır. Yöntemin uygulanabilmesi için uzaklık matrisine ihtiyaç duyulduğundan, her bir ölçüye bağlı olarak uzaklık matrisleri hesaplanmıştır. Bu matrisler kullanılarak MATLAB programı aracılığıyla üretilen algoritmalar ile

sonuçlar elde edilmiştir. Bu işlem sırasında aşamalı kümeleme yönteminde her bir ölçü ve bağlantı yöntemi (tek, tam, ortalama, ağırlıklı ortalama) için süreç tekrarlanmıştır. Bulguların geçerliğini ölçmek amacıyla indeks olarak kofenetik korelasyon katsayısı kullanılmıştır. Elde edilen sonuçlar, en iyi kümelemenin kullanılan bağlantı yöntemi ve ölçü ikilisine bağlı değiştiğini göstermektedir. Biyoloji alanıyla ilgili gerçek veri kümesi kullanıldığında, en iyi sonuç ortalama bağlantı yönteminde Hamann benzerlik ölçüsüyle elde edilmiştir. Tamamen rastgele türetilmiş veri kümesinde ortalama bağlantı yönteminde Russell&Rao ölçüsü en iyi kümelenme yapısını ortaya koyan ölçü olmuştur. %90 oranında '1' içeren veri kümesinde bu sonuca Chord ölçüsü kullanıldığı ortalama bağlantı yönteminde ulaşılmıştır. %10 ve %50 oranlarında '1' içeren veri kümesi için en yüksek kofenetik korelasyon katsayısı Sokal&Sneath-1 ölçüsü ve ortalama bağlantı yöntemi kullanılarak elde edilmiştir. Elde edilen bu sonuçlar ile Bölüm 4'de bulunan sonuçlar bir arada incelendiğinde yüksek oranda uyum söz konusudur. Örneğin, Hamann ölçüsü teorik olarak en iyi ölçüler arasındayken buna paralel olarak kullanılan bitki verisinde de en yüksek kofenetik korelasyon katsayı değerini veren ölçü olmuştur. Benzer şekilde, Russell&Rao ve Sokal&Sneath-1 ölçüleri teorik olarak en iyi ölçüler arasında gösterilmişti. Sayısal sonuçlarda da bu ölçüler en iyi kümelenme yapısını ortaya koyan ölçüler olmuştur. Ayrıca en iyi beş ölçünün neredeyse tamamı teorik olarak da en iyi ölçüler olarak karşımıza çıkmaktadır. Bu ölçülere, Jaccard, Dice, Sorgenfrei, Roger&Tanimoto ve Braun&Blanquet ölçüleri örnek olarak verilebilir. Dolayısıyla hem teorik de hem de uygulamada elde edilen sonuçlar arasında yüksek derece uyum vardır.

İkili uzaklık ve benzerlik ölçülerinin seçimine ilişkin kriterlerin daha teorik bir alt yapıya dayanması daha güvenilir bir yaklaşım olacağından teorik bir takım koşullar ortaya koymak amacıyla araştırma yapılması faydalı olabilir. Ayrıca ikili veri kümeleri üzerinde yapılan kümeleme analizi sonrası elde edilen kümeleme yapısının geçerliğini ölçen alternatif indeksler geliştirmek de bir diğer araştırma konusu olabilir.

KAYNAKLAR

- ABDI, H., 2007. Distance, In Salkind, N.J. (eds.) Encyclopedia of Measurement and Statistics. 280–284. Sage Thousand Oaks. 1136p.
- ABREU, R., ZOETEWELJ, P. and ARJAN J.C., 2006, An Evaluation of Similarity Coefficients for Software Fault Localization, 12th Pacific Rim International Symposium. pp. 39 - 46
- ALDENFERDER, M.S., BLASHFIELD, R.K., 1984. The Literature on Cluster Analysis, Multivariate Behavioral Research.
- ALPAR, R., 2011. Uygulamalı Çok Değişkenli İstatistiksel Teknikler, Detay Yayıncılık. 853s.
- ANDERBERG, M. R., 1973. Cluster Analysis for Applications. Academic Press Inc., London. 359p.
- BAULIEU, F. B., 1989. A Classification of Presence/Absence Based Dissimilarity Coefficients. Journal of Classification. 6:233-246
- BARONI-URBANI, C. and BUSER, M.W., 1976, Similarity of Binary Data, Systematic Zoology, Vol. 25, No. 3. pp. 251-259.
- BHATTACHARYYA, A., 1943, On a measure of divergence between two statistical populations defined by probability distributions, Bull. Calcutta Math. Soc., vol. 35. pp. 99-109.
- BRAUN-BLANQUET, J., 1932, Plant Sociology: the Study of Plant Communities, McGraw-Hill, New York.
- BRAY, J.R. and CURTIS, J.T., 1957, An ordination of the upland forest of the southern Wisconsin, Ecological Monographies, 27. pp. 325-349.
- BÜYÜKÖZTÜRK, Ş. , ŞEKERCİOĞLU, G., ÇOKLUK, Ö., 2014. Sosyal Bilimler İçin Çok Değişkenli İstatistik: SPSS ve LISREL Uygulamaları, Pegem Akademi Yayıncılık. 424s.
- CHA, S. H., YOON S., TAPPERT, C.C., 2006. Enhancing Binary Feature Vector Similarity Measures. Journal of Pattern Recognition Research I.

- CHOI, S.-S., 2008. Correlation Analysis of Binary Similarity and Dissimilarity Measures. Phd Thesis. Seidenberg School of Computer Science and Information Systems, Pace University. p.162
- CLEMENTS, F. E., 1954, Use of Cluster Analysis with Anthropological Data, American Anthropologist. 56 (2): pp. 180-199.
- COX, D. R., 1957. Note on grouping. Journal of the American Statistical Association. 52 (280): pp. 543-547.
- DALLMEIER, V., LINDIG, C., ZELLER, A, (2005). Lightweight Defect Localization for Java, Proceedings of 19th European Conference on Object-Oriented Programming. ECOOP 2005. pp. 528-550.
- DEZA, M. M. and DEZA, E., 2006. Dictionary of Distances. Elsevier. 391p.
- DICE, L.R., 1945. Measures of the amount of ecological association between species, Ecology. (26): pp. 297-302.
- DİNÇER, K. S., 1992. Kümeleme Çözümlemesinde Uygun Kümeleme Ölçütlerinin Karşılaştırılması. Doktora Tezi, Anadolu Üniversitesi Sağlık Bilimleri Enstitüsü, Eskişehir. 128s.
- DRIVER, H.E. and KROEBER, A.L., 1932. Quantitative Expression of Cultural Relationships, University of California Press.
- DURAN , B. and ODELL , P. 1974. Cluster Analysis: A Survey, New York, NY: Springer - Verlag.
- ERİŞOĞLU, M., 2011. Uzaklık Ölçülerinin Kümeleme Analizine Olan Etkilerinin İncelenmesi ve Geliştirilmesi. Doktora Tezi - Çukurova Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, Adana. 199s.
- EVERITT, B.S., LANDAU, S., LEESE, M., STAHL, D., 2011. Cluster Analysis. Wiley, London. 348p.
- FAITH, D.P., MINCHIN and P.R., BELBIN, L., 1987. Compositional dissimilarity as a robust measure of ecological distance, Journal of Plant Ecology. Volume 69, Numbers 1-3.
- FINCH, H. and HUYNH H., 2000. Comparison of Similarity Measures in Cluster Analysis with Binary Data. Annual Meeting of the American Educational Research Association in New Orleans.

- FINELY, J.P., 1884. Tornado prediction. *The American Meteorological Journal*. 1, 85-8.
- FISHER, W. D., 1958. On grouping for maximum homogeneity. *Journal of the American Statistical Association*. 53 (284): pp. 789-798.
- FLOREK, K., LUKASZEWICZ, J., STEINHAUS, H., and ZUBRZYCKI, S., 1951. Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum*. 2: pp. 282-285.
- FORBES, S.A., 1907. On the local distribution of certain Illinois fishes. An essay in statistical ecology, *Bulletin of the Illinois State Laboratory of Natural History*.
- FORBES, S.A., 1925. Method of determining and measuring the associative relations of species, *Science*. 61, 524.
- FOWLKES, E. B. and MALLOWS, C. L., 1983. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*. 78: pp. 553-584.
- FRIEDMAN, H. P. and RUBIN, J., 1967. On some invariant criteria for grouping data. *Journal of American Statistical Association*. 62: pp. 1159-1178.
- GILBERT, G.K., 1884. Finley's tornado predictions, *The American Meteorological Journal*, 1, pp. 166-72.
- GOODMAN, L.A. and KRUSKAL, W.H., 1954. Measures of association for cross classifications, *Journal of the American Statistical Association*. 49, pp. 732-764.
- GOODMAN, L.A. and KRUSKAL, W.H., 1959. Measures of association for cross classifications II. Further discussion and references, *Journal of the American Statistical Association*. 54, 123-163 pp. 35-75.
- GOODMAN, L.A. and KRUSKAL, W.H., 1963. Measures of association for cross classifications III. Approximate sampling theory, *Journal of the American Statistical Association*. 58, pp. 310-364.
- GOODMAN, L.A. and KRUSKAL, W.H., 1972. Measures of association for cross classifications IV. Simplification of asymptotic variances, *Journal of the American Statistical Association*. 67, No. 338, pp. 315-321.

- GORDON, A. D., 1999. Classification. 2nd edn. Chapman & Hall, NewYork, NY. 256p.
- GOWER, J. C., 1966. Some distance properties of latent root and vactor methods used in multivariate analysis. *Biometrika*. 53 (3/4): pp. 325-338.
- GOWER, J. C., 1971. A general coefficient of similarity and some of its properties, *Biometrics*. Vol. 27, No. 4, pp. 857-871
- GOWER, J.C. and LEGENDRE, P., 1986. Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*. 3: 5-48.
- HAIR, J. F. and BLACK, W.C., 2009. Cluster Analysis. In Laurence G.Grimm & Paul R. Yarnold (Eds.). *Reading and Understanding More Multivariate Statistics*. Washington, D.C.: American Psychological Association.
- HAMANN, V., 1961. Merkmalbestand und Verwandtschaftsbeziehungen der Farinosae, Ein Beitrag zum System der Monokotyledonen, *Willdenowia*. Vol. 2. pp. 639-768.
- HANDS, S. and EVERITT, B., 1987. A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques. *Multivariate Behavioral Research*. 22, pp. 235-243.
- HARTIGAN, J. and WONG, M., 1979. Algorithm AS136: A k-means clustering algorithm. *Applied Statistics*. 28 (1): pp. 100–108.
- HUBALEK, Z., 1982. Coefficients of Association and Similarity, Based on Binary (Presence-Absence) Data: An Evaluation, *Biological Reviews*. Vol.57-4, pp. 669-689.
- JACCARD, P., 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura , *Bull Soc Vandoise Sci Nat*. 37: pp. 547-579.
- JACKSON, D.A., SOMERS, K.M. and HARVEY, H.H., 1989. Similarity Coefficients: Measures of Co-Occurrence and Association or Simply Measures of Occurrence?, *The American Nat1uralist*, Vol. 133, No. 3, pp. 436-453.
- JAIN, A.K. and DUBES, R.C. 1988. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ. 320p.

- JANSON, Svante, VEGELIUS, Jan, 1981. Measures of ecological association, *Oecologia*. Vol. 49, No. 3, pp 371-376.
- JOHNSON, S.C., 1967. Hierarchical clustering schemes, *Psychometrika*, Springer. vol. 32(3), pp. 241-254.
- JOHNSON, R. A. and WICHERN, D. W., 2014. Applied Multivariate Statistical Analysis. Sixth Edition, Pearson Prentice Hall, New Jersey.
- KENDALL, M. G., STUART, A. 1967. The Advanced Theory of Statistics, Second Edition, Charles Griffin, London.
- KRAUSE, E. F., 1986. Taxicab Geometry An Adventure in Non-Euclidean Geometry, Courier Dover Publications.
- KULCZYNSKI, S., 1927. Classedes Sciences Mathmatiques et Naturelles, Bulletin International de l'Acadamie Polonaise des Sciences et des Lettres Serie B (Sciences Naturelles) (Supplement II): pp. 57-203.
- LANCE, G.N. and WILLIAMS, W.T., 1967. A general theory of classificatory sorting strategies. I. Hierarchical systems. *Comput. J.* 9: 60–64.
- LEGENDRE, P., LEGENDRE, L., 1998. Numerical Ecology, Second English Edition Ed., Elsevier.
- MAGURRAN, A. E., 2003. Measuring Biological Diversity, Wiley-Blackwell, 264 p.
- MCCONNAUGHEY, B.H., 1964. The determination and analysis of plankton communities, *Marine Research in Indonesia*. Special number.
- MICHAEL, E.L., 1920. Marine ecology and the coefficient of association: a plea in behalf of quantitative biology, *Ecology*. 8, pp. 54-59.
- MOUNTFORD, M.D., 1962. An index of similarity and its application to classification problems, In: P.W.Murphy (ed.), *Progress in Soil Zoology*, 43-50, Butterworths.
- NEI, M. and LI, W.H., 1979. Mathematical Model for Studying Genetic Variation in Terms of Restriction Endonucleases, *Proceedings of the National Academy of Sciences of the United States of America*. Vol. 76, No. 10, pp. 5269-5273.

- OCHIAI, A., 1957. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions, *Bulletin of the Japanese Society for Sci Fish Science*. 22:526-530.
- OVERALL, J. E. and KLETT, J. C., 1972. *Applied Multivariate Analysis*. McGraw-Hill Book Company, New York. 362p.
- ÖZDAMAR, K., 2013. *Paket Programlarla İstatistiksel Veri Analizi*, Nisan Yayıncılık. 474s.
- PEARSON, K., 1905. On the general theory of skew correlation and non-linear regression, *Drapers' Company Research Memoirs, Biometric Series II*
- PEARSON, K., 1913. On the probable error of a coefficient of correlation as found from a fourfold table, *Biometrika*. 9, 22-27.
- PEARSON, K., HERON, D., 1913. On Theories of Association, *Biometrika*. 9 (12):159-315.
- PEIRCE, C.S., 1884. The numerical measure of the success of predictions, *Science*. 4, 453-454.
- PIELOU, E.C., 1969. *An Introduction to Mathematical Ecology*. Wiley-Interscience. New York
- RAND, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*. 66: 846–850.
- RENCHE, A. C., 2002. *Methods of Multivariate Analysis*. Second Edition, John Wiley&Sons Inc. Publication. 708p.
- ROGERS, J.S., TANIMOTO, T.T., 1960. A computer program for classifying plants, *Science*. (132): 1115-1118.
- ROHLF, F. and FISHER, D., 1968. Tests for hierarchical structure in random data sets. *Systematic Zoology*. 17: 407 – 412.
- ROUSSEUW, P., 1990. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 20: 53–65.
- RUSSELL, P.F., RAO, T.R., 1940. On habitat and association of species of anopheline larvae in south-eastern Madras, *Journal of Malaria India Institute*. (3): 153-178.

- SEO, Jinwook, 2005. Hierarchical Clustering Explorer 3.0 User's Guide, http://www.cs.umd.edu/hcil/hce/hce3-manual/hce3_manual.html
- SERVİ, T., 2009. Çok Değişkenli Karma Dağılım Modeline Dayalı Kümeleme Analizi. Doktora Tezi - Çukurova Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, Adana. 239s.
- SHARMA, S., 1996. Applied Multivariate Techniques. New York, JohnWiley & Sons. 493p.
- SANTINI, S., JAIN, R, 1999. Similarity Measures, IEEE Transactions on Pattern Analysis and Machine Intelligence. 21/9: 871-883
- SIMPSON, G.G., 1943. Mammals and the nature of continents, American Journal of Science. 241, 1-31.
- SNEATH, P. H. A. 1957. Some thoughts on bacterial classification. J. Gen. Microbiology. 17:184- 200.
- SNEATH, P.H.A., SOKAL, R.R., 1973. Numerical Taxonomy: The Principles and Practice of Numerical Classification. W.H. Freeman and Company, San Francisco.
- SOKAL, R.R., MICHENER, C.D., 1958. A statistical method for evaluating systematic relationships, Bulletin of the Society of University of Kansas. 38: 1409 1438.
- SOKAL, R. and ROHLF, F., 1962. The comparison of dendrograms by objective methods. Taxonomy. 11: 33–40.
- SOKAL, R.R., SNEATH P.H., 1963. Principles of numeric taxonomy. San Francisco, W.H. Freeman.
- SORENSEN, T., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Vidensk Selsk Biol Skr. 5:1-34.
- SORGENFREI, T., 1959. Molluscan assemblages from the marine middle Miocene of south Jutland and their environments. Denmark Geol. Unders. Ser. 2, no. 79, 403.
- TANIMOTO, T.T., 1957. IBM Internal Report 17th. IBM.

- TARWID, K., 1960. Szacowanie zbieżności ekologicznych gatunków drogi oceny prawdo- podobieństwa spotkania się ich w polowach, *Ekologia Polska* Ser. B. 6: 115-130.
- TATLIDİL, H. 1996. Uygulamalı Çok Değişkenli Analiz, Ankara. S.329-343.
- TEKNOMO, K., 2006. <http://people.revoledu.com/kardi/index.html> (Erişim tarihi: 20 Ağustos 2014)
- TIMM, N. H. 2002. *Applied Multivariate Analysis*. Springer. Second Edition. Pittsburgh. p718.
- TRYON, R. C., 1939. *Cluster Analysis*. Edwards Brothers. Ann Arbor, MI. 347p.
- TUBBS, J.D., 1989. A note on binary template matching, *Pattern Recognition*. 22(4):359-365.
- XU, R. and WUNSCH, D. C. II., 2009. *Clustering*, John Wiley & Sons, Inc. Hoboken, New Jersey. 358p.
- WARD, J., 1963. Hierarchical groupings to optimize an objective function. *Journal of the American statistical association*. 58: pp. 236 – 244.
- WILLETT, P., 2003. Similarity-Based Approaches to Virtual Screening, *Biochemical Society Transactions*. 31, pp. 603–606.
- WISHART, D., 2001. Gower's Similarity Coefficient.
- YULE, G.U., 1900. On the Association of Attributes in Statistics: With Illustrations from the Material of the Childhood Society, *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*. Vol. 194, pp. 257-319.
- YULE, G.U., 1912. On the methods of measuring association between two attributes, *Journal of the Royal Statistical Society*. Vol. 75, No. 6, pp. 579-652.
- ZHANG, B. and SRIHARI, S.N., 2003. Binary vector dissimilarities for handwriting identification. *Proceedings of SPIE, Document Recognition and Retrieval X*. p 15 166

ÖZGEÇMİŞ

1988 yılında Gaziantep’te doğdu. İlk ve orta öğrenimini Gaziantep’te, lise öğrenimini Adana’da tamamladı. 2007 yılında Eskişehir Anadolu Üniversitesi Fen Fakültesi İstatistik Bölümü’nü kazandı ve 2012 yılında mezun oldu. Aynı yıl, Karamanoğlu Mehmetbey Üniversitesi’ne araştırma görevlisi atanıp, ardından 2547 Sayılı Yükseköğretim Kanunu’nun 35. Maddesi gereği yüksek lisans eğitime Çukurova Üniversitesi’nde başladı. Halen burada araştırma görevlisi olarak görevine devam ediyor.