**ANKARA ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

**YÜKSEK LİSANS TEZİ**

# KÜMELEME ANALİZİ YÖNTEMLERİ İLE COVID-19 VERİLERİNİN İNCELENMESİ

**Ezgi Seren CANBAY**

**İSTATİSTİK ANABİLİM DALI**

**ANKARA**
**2022**

# ÖZET


Yüksek Lisans Tezi

KÜMELEME ANALİZİ YÖNTEMLERİ İLE COVID-19 VERİLERİNİN İNCELENMESİ

Ezgi Seren CANBAY

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Doç. Dr. Esin KÖKSAL BABACAN

Kümeleme analizi, çok değişkenli istatistiksel yöntemlerden biridir. Veri madenciliğinde sıkça kullanılan bu analiz, çok sayıda değişkene ait gözlem birimlerini gruplar içi homojen ve gruplar arası heterojen bir yapı oluşturacak şekilde kümelere ayırma işlemidir.

Bu çalışmada kümeleme analizinin çalışma alanlarına, analiz yöntemlerinin sınıflandırılmasına ve bu yöntemlerin çalışma prensiplerine yer verilmiştir. Dünyada büyük etki gösteren Covid-19 salgınına ilişkin 52 ülkeye ait veri internet üzerinden elde edilmiştir. Söz konusu veri seti ilkbahar, yaz, sonbahar ve kış olarak dört dönem biçiminde alınarak açık kaynak kodlu Python programı ile ülkelerin sınıflandırılmasına yönelik kümeleme analizi gerçekleştirilmiştir. Kümeleme analizi, $K$-Ortalamalar algoritması ve Ward yöntemi ile gerçekleştirilmiş olup, küme sayısının belirlenmesinde Elbow yönteminden faydalanılmıştır. Analiz sonucunda elde edilen kümeler karşılaştırılmış ve her iki yöntem için farklı kümelemelerin elde edildiği gözlemlenmiştir.

**Temmuz 2022, 86 sayfa**



**Anahtar Kelimeler:** Kümeleme analizi, hiyerarşik yöntemler, hiyerarşik olmayan yöntemler, $K$-Ortalamalar, Ward, Covid-19.

## ABSTRACT

Master Thesis

INVESTIGATION OF COVID-19 DATA USING WITH CLUSTERING ANALYSIS METHODS

Ezgi Seren CANBAY

Ankara University
Graduate School of Natural and Applied Science
Department of Statistics

Supervisor: Assoc. Prof. Dr. Esin KÖKSAL BABACAN

Cluster analysis is one of the multivariate statistical methods. This analysis, which is frequently used in data mining, is the process of dividing the observation units of many variables into clusters to form a homogeneous structure within groups and heterogeneous between groups.

In this study, the working areas of cluster analysis, the classification of analysis methods and the working principles of these methods are given. The data of 52 countries regarding the Covid-19 pandemic, which has a great impact in the world, was obtained from the internet. Clustering analysis for the classification of countries was carried out with the open source Python program by creating four periods as spring, summer, autumn and winter for the data set. Cluster analysis was performed with the $K$-Means algorithm and Ward's method, and the Elbow method was used to determine the number of clusters. The clusters obtained as a result of the analysis were compared and it was observed that different groups were formed.

**July 2022, 86 pages**



**Key Words:** Cluster analysis, hierarchical methods, non-hierarchical methods, $K$-Means, Ward, Covid-19.

## ÖNSÖZ ve TEŞEKKÜR

Tez çalışmam boyunca akademik bilgi ve birikimiyle her zaman yanımda olan, her konuda bana destek olarak karşılaştığım tüm sorunları sabırla ve güler yüzle çözmek için elinden geleni yapan, çalışmaktan büyük bir keyif aldığım çok değerli danışman hocam Doç. Dr. Esin Köksal Babacan' a sonsuz teşekkürlerimi sunarım.

Hayatımın tüm aşamalarında bana destek olan, çalışmalarımı benimseyerek tüm birikimiyle sonsuz destek veren, yorulmadan, bıkmadan her türlü sorunuma çözüm arayarak süreç boyunca beni hiç yalnız bırakmayan eşim Özgür Canbay'a hoşgörü ve anlayışı için teşekkür ederim.

Eğitim hayatımın başından itibaren desteğini esirgemeyen, akademik çalışmalara verdikleri önemle beni cesaretlendiren annem Türkan Ertürk, babam Talat Ertürk ve kardeşim Sude Ertürk'e teşekkür ederim.

Ezgi Seren CANBAY
Ankara, Temmuz 2022

# İÇİNDEKİLER

# ŞEKİLLER DİZİNİ

# ÇİZELGELER DİZİNİ

# 1. GİRİŞ

Kümeleme analizi, bir veri matrisinde yer alan ve doğal gruplamaları kesin olarak bilinmeyen birimleri, değişkenleri ya da birim ve değişkenleri, birbirlerine benzer olan alt kümelere (grup, sınıf) ayırmaya yardımcı olan yöntemler topluluğudur (Özdamar 1999). Kümeleme analizi, benzerlik ölçüsü olarak adlandırılan bazı ölçüler yardımıyla birimleri homojen gruplara bölmek amacıyla kullanılan çok değişkenli istatistiksel bir analiz yöntemidir. Çok değişkenli analiz yöntemleri ile verilere ait birden fazla özellik bir arada incelenebildiği için bu yöntemler veri madenciliğinde sıklıkla tercih edilmektedir. Bu şekilde bakıldığında, kümeleme analizi büyük boyutlu verileri gruplandırmak için kullanılan önemli veri madenciliği tekniklerindendir. Kümeleme analizi her bir nesne için çeşitli değişkenlerin gözlenen değerlerine dayanarak benzer birimleri, nesneleri birkaç kümede gruplandırma tekniğidir. Oluşturulan kümelerdeki nesneler, kendi içinde benzer olup diğer kümelere benzemezler. Bir kümeleme analizinden elde edilen sonuç, kümeleme olarak adlandırılır. Farklı kümeleme yöntemleri aynı veri seti üzerinde farklı kümeler üretilmesini sağlayabilir. Bu durum algoritmalara bağlı olduğundan daha önce bilinmeyen ve tahmin edilmesi güç grupların keşfedilmesine yol açmaktadır. Kümeleme analizi, web araması, coğrafi bilişim sistemleri, biyoloji ve güvenlik, iş zekası, görüntü deseni tanıma ve çok sayıda müşteriyi gruplar halinde organize etme gibi işlerde kullanılabilir (Han vd. 2012).

Kümeleme analizi uygulanırken ilk olarak değişkenler seçilerek veri matrisi belirlenir. Eğer varsa aykırı değerler incelenir. Veri seti incelenerek değişkenlere standartlaştırma yapılıp yapılmayacağına karar verilir. Nesnelerin benzerliklerini veya uzaklıklarını ölçmeye yarayan ölçülerden hangisinin kullanılacağına karar verilerek benzerlik/uzaklık matrisi elde edilir. Veriye uygun kümeleme yöntemi belirlenir. Optimum küme sayısının belirlenmesinde kullanılan yöntemler aracılığıyla uygun küme sayısı elde edilir. Son aşamada ise oluşturulan kümeler yorumlanır.

Veri madenciliği ile ilgili literatürde birçok farklı tanım yer almaktadır. Amerikan Pazarlama Birliği (AMA) veri madenciliğini şu şekilde tanımlamıştır: “Verilerin, yeni ve potansiyel olarak yararlı bilgi bulma amaçlı analiz süreci. Bu süreç bulunması zor

örüntülerin ortaya çıkarılması için matematiksel araçların kullanımını içerir." Gartner tarafından verilen bir başka tanım ise şu şekildedir, "Veri depolarında saklanan büyük miktarda veri üzerinden eleme ile anlamlı korelasyonlar, örüntüler ve trendler keşfetme süreci. Veri madenciliği, örüntü tanıma teknolojilerinin yanı sıra matematiksel ve istatistiksel teknikleri kullanmaktadır."

Literatürde yer alan farklı tanımlardan bahsedilebilir ancak yukarıda belirtilen iki tanım veri madenciliği sürecinin üç ana temelini ortaya koymaktadır:

- Miktarı büyük veri seti
- Potansiyel olarak yararlı bilgi
- Matematiksel ve istatistiksel teknikler (Akküçük 2011).

Veri madenciliğinde altı çizilmesi gereken önemli unsurlardan biri de elde edilecek bilginin "*önceden bilinmeyen"* olmasıdır. Önceden bilinmeyen olması ile bahsedilmek istenen, elde edilecek sonucun tahmin edilememesidir. Tahmin edilebilen sonuçlar veya çıkarılmış sonuçların ispatını yapmak için veri madenciliği araçlarını kullanmak mantıklı değildir. Veri madenciliğinin düşünülmemiş ve akla gelmesi zor olan sonuçları ortaya koyması, diğer yöntemlerden ayrılmasına neden olur. Akla gelmesi zor olan sonuçları ortaya koyması ile ilgili en bilinen örnek klasikleşmiş çocuk bezi-bira örneğidir (Akküçük 2011). Bir perakende mağazalar zinciri, müşteri kartı sisteminden veri toplayarak (müşterilerin hangi gün, hangi saat, neler satın alabildiği bilgisine ulaşarak) yaptığı veri madenciliği araştırmasının sonucunda farklı ürünlerin satışında, birden çok bağlantı olduğunu ortaya koymuştur. Bazıları tahmin edilebilir ürünler iken (cips alan müşterilerin aynı zamanda kola satın alması gibi) bazıları çocuk bezi ve bira satışları arasında, özellikle cuma günleri, güçlü bir ilişki çıkması gibi tahmin edilebilir değildir. Bu durum sorgulanması akla gelecek bir bağlantı değildir. Bu nedenle satışları daha fazla arttırmak için bira ile çocuk bezinin mağazada yan yana tezgahlara koyulması fikri düşünülerek işletme açısından karlı olacak kararlar alınabilir.

İstatistik ve veri madenciliği arasında nasıl bir ilişki olduğu merak konusu olmuştur. İstatistik, verilerin toplanması, işlenmesi, analiz edilmesi ve raporlanması

aşamalarından oluşur. Veri madenciliği ise istatistik biliminden doğmuş ve kendini makine algoritmalarıyla geliştirmiş bir bilimdir (Silahtaroğlu 2020). Veri madenciliği ve istatistik, verinin ön işlemesi, bilgiye dönüştürülmesi, tahmin yapılması, verinin analiz edilmesi ve kullanılan analiz çeşitleri konularında benzerlik gösterir. Farklılıklardan en önemlisi veri madenciliğinin milyonlarca hatta milyarlarca veri ve çok fazla değişkenle ilgilenmesidir. Çoğunlukla analizler bu verilerin tamamıyla yapılmaktadır. Bu nedenle veri madenciliğinin bilgisayar kullanılmadan uygulanması düşünülemez. İstatistiksel analizlerde ise genellikle yığın diye adlandırılan verilerin bütününü en iyi temsil eden örneklem belirlenir. İstatistiksel yöntemler daha eskiye dayandığından örneklemin büyüklüğüne göre bilgisayar kullanılmadan gerçekleştirilebilir. Buradan yola çıkarak veri madenciliği analizlerinin verinin tamamıyla gerçekleştirilmesi ile tümdengelim yaklaşımı, istatistiksel analizlerin ise örneklem analizi ile kitleye ilişkin çıkarımlar yapması sonucu tümevarım yaklaşımı ile hareket ettiği ortaya çıkmaktadır. Bir diğer fark ise istatistiksel çalışmaların belli bir amaçla yapılması, buna bağlı olarak hipotez kurulması ve çalışma başlamadan belirlenen sorulara cevap aranmasıdır. Veri madenciliğinde hipotez kurulmasına ihtiyaç yoktur, tam aksine çalışmanın sonucunda elde edilen bulgular yorumlanır. Bu durumda istatistiksel analizlerin çıkış noktasında deneye dayanmayan ön bilgiden söz edilebilir. Veri madenciliği ve istatistik birbirinden farklı iki kavram olsa da veri madenciliği teknikleri istatistiksel analizlerin temeline dayanmaktadır (Emre vd. 2017).

Çok değişkenli analiz yöntemleri ile verilere ait birden fazla özellik bir arada incelenebildiği için bu yöntemler veri madenciliğinde sıklıkla tercih edilmektedir. Bu şekilde bakıldığında, kümeleme analizi büyük boyutlu verileri gruplandırmak için kullanılan önemli veri madenciliği tekniklerindendir. Kümeleme analizi her bir nesne için çeşitli değişkenlerin gözlenen değerlerine dayanarak benzer birimleri, nesneleri birkaç kümede gruplandırma tekniğidir. Oluşturulan kümelerdeki nesneler, kendi içinde benzer olup diğer kümelere benzemezler. Bir kümeleme analizinden elde edilen sonuç, kümeleme olarak adlandırılır. Farklı kümeleme yöntemleri aynı veri seti üzerinde farklı kümeler üretilmesini sağlayabilir. Bu durum algoritmalara bağlı olduğundan daha önce bilinmeyen ve tahmin edilmesi güç grupların keşfedilmesine yol açmaktadır. Kümeleme analizi, web araması, coğrafi bilişim sistemleri, biyoloji ve güvenlik, iş zekası, görüntü

deseni tanıma ve çok sayıda müşteriyi gruplar halinde organize etme gibi işlerde kullanılabilir (Han vd. 2012).

Kümeleme analizi uygulanırken ilk olarak değişkenler seçilerek veri matrisi belirlenir. Eğer varsa aykırı değerler incelenir. Veri seti incelenerek değişkenlere standartlaştırma yapılıp yapılmayacağına karar verilir. Nesnelerin benzerliklerini veya uzaklıklarını ölçmeye yarayan ölçülerden hangisinin kullanılacağına karar verilerek benzerlik/uzaklık matrisi elde edilir. Veriye uygun kümeleme yöntemi belirlenir. Optimum küme sayısının belirlenmesinde kullanılan yöntemler aracılığıyla uygun küme sayısı elde edilir. Son aşamada ise oluşturulan kümeler yorumlanır.

Bu çalışmanın amacı, çok değişkenli analiz yöntemlerinden olan ve veri madenciliğinde sıklıkla kullanılan kümeleme analizi yöntemlerini araştırmak ve kullanım alanlarına dair bilgi vermektir. Çalışmanın ikinci bölümünde, kümeleme analizi konusuna giriş yapılarak detaylı literatür taramasına yer verilmiştir. Üçüncü bölümde, kümeleme analizinin anlaşılması için gerekli tanım ve kavramlar açıklanmıştır. Ayrıca kümeleme analizi yöntemlerine açıklamalarıyla yer verilmiştir. Dördüncü bölümde, kümeleme analizinin en yaygın kullanılan yöntemlerinden $K$-Ortalamalar ve Ward algoritmaları, 52 ülkeye ait Covid-19 verilerine uygulanarak ülkelerin benzerliklerine göre kümelere ayrılması ve sonuçların değerlendirilmesi yapılmıştır. Son bölüm ise, sonuç ve tartışmalara ayrılmıştır.

## 2. LİTERATÜR TARAMASI

Bu bölümde literatürde kümeleme analizi ile ilgili yapılan çalışmaların bir kısmı incelenmiştir.

Sığırlı ve arkadaşları, 2006 yılında yaptıkları çalışma ile Avrupa Birliği üyesi ülkeleri sağlık düzeyi ölçütlerine göre gruplandırmışlardır. Yaptıkları çalışmada çok boyutlu ölçekleme metodu aracılığıyla ülkelerin iki boyutlu uzayda farklı üç grup meydana getirdikleri tespit edilmiştir. Çalışmanın sonucunda Türkiye'nin Slovakya, Macaristan ve Çek Cumhuriyeti dışında kalan diğer ülkelerden sağlık alanındaki temel göstergeler açısından, ikinci boyutta ise milli gelirin içerisinde sağlığa ayrılan pay ve sağlık alanındaki harcamalar açısından farklılık gösterdiğini ifade etmişlerdir.

Erkekoğlu, 2007 yılında yaptığı çalışmada ülkelerin gelişmişlik düzeylerini kümeleme analizi yöntemiyle incelemiştir. Çalışmanın sonucunda, ülkelerin beş gruba ayrıldığı görülmüştür. Türkiye, Litvanya, Letonya, Polonya, Bulgaristan ve Romanya aynı kümede yer almıştır. Bu durumda aynı kümede yer alan bu altı ülkenin aynı gelişmişlik düzeyinde oldukları söylenebilir.

Öz ve arkadaşları, 2009 yılında yaptıkları çalışmalarında Türkiye ve Avrupa Birliği üyesi ülkeleri beşeri sermaye bileşenleri olan eğitim, sağlık, işgücü piyasalarına ilişkin göstergeler bakımından kıyaslamak için kümeleme analizi gerçekleştirmiştir. Çalışmanın sonucuna göre, söz konusu göstergeler için Türkiye'nin Avrupa Birliği üyesi ülkelerle benzerlik göstermediği ortaya çıkmıştır.

Ersöz, 2009 yılında yaptığı çalışmada OECD ülkelerini bazı sağlık göstergeleri bakımından hiyerarşik kümeleme yöntemleri ile hiyerarşik olmayan kümeleme yöntemlerinden $K$-Ortalamalar ve $K$-Medoids algoritmaları kullanarak analizlerin karşılaştırılmasını yapmıştır. Çalışmaya göre, Türkiye'nin, Kore Cumhuriyeti, Meksika, Polonya ve Slovakya ile birçok sağlık göstergesi bakımından benzer oldukları söylenebilir.

McCuddy, 2010 yılında yaptığı çalışmasında polinom regresyon modeli ile ekonomik özgürlükler ve kamuya ait yolsuzluklar arasındaki ilişkiyi incelemiştir. Analiz sonucunda yolsuzluk algılama endeksi ile ekonomik özgürlük endeksi arasında karesel bir ilişkinin olduğu gözlemlenmiştir.

Kılıç ve arkadaşları, 2011 yılında yaptıkları çalışmalarında 30 ülkeyi turizm istatistikleri bakımından bulanık kümeleme analizi yöntemi kullanarak sınıflandırmışlardır. Analiz sonucunda küme sayısı üç olarak belirlenmiş olup, kümelerin ilkinde Paraguay, Slovakya, Norveç, Şili, Macaristan; ikincisinde Almanya, Avusturya, ABD, Çin, Fransa, Polonya, İtalya, Hollanda, Portekiz ve son kümede ise Arabistan, Avustralya, Belarus, Filipinler, Panama, Tunus, Jamaika, Mısır, Ürdün, Hırvatistan, Fas, İsrail, Finlandiya, Türkiye, İspanya ve Slovenya ülkeleri yer almıştır.

Hemant ve Pushpavathi, 2012 yılında yaptıkları çalışmalarında hastaneden elde ettikleri 768 adet diyabet verisi ile kümeleme analizi gerçekleştirmişlerdir. Bu analizde sınıflandırma yapmak için $K$-Ortalamalar algoritması kullandıktan sonra SMO (Sequential Minimal Optimization), Naive Bayes, Bagging, AdaBoost, J48, Rotation Forest and Random Forest gibi çeşitli sınıflandırma algoritmaları kullanmışlardır. Çalışmanın sonunda en iyi sonucun Bagging algoritması ile elde edildiği görülmüştür.

Balasubramanian ve Umarani'nin 2012 yılında yaptıkları çalışma, Hindistan'ın Tamil Nadu Eyaletinin Krishnagiri Bölgesinde florür oranı yüksek içme suyu kullanan insanların oluşturduğu bir örneklem ile gerçekleştirilmiştir. Çalışmada sudaki florür seviyesinin, farklı değişkenlerin etkisinin de incelenmesiyle, ne tür diş hastalıklarına sebep olduğu tespit edilmek istenmiştir. Analizde $K$-Ortalamalar algoritması kullanılmıştır.

İşler ve Narin, 2012 yılında yaptıkları çalışmada konjestif kalp yetmezliği rahatsızlığı bulunan hastalar ve kontrol grubunda yer alan toplam 83 kişiden elde edilen kalp hızı değişkenliği verileri ile kümeleme analizi gerçekleştirmiştir. Weka yazılımıyla elde edilen $K$-Ortalamalar algoritmasının sonuçlarına göre konjestif kalp yetmezliği hastalığının teşhisinde bu yöntemin başarılı olduğu değerlendirilmiştir. Bunun yanı sıra

$K$-Ortalamalar algoritması uygulanırken başlangıç merkez değerlerinin başarıyı etkilediği bilindiğinden, farklı değerler ile algoritma tekrarlanmıştır. Çalışmanın sonucunda dört kümenin yer aldığı durum için en yüksek %98.79 başarı elde edilmiştir.

Bulum ve arkadaşları, 2013 yılında yaptıkları çalışmalarında uluslararası ticaret hacminin ülkelerin coğrafi konumuyla ilişkisini kümeleme analizi yöntemlerinden Ward ve $K$-Ortalamalar kullanarak analiz etmişlerdir. 2004-2012 yılları arasında 200 ülkeye ait ithalat ve ihracat verilerinin kullanıldığı çalışmadan elde edilen bulgulara göre uluslararası ticarette ABD diğer ülkelere göre öne çıkmıştır. Daha sonrasında ise Çin ve Almanya'nın geldiği görülmüştür.

Çelik, 2013 yılında yaptığı çalışmasında tüm illere ait çeşitli sağlık değişkenlerini içeren TÜİK'in 2010 yılına ait sağlık verilerini, illerin bu verilere göre gelişmişliğini veya farklılığını incelemek amacıyla kullanmıştır. SPSS ile yapılan analizde veriler 7, 10 ve 15 kümeye ayrılarak elde edilen sonuçlar değerlendirilmiştir. Hiyerarşik kümeleme tekniklerinden tek bağlantı tekniği ile Ward yöntemi ve hiyerarşik olmayan kümeleme yöntemlerinden ise $K$-Ortalamalar algoritması kullanılmıştır. Analize göre sağlık verileri bakımından en sorunlu ve sorunsuz iller de tespit edilmiştir.

Kangallı ve arkadaşları, 2014 yılında yaptıkları çalışmalarında, 2011 yılına ait ekonomik özgürlük endeksinde yer alan verilerden yararlanarak OECD üyesi ülkeleri kümeleme analizi ile sınıflandırmışlardır. Analizde $K$-Ortalamalar algoritması ve Ward yöntemi kullanılmıştır. Çalışmanın sonucunda bu ülkeler, ekonomik özgürlük ve gelişmişlik düzeyi bakımından üç kümeye ayrılarak sonuçlar incelenmiştir. İkinci kümede yer alan (Türkiye, Meksika, Slovak Cumhuriyeti, Güney Kore, İsrail, İspanya, Çek Cumhuriyeti, Macaristan, Polonya, Portekiz, Slovenya, Yunanistan ve İtalya) ülkelerin diğer kümelere kıyasla mali özgürlük ve kamu harcaması dışındaki ekonomik özgürlük endeksini oluşturan diğer tüm değişkenler bakımından en az ortalamaya sahip olduğu değerlendirilmiştir.

Kolay ve Erdoğmuş, 2016 yılında yaptıkları çalışmalarında UCI (University of California-Irvine) makine öğrenmesi laboratuvarından alınan 286 hastaya ait 9 özellik

kullanarak göğüs kanserinin tekrar etme durumunu sınıflandırmışlardır. Buna göre, Matlab programı ile uzaklık ölçütlerine göre $K$-Ortalamalar algoritması başarısı test edilmiştir. Daha sonra Weka yazılımı ile verilere hiçbir ön işleme yapılmadan çeşitli makine öğrenmesi yöntemleri ile sınıflandırma yapılmıştır. Çalışmanın sonucunda, kümeleme analizine ilişkin uzaklık ölçülerinin başarıyı yaklaşık %12 değiştirdiği, bununla birlikte sınıflandırma başarısının ise yöntemlere göre %45 ile %79 arasında değiştiği görülmüştür.

Şahin, 2017 yılında yaptığı çalışmasında 23 Doğu Avrupa ülkesini 10 ekonomik özgürlük endeksi kullanarak Ward yöntemi ile kümelere ayırmıştır. Kümeleme analizi ile oluşturulan dört küme içerisinde Türkiye'nin yeri incelendiğinde Bosna-Hersek, Çek Cumhuriyeti, Hırvatistan, Karadağ, Macaristan, Moldava, Polonya, Sırbistan, Slovakya ve Kıbrıs ile kümelerinin aynı olduğu görülmüştür.

Mut ve Akyürek, 2017 yılında yaptıkları çalışmalarında sağlık göstergeleri bakımından OECD ülkelerini sınıflandırarak Türkiye'nin hangi ülkelerle benzerlik gösterdiğini incelemişlerdir. Küme sayısının belirlenmesi için Ward yönteminden yararlanılmıştır. Küme sayısının belirlenmesinden sonra ise $K$-Ortalamalar algoritması uygulanmıştır. Kümeleme analizinde kullanılan söz konusu tüm sağlık göstergelerinin önemli düzeyde etkin olduğu değerlendirilmiştir. Her iki yöntemden elde edilen sonuçlara göre Türkiye'nin aynı kümede bulunduğu ülkelerin Meksika ve Şili olduğu görülmüştür.

Silitonga, 2017 yılında yaptığı çalışmasında Hindistan'daki bir hastaneden elde ettiği verileri kullanarak bu hastaneye gelen hastalara ilişkin çeşitli değişkenlerle $K$-Ortalamalar yöntemi ile kümeleme analizi yapmıştır. Araştırmanın amacı, hastaneye gelen hastaların hastalığa eğilim örüntüsünü bulmaktır. Analiz sonucuna göre, gelecekte hastanenin elde edilen eğilim modeline dayalı olarak hizmet önceliğini tahmin etmesi beklenmektedir.

Çınaroğlu ve Avcı, 2018 yılında yaptıkları çalışmalarında hizmet bölgelerini belirlemede Sağlık Bakanlığınca belirlenen coğrafi yakınlık kriterinin haricinde araştırmanın perspektifini genişleterek sağlık alanında bölgesel planlama yapılmasına

ilişkin çalışmaların etkinliğini arttırmak için neler yapılması gerektiği konusunda tavsiyede bulunmuşlardır. Gerçekleştirilen kümeleme analizi sonucunda 10 küme elde edilmiştir. Çalışma sonucunda sağlık hizmet bölgelerini belirlemede etkinliğin arttırılabilmesi için öncelikli ele alınması gereken faktörlerin ekonomi, sağlık ve nüfus yoğunluğu olduğu değerlendirilmiştir.

Veena ve arkadaşlarının 2018 yılında yaptıkları çalışmaya ilişkin veri seti Hindistan'ın kuzeydoğu kesimine ait karaciğer hastalığı olan 416 kişi ve sağlıklı 167 kişiden oluşmaktadır. Kümeleme analizi yapılırken C4.5, Boosting in C5.0, Random Forest, KNN (K Nearest Neighbours), $K$-Ortalamalar ve Naive Bayes algoritmaları kullanılmış olup bulunan sonuçlar karşılaştırılmak istenmiştir. Algoritmaların iki veya daha fazla kombinasyonunun en iyi sonuçların belirlenmesinde son derece yardımcı olduğu öne sürülmüştür. Bu çalışma, karaciğerde meydana gelen hastalıkların erken dönemde tespit edilmesini amaçlamıştır.

Akdamar, 2019 yılında yaptığı çalışmasında iş gücü piyasası göstergelerine göre OECD ülkelerini, kümeleme analizi ve çok boyutlu ölçekleme analizi yöntemleriyle gruplara ayırmıştır. Hiyerarşik kümeleme analizinden elde edilen sonuçlara göre ülkeler 4 grupta incelenmiştir. Çok boyutlu ölçekleme analizi sonuçları değerlendirildiğinde OECD ülkeleri, iyi ile mükemmel arası bir uyumla iki boyutlu uzayda görselleştirilmiştir. Yapılan analizlerde, 2009 mortgage krizi nedeniyle iş gücü piyasalarına dair bir takım sert tedbirler alan Yunanistan ve İspanya'nın geri kalan ülkelerden ayrıştığı gözlemlenmiştir.

Tekin, 2020 yılında yaptığı çalışmasında hiyerarşik kümeleme analizi ile ülkeleri Covid-19 salgının etkilerine göre gruplara ayırmıştır. Çalışmada değişken olarak vaka sayısı, test sayısı, ölüm sayısı, bazı finansal göstergeler ve ülkelerin salgından önceki sağlık göstergeleri düzeyi kullanılmıştır. Hiyerarşik kümeleme yöntemiyle dört ayrı veri setine, dört ayrı kümeleme yöntemi uygulanarak küme sayıları üçlü, dörtlü, beşli ve yedili olacak şekilde analiz edilmiş olup sonuçlara ilişkin karşılaştırmalar yapılmıştır.

Demircioğlu ve Eşiyok, 2020 yılında yaptıkları çalışmalarında Covid-19 salgınının dünyaya etkilerini ülkelere göre incelemiştir. Yapılan çalışmada, kümeleme analizi yöntemlerinden $K$-Ortalamalar algoritması kullanılmış olup OECD ve Avrupa Birliği'ne üye olan 36 ülke, sağlık alanındaki verileri incelenerek sınıflandırılmıştır. Çalışmanın sonunda ülkeler ikili, üçlü ve dörtlü kümelere ayrılarak karşılaştırmalar yapılmıştır.

Değirmenci ve Ayan, 2020 yılında yaptıkları çalışmalarında Ekonomik İşbirliği ve Kalkınma Örgütü'ne üye ülkeleri, sağlık alanındaki göstergelerini baz alarak bulanık kümeleme analizi ve TOPSIS yöntemi ile sınıflandırmıştır. Yapılan bulanık kümeleme analizinin sonucuna göre Türkiye, Kore, Meksika ve Polonya aynı kümede yer almıştır. Bununla birlikte TOPSİS analizinde, Türkiye'nin Meksika'nın ardından son sırada yer aldığı görülmüş olup ilk üç sırayı Amerika Birleşik Devletleri, Japonya ve Kanada'nın oluşturduğu anlaşılmıştır. Türkiye'nin örgütteki diğer ülkeler ile kıyaslanması neticesinde ise hastanede kişi başına düşen yatak sayısı haricindeki diğer sağlık imkanlarının yeterli olmadığı değerlendirilmiştir.

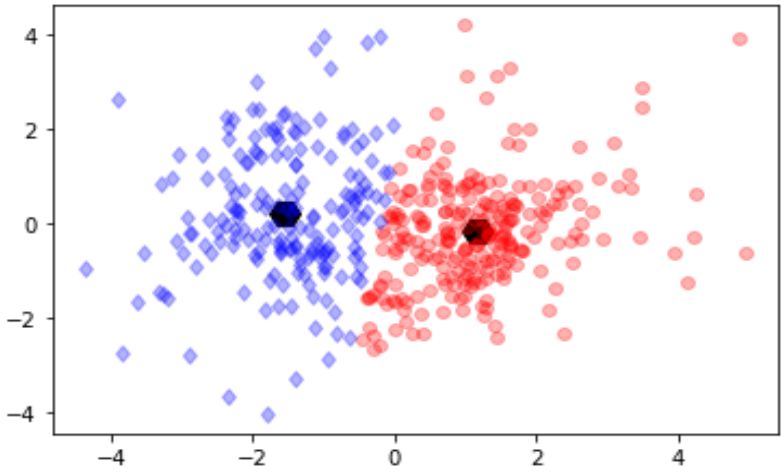
Demirkale ve Özarı, 2020 yılında yaptıkları çalışmalarında hiyerarşik olmayan $K$-Ortalamalar yöntemi ile 2007-2019 dönem aralığını baz alarak kırılgan beşli ülkelerini 5 temel makro-ekonomik ve finansal göstergeler açısından gruplandırmışlardır. Yapılan analizlerin sonucunda, Türkiye ve Brezilya'nın genellikle aynı kümede, diğer üç ülkenin (Hindistan, Endonezya ve Güney Afrika) ise genellikle farklı bir kümede birlikte yer aldıkları görülmüştür.

Çetintürk ve Gençtürk, 2020 yılında yaptıkları çalışmalarında 2003 ve 2017 yılları arasında OECD ülkelerinin sağlık alanındaki harcamalarını, miktar ve oranını dikkate alarak değerlendirmiştir. Çalışmaya konu olan ülkeler harcama türleri dikkate alınarak kümelere ayrılmış ve kümeleme analizinin sonucunda, Türkiye'nin sağlık alanındaki harcamalarının en fazla Estonya, Letonya, Meksika, Çek Cumhuriyeti, Lüksemburg, Belçika ve Avustralya ile benzediği tespit edilmiştir.

# 3. KÜMELEME ANALİZİ

Kümeleme analizi, verilerin benzerliklerini belirleme ve bu benzerlikleri baz alarak nesneleri doğru bir şekilde kümelere ayırma işlemidir. Böylece küme içinde yüksek benzerlik ve kümeler arası düşük benzerlik gözlemlenir (Şekil 3.1). Analiz bu yönüyle, çok değişkenli analiz yöntemleri arasında yer alan diskriminant analizi ile benzerlik göstermektedir. Diskriminant analizi, hangi gruptan geldiği bilinmeyen bir gözlemin dahil edileceği grubu bulmaya yarayan bir yöntemdir. Diskriminant analizinde gruplar bulunmaktadır. Diskriminant analizinden farklı olarak kümeleme analizinde birimlerin anlık durumu gözlenirken, tahmin yapmak mümkün olmamaktadır. Diskriminant analizi, bir gözlemin daha önceden bilinen bir gruba eklenmesi için kullanılırken tahminde bulunulmasına olanak sağlamaktadır. Veri setinin olağan sınıflamaları hakkında gerekli bilginin bulunmaması durumunda, alt gruplara ilişkin yapıları oluşturmada kümeleme analizi kullanılmaktadır. Buna karşın olağan sınıflamaların bilinmesi durumunda alt kümelerin tetkik edilmesi diskriminant analizi aracılığıyla gerçekleştirilir. Kümeleme analizi ile faktör analizi kıyaslandığında kümeleme analizinin amacının nesneleri gruplamak, faktör analizinin amacının ise değişkenleri gruplamak olduğu söylenebilir. Bununla birlikte faktör analizi modelinde gruplandırmalar, değişkenler arasındaki korelasyona göre gerçekleştirilirken, kümeleme analizindeki gruplandırmalar uzaklık ölçülerine göre gerçekleştirilir (Çokluk vd. 2014).

Veri madenciliği, istatistik, biyoloji ve makine öğrenmesi gibi birçok alanda kümeleme analizinin yaygın olarak yer aldığı görülmektedir. Kümeleme analizinin kullanıldığı birçok farklı alan göz önüne alındığında, bir kümenin genel kabul görmüş bir tanımının olmaması şaşırtıcı değildir. Biyolojide bitki veya hayvan türlerinin sınıflandırılması, hastalıkların tıbbi sınıflandırılması, arkeolojide yerleşim yerlerinin ve dönemlerin keşfi ve segmentasyonu, görüntü segmentasyonu ve nesne tanıma, sosyal tabakalaşma, pazar segmentasyonu, verimli arama sorguları için veri tabanlarının organizasyonu gibi örnekler verilebilir. Bununla birlikte kümeleme analizinin uygulandığı genel konular; veri kümelerinin yapılandırılması ve veri analizinde yaşanacak karışıklığın önüne geçilmesi şeklinde kısaca ifade edilebilir.




Şekil 3.1 Nesnelerin benzerliklerine göre kümelere ayrılması

Kümelerin istenebilecek ve mevcut veriler kullanılarak kontrol edilebilecek potansiyel özelliklerinin bir listesi aşağıdaki gibi verilebilir.

- Küme içi farklılıklar küçük olmalıdır.
- Kümeler arası farklılıklar büyük olmalıdır.
- Bir kümenin üyeleri, ağırlık merkezi tarafından iyi temsil edilmelidir.
- Kümeler, veri kümesindeki yüksek yoğunluklu alanlara karşılık gelmelidir.
- Kümelere karşılık gelen veri kümesindeki alanlar belirli özelliklere sahip olmalıdır (dışbükey veya doğrusal olma gibi).
- Kümeleri az sayıda değişken kullanarak karakterize etmek mümkün olmalıdır.
- Kümeler, harici olarak verilen bir bölüme veya kümelemeyi hesaplamak için kullanılmayan bir veya daha fazla değişkenin değerlerine uygun olmalıdır.
- Küme sayısı az olmalıdır.

Bu özellikler ölçülürken kesin hale getirilmeleri ve detaylı tanımlanmaları önemlidir. Örneğin, birkaç kümeyle istikrara ulaşmak genellikle daha kolaydır ancak kümelerin daha homojen olması gereken durumlarda fazla küme gerekebilir. Bu tür özellikler hakkında karar vermek, kümeleme amacına uygun bir kümeleme yöntemini seçmekle mümkündür.

Kümeleme analizi tahmin amaçlı değildir ve varsayımları bulunmamaktadır. Değişkenler arasında oluşabilecek çoklu bağlantı ve aşırı gözlem sorununa dikkat etmek gerekir (Rençber 2020).

Kümeleme analizi veri matrisinin oluşturulması, benzerlik/mesafe matrislerinin hesaplanması, kümeleme yönteminin belirlenmesi ve sonuçların değerlendirilmesi şeklinde dört aşamadan oluşmaktadır. Kümelerin belirlenmesinde yöntemler hiyerarşik (aşamalı) ve hiyerarşik olmayan (aşamalı olmayan) yöntemler olmak üzere ikiye ayrılır. Hiyerarşik kümeleme yöntemlerinde, ya birimlerin her biri ilk başta tek başına bir küme olarak düşünülüp sonrasında benzerlik veya farklılıklarına göre kümeler oluşturulur (toplamalı kümeleme yöntemleri) ya da tüm birimler ilk başta tek bir küme olarak düşünülüp sonrasında birbirine benzemeyen noktalar kümeden çıkarılarak kümelemeye devam edilir (bölümlemeli kümeleme yöntemleri) (Özdamar 1999). Buna karşılık, hiyerarşik olmayan kümeleme yöntemlerinde $n$ birimin $k$ adet kümeye parçalanması hedeflenmektedir, yani bu yöntemlerde kümelerin sayısı önceden bilinmektedir (Özdamar 1999).

### 3.1 Uzaklık Ölçüleri

Kümeleme analizinde amaç, grupları birbirine benzer nesneleri içerecek şekilde oluşturmaktır. Bu nedenle kümeleme algoritmaları nesneler arasındaki uzaklıkları hesaplar. Uzaklık, bir çeşit benzerlik ölçüsüdür, yani nesneler arasındaki mesafe fazla ise birbirlerine benzemedikleri düşünülebilir (Akküçük 2011).

$X$ ve $Z$ iki farklı vektör olarak tanımlanarak $X = \left(x_1, x_2, \dots, x_p\right)^T$ ve $Z = \left(z_1, z_2, \dots, z_p\right)^T$ şeklinde ifade edilsin. Bu durumda $X$ ve $Z$ arasındaki uzaklık $D = \|X - Z\|$ şeklinde gösterilir. $D$ değeri ne kadar küçükse $X$ ve $Z$'nin o kadar benzer olduğu söylenebilir. ($D$, $p$ boyutlu uzayda $X$ ve $Z'$nin mesafesidir.)

Literatürde oldukça fazla uzaklık ölçüsü tanımlanmış olmasına karşın bu çalışmada kümelem

### 3.1.1 Minkowski uzaklığı

Minkowski, genel bir uzaklık ölçüsüdür. $X$ ve $Z$ iki farklı vektör olarak tanımlansın ve $X = \left(x_1, x_2, \dots, x_p\right)^T$ ve $Z = \left(z_1, z_2, \dots, z_p\right)^T$ şeklinde ifade edilsin. Bu durumda $X$ ve $Z$ arasındaki uzaklık $D = \|X - Z\|$ şeklinde gösterilir (Tatlıdil 1996).

$$D = \|X - Z\| = \left[\sum_{i=1}^{p} |x_i - z_i|^{\lambda}\right]^{1/\lambda}, \quad \lambda \geq 1 \tag{3.1}$$

### 3.1.2 Manhattan City Block uzaklığı

Minkowski uzaklığında $\lambda$=1 olması durumudur. Manhattan City Block uzaklığı birimler arasındaki mutlak fark değerlerinin toplamından elde edilir.

$$D = \|X - Z\| = \sum_{i=1}^{p} |x_i - z_i| \tag{3.2}$$

ile gösterilir (Tatlıdil 1996).

### 3.1.3 Öklid uzaklığı

Nicel veriler için en sık kullanılan uzaklık ölçüsüdür. Minkowski uzaklığında $\lambda$=2 olması durumudur. Gözlemler yoğun, kümeler arası ayrıklık fazla olduğunda Öklid uzaklığı daha iyi çalışmaktadır. Veri noktaları arasındaki uzaklık azaldıkça Öklid uzaklığı da sıfıra yaklaşır. Öklid uzaklığı,

$$D = \|X - Z\| = \left[\sum_{i=1}^{p} (x_i - z_i)^2\right]^{1/2} \tag{3.3}$$

ile gösterilir (Everitt vd. 2011).

### 3.1.4 Karesi alınmış Öklid uzaklığı

Öklid uzaklığının karesinin alınmasıyla elde edilmektedir. Değişkenler farklı ölçeklendiğinde kullanılan uzaklık ölçüsüdür. Öklid uzaklığından farklı olarak toplamın karekökü alınmadığı için uç değerlere karşı daha hassastır. Karesi alınmış Öklid uzaklığı,

$$D = \|X - Z\| = \sum_{i=1}^{p}(x_i - z_i)^2 \tag{3.4}$$

ile gösterilir (Özdamar 1999).

### 3.1.5 Mahalanobis uzaklığı

İki değişken arasındaki kovaryans veya korelasyon katsayısını dikkate alan bir uzaklık ölçüsüdür. Bu nedenle değişkenler arasında korelasyon bulunduğunda kullanılması önerilen bir uzaklık ölçüsüdür. Mahalanobis uzaklığı,

$$D = \|X - Z\| = (x_i - z_i)^T S^{-1}(x_i - z_i) \tag{3.5}$$

ile gösterilir. Burada S, örnekleme ait varyans-kovaryans matrisidir (Tatlıdil 1996).

### 3.1.6 Canberra uzaklığı

Veri matrisi değişkenlerinin pozitif olduğu durumlarda kullanılan ve ölçek duyarlılığı olmayan bir uzaklık ölçüsüdür. Canberra uzaklığı,

$$D = \|X - Z\| = \frac{\sum_{i=1}^{p}|x_i - z_i|}{\sum_{i=1}^{p}(x_i + z_i)} \tag{3.6}$$

ile gösterilir.

Bu uzaklık ölçüleri nicel değişkenler ile kullanılmakta olup nitel değişkenler söz konusu olduğunda yukarıdaki formüllere aşağıda yer alan $w_i$ katsayısı eklenmelidir (Tatlıdil 1996).

$$w_i = \begin{cases} 1 & ,nicel\ veriler\ için \\ \frac{1}{i'inci\ değişkenin\ dağılım\ aralığı} & ,nitel\ veriler\ için \end{cases} \quad (3.7)$$

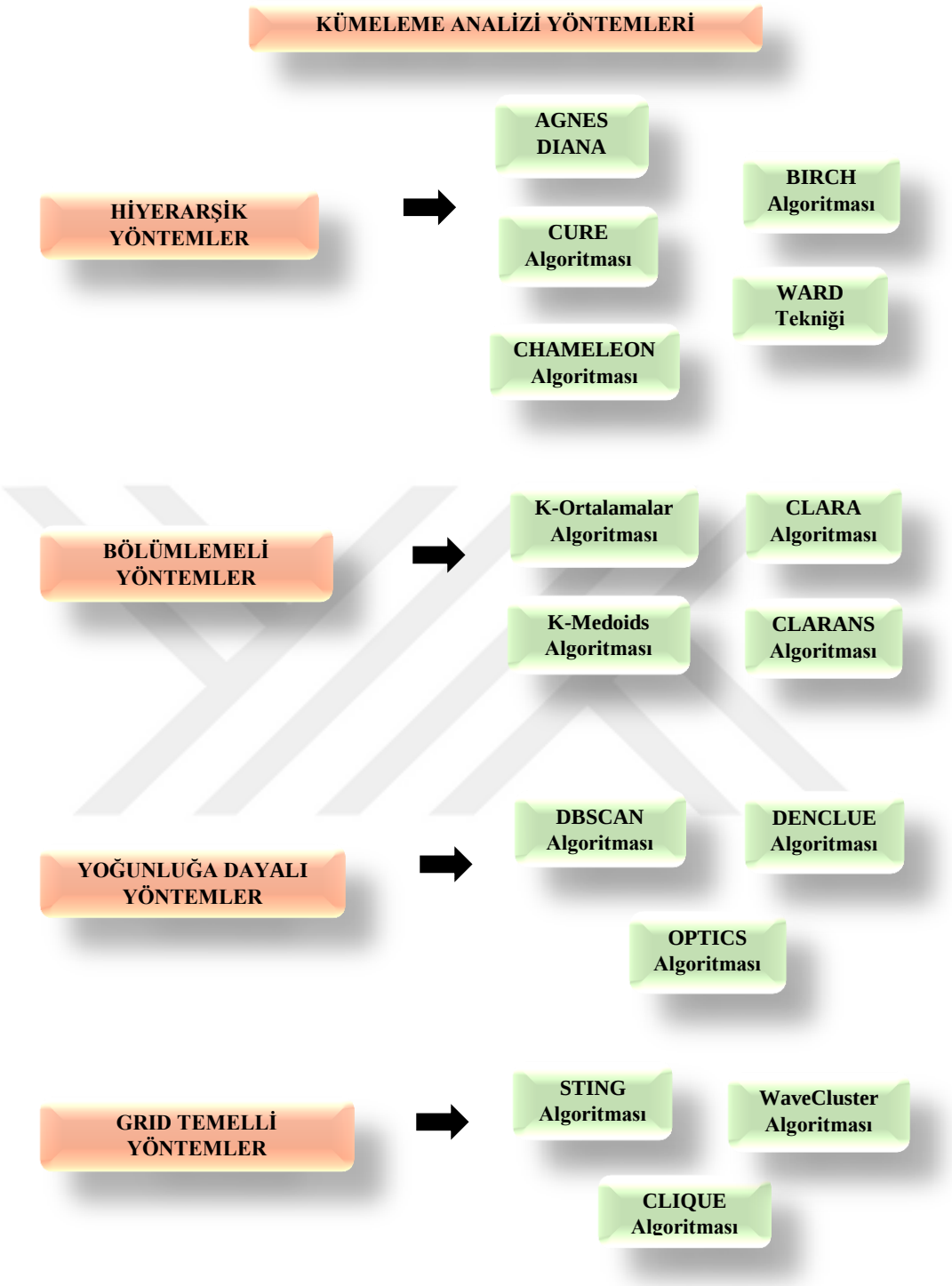
olmak üzere

$$D = \|X - Z\| = \frac{1}{p} \sum_{i=1}^{p} w_i \, |x_i - z_i|$$

ile bulunur.

## 3.2 Kümeleme Analizi Yöntemleri

Kümeleme algoritmalarının genel anlamda hiyerarşik ve hiyerarşik olmayan yöntemler olarak ikiye ayrılır. Ancak literatürde Şekil 3.2'de görüldüğü üzere birçok algoritmanın adı geçmektedir. Algoritmalar kümelemenin oluşturulma şekline, kullanılan verinin çeşidine ve yapılacak çalışmanın amacına göre farklılık gösterir (Silahtaroğlu 2020).




Şekil 3.2 Kümeleme analizinin sınıflandırılması

### 3.2.1 Hiyerarşik yöntemler

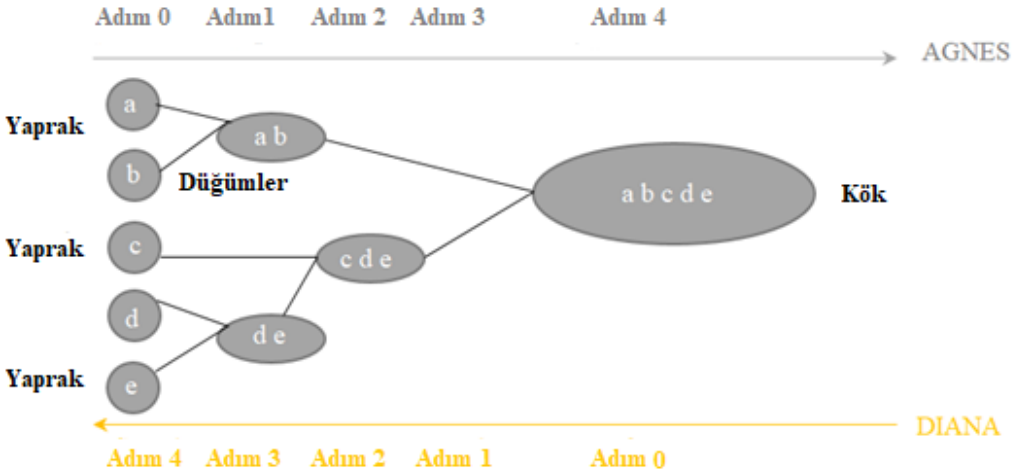
Hiyerarşik kümeleme yöntemi, verilerin benzer özelliklere göre birleştirilmesi veya bölünerek iç içe kümeler oluşturulmasıyla elde edilen genel bir kümeleme algoritmasıdır. (Rençber 2020). Hiyerarşik yöntemler ile bir küme ağacı oluşturulur. Hiyerarşik ağaç kümeleri oluşum şekline göre iki grupta incelenebilir. Bunlardan birincisi aşağıdan yukarıya toplamalı kümeleme algoritmaları, ikincisi ise yukarıdan aşağıya bölünür kümeleme algoritmalarıdır (Silahtaroğlu 2020).

Hiyerarşik kümeleme uzaklık ölçülerini kullandığından neredeyse her türlü veri türüne uygundur. Ancak bölünür algoritmalar için $k$ küme sayısının önceden belirlenmemesi ve bu algoritmaların oluşturulan kümeyi bir daha kontrol edememesi dezavantajdır. Yani hiyerarşik kümeleme yönteminde bir kez işlem yapıldığında, (toplamalı veya bölünür) bu işlem geri alınamaz ve aralarında nesne değişimi yapılamaz. Bu durumda hatalı kararlar da düzeltilemez. Bu katılık, farklı seçeneklerin birleşim sayısı hakkında endişe duymadan daha düşük hesaplama maliyetlerine yol açtığından yararlı olabilmektedir. Hiyerarşik yöntemler alt başlığında incelenecek algoritmalardan bazıları işlemlerin geri alınamaması durumuna çözümler geliştirmiştir.

#### 3.2.1.1 AGNES – DIANA hiyerarşik kümeleme

Toplamalı kümeleme algoritmaları (AGNES), veri tabanında yer alan her bir noktayı küme olarak görür. Her bir adımda nesneleri benzerliklerine göre birleştirerek ilerler. Benzer olanlar tek bir küme oluncaya kadar veya sonlandırma şartı sağlanana kadar ilerleme devam eder.

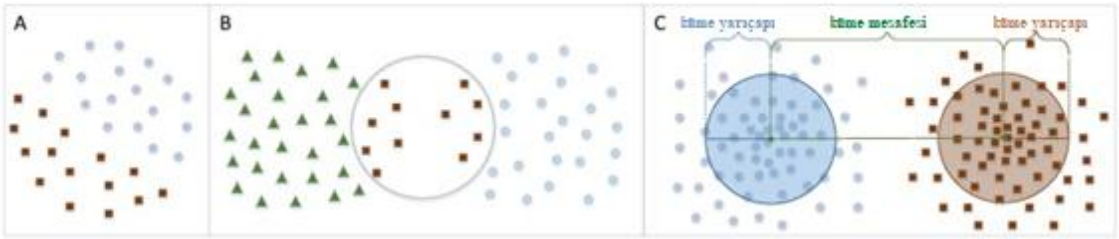
Bölünür kümeleme algoritmaları (DIANA) ise toplamalı kümeleme algoritmalarının aksine başlangıçta veri tabanındaki bütün noktaları tek bir küme olarak görür. Sonraki her adımda birbirine benzemeyen noktaları kümeden çıkararak her nesne tek bir küme oluncaya kadar veya sonlandırma şartı sağlanana kadar ilerler.




Şekil 3.3 AGNES – DIANA hiyerarşik kümeleme

### 3.2.1.2 BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies) algoritması

BIRCH algoritması hiyerarşik kümelemeyi (ilk mikro kümeleme aşamasında) ve yinelemeli bölümleme gibi diğer kümeleme yöntemlerini entegre ederek büyük miktarda sayısal veriyi kümelemek için tasarlanmıştır. Kümeleme analizinde karşılaşılan ölçekleme ve önceki adıma dönmenin mümkün olmaması sorunlarına çözümler getirmiştir.

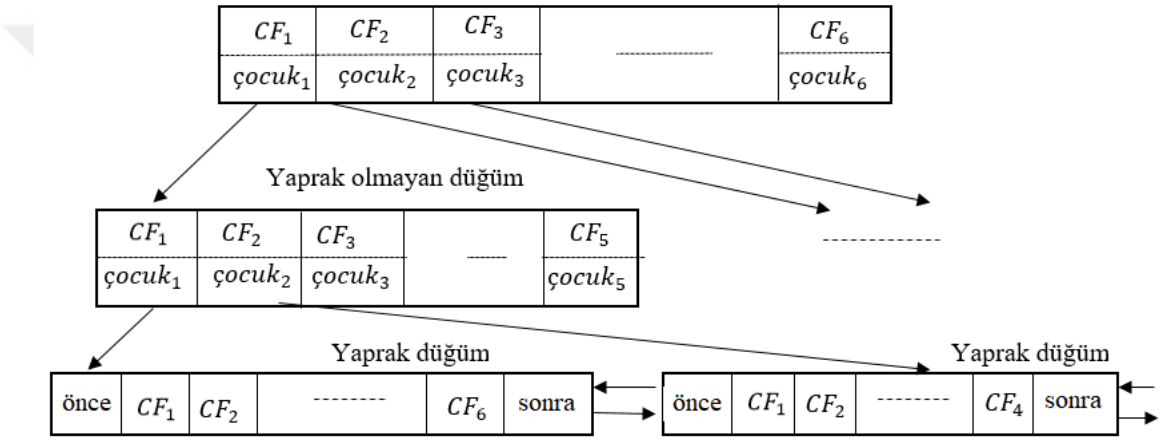



Şekil 3.4 Kümenin bölünmesi (A), daire içine alınan alanda kümenin oluşması (B), küme uzaklığının ve yarıçapının gösterilmesi (C) (Lorbeer vd. 2018)

Gözlemlerin farklı biçimleri ve renkleri, ait oldukları kümelere karşılık gelir. (Şekil 3.4) BIRCH algoritması kümeye ait sınırın kontrolü için yarıçap veya çap kavramını

kullandığından kümelerin şekli küresel değilse iyi performans göstermez ve sayısal olmayan verileri işlemede başarılı değildir.

Algoritmada kümedeki tüm nesnelere ilişkin bilgileri içeren kümeleme özelliği, CF (clustering feature) ile özetlenmektedir. Hiyerarşik kümeleme analizi için CF ağacı oluşturulur. Her düğüm, küme hiyerarşisinde bir kümeyi temsil eder, ara düğümler üst kümelerdir ve yaprak düğümler gerçek kümelerdir. Dallanma faktörü, bir düğümün sahip olabileceği maksimum çocuk sayısıdır. Bu evrensel bir parametredir. Her düğüm, ait olduğu kümenin en önemli bilgisini, küme özelliklerini (CF) içerir.




Şekil 3.5 BIRCH algoritmasında kullanılan CF ağacı (Andritsos 2002)

BIRCH algoritması şu şekilde çalışır:

Adım 1: Başlangıç CF Ağacı: Veriler birer birer yüklenir ve ilk CF ağacı oluşturulur. En yakın yaprak girişine, yani alt kümeye bir nesne eklenir. Bu alt kümenin çapı daha büyük olursa, mümkün durumda yaprak düğüm bölünür. Veri bir yaprak düğüme uygun şekilde yerleştirildiğinde, ağacın köküne doğru olan tüm düğümler gerekli bilgilerle güncellenir.

Adım 2: Küçültülmüş CF Ağacı: CF ağacı belleğe sığmıyorsa, bazı alt kümeler birleştirilerek ağaç küçültülebilir. Bu aşama isteğe bağlıdır.

Adım 3: Kümeleme: CF ağacının yaprak düğümleri alt küme istatistiklerini tutar. Algoritma, bu istatistikleri bazı kümeleme tekniklerini uygulamak için kullanır.

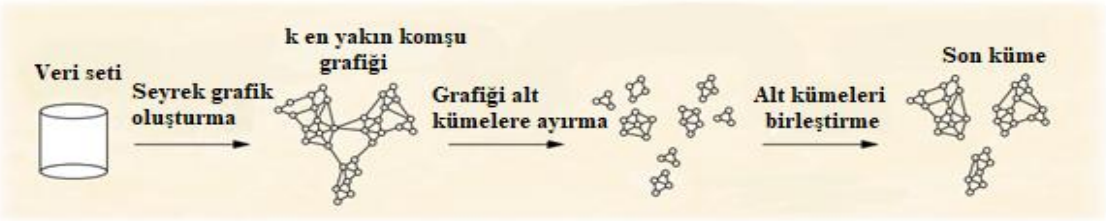
Adım 4: Kümelerin Revize Edilmesi: Keşfedilen kümelerin ağırlık merkezleri kullanılarak veri nesneleri yeniden dağıtılır. Bu isteğe bağlı bir aşamadır. Veri kümesinin ek bir taramasını gerektirir ve nesneleri en yakın merkezlerine yeniden atar. İsteğe bağlı olarak gerçekleştirilen bu aşama ayrıca aykırı değerlerin atılmasını da içerir (Andritsos 2002).

### 3.2.1.3 CHAMELEON algoritması

CHAMELEON algoritmasının temel özelliği en benzer küme çiftinin belirlenmesinde aradaki bağlantısallığı ve yakınlığı hesaba katmasıdır. Ayrıca her bir küme çifti arasında yakınlık ve karşılıklı bağlantının derecesini modelleyebilmek için yeni bir yaklaşım kullanır. Bu yaklaşımda kümelerin kendilerine ait özellikleri dikkate alınır. Böylece kullanıcı kaynaklı ve statik olmayan bir modele bağlıdır ve birleştirilmiş kümelerin iç özelliklerine otomatik olarak uyum sağlayabilir.

Algoritma, seyrek bir grafik üzerinde çalışır. Düğümler veri öğelerini, ağırlıklı kenarlar ise veri öğeleri arasındaki benzerlikleri temsil eder. Bu seyrek grafik gösterimi, algoritmanın büyük veri kümelerine ölçeklenmesine ve yalnızca benzerlik alanında bulunan ve metrik uzaylarda olmayan veri kümelerini başarıyla kullanmasına olanak tanır.

CHAMELEON kümeleri meydana getirirken iki aşamadan oluşan bir algoritma kullanır. Birinci aşamada, veri öğelerini görece küçük alt kümelere ayıran grafik bölümleme algoritması kullanır. İkinci aşamada ise bu alt kümeleri birleştirerek gerçek kümelere ulaşmak için farklı bir algoritma kullanır. (Şekil 3.6) ile CHAMELEON algoritmasının aşamaları verilmiştir.




Şekil 3.6 CHAMELEON algoritmasının aşamaları (Karypis vd. 1999)

$C_i$ ve $C_j$ kümelerine ait tüm bağlantıların (kenarların) ağırlıklarının toplamı, $C_i$ ve $C_j$ arasındaki mutlak bağlantısallık olarak adlandırılır ve $\left|EC\left(C_i, C_j\right)\right|$ ile ifade edilir. Göreli bağlantı ise kümeler arasında, kendi iç bağlantılarına göre normlaştırılmış mutlak bağlantısallıklardır. $RI\left(C_i, C_j\right)$ ile ifade edilerek

$$RI\left(C_i, C_j\right) = \frac{\left|EC\left(C_i, C_j\right)\right|}{\frac{\left|EC\left(C_i\right)\right| + \left|EC\left(C_j\right)\right|}{2}} \tag{3.8}$$

biçiminde formülize edilir.

Göreli bağlantısallık, farklı kümeler için bağlantı derecesindeki farklılıkların yanı sıra küme şekillerindeki farklılıkları da hesaba katabilir.

İki kümenin birleştirilmesinde CHAMELEON algoritmasının kullandığı diğer ölçü, göreli yakınlık ölçüsüdür. Göreli yakınlık, mutlak yakınlık değerinin iki kümenin iç yakınlığına göre normlaştırılmış halidir. $C_i$ ve $C_j$ kümeleri arasındaki göreli yakınlık

$$RC\left(C_j, C_j\right) = \frac{\bar{S}EC\left(C_i, C_j\right)}{\frac{\left|C_i\right|}{\left|C_i\right| + \left|C_j\right|}\bar{S}EC\left(C_i\right) + \frac{\left|C_j\right|}{\left|C_i\right| + \left|C_j\right|}\bar{S}EC\left(C_j\right)} \tag{3.9}$$

olarak hesaplanır. Burada,

$|C_i|$: $i$. kümedeki kenarların toplamı

$\left|C_j\right|$: $j$. kümedeki kenarların toplamı

$\bar{S}EC\left(C_i\right)$: iç yakınlık ($C_i$ kümesini kendi içinde kabaca ikiye ayıran kenarların ortalama ağırlığı)

$\bar{S}EC\left(C_j\right)$: iç yakınlık ($C_j$ kümesini kendi içinde kabaca ikiye ayıran kenarların ortalama ağırlığı)

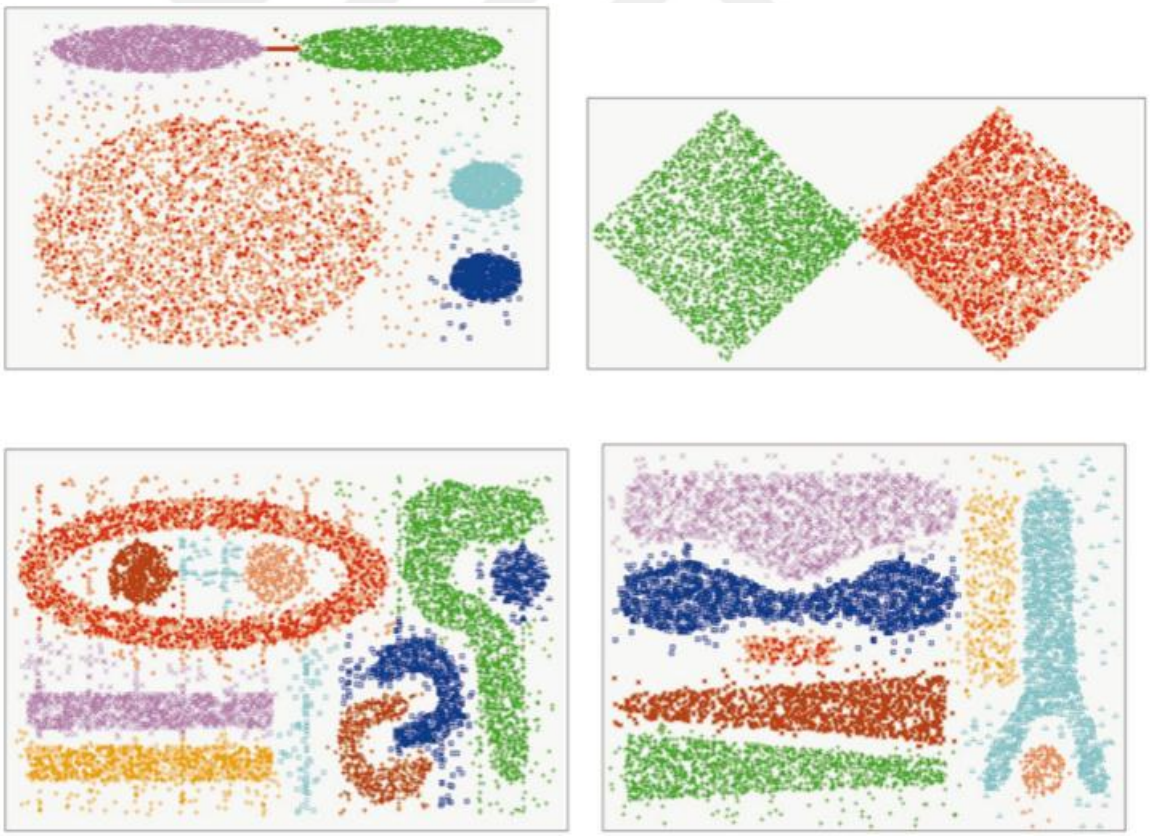
$\bar{S}EC\left(C_i, C_j\right)$: mutlak yakınlık ($C_i$ ve $C_j$ kümelerini birbirine bağlayan kenarların ortalama ağırlığı)

şeklinde ifade edilir.

$C_i$ ve $C_j$ kümelerinin birleştirilmesi için bu kümeler arasında yüksek bir göreli yakınlık değeri ve göreli bağlantısallık hesaplanması gerekmektedir. Yani $RI(C_i, C_j) \times RC(C_i, C_j)$ formülünü maksimize edecek kümeler, birleştirmek üzere seçilmelidir. Bu formül her iki parametreye $(RI, RC)$ de eşit önem verir. Fakat iki ölçümden birine daha yüksek eğilim gösteren kümeler tercih edilebilir. Bu nedenle CHAMALEON, kullanıcı tarafından belirlenen $\alpha$ değerinin maksimize olduğu yerdeki küme çiftini seçer.

$$\max\left(RI(C_i, C_j) \times RC(C_i, C_j)^{\alpha}\right) \tag{3.10}$$

$\alpha$>1 olduğunda göreli yakınlığa, $\alpha$<1 olduğunda ise göreli bağlantısallığa daha fazla önem verilmiş anlamına gelir (Karypis vd. 1999).



Şekil 3.7 CHAMELEON algoritmasının dört farklı veri setinde çalıştırılması (Karypis vd. 1999)

Şekil 3.7'de görüldüğü gibi algoritma tarafından veri nesnelerinin kümelere ayrılması başarılı bir şekilde gerçekleştirilmiştir. Birbirinden farklı küme şekillerinin ortaya çıkmasının kümeleme başarısını etkilemediği söylenebilir.

### 3.2.1.4 Ward tekniği

J.H. Ward tarafından 1963 yılında ortaya atılan Ward tekniği, hata kareler toplamına dayanmaktadır. Amaç, her aşamada küme içi hata kareler toplamındaki artışı en aza indirmektir (Ward 1963). Kümelere ait hata kareler toplamını veren '$E$' toplamı şu şekilde bulunmaktadır:

Adım 1:

$$\bar{x}_{m,k} = 1/n_m \sum_{i=1}^{n_m} x_{m,k} \tag{3.11}$$

$\bar{x}_{m,k}$: $m$. kümedeki $k$. değişene ait ortalamayı verir.
$n_m$: $m$. kümedeki birim sayısını ifade eder.

Adım 2:

$$E_m = \sum_{i=1}^{n_m} \sum_{k=1}^{p_k} \left(x_{m,k} - \bar{x}_{m,k}\right)^2 \tag{3.12}$$

$E_m$ : Küme içi hata kareler toplamını ifade eder.
$p_k$ : Değişken sayısını ifade eder.

Adım 3:

$$E = \sum_{m=1}^{g} E_m \tag{3.13}$$

$E$: Tüm kümelere ait hata kareler toplamını verir.

Kümeler her aşamada birleştirilirken hata kareler toplamı $E$'deki değişimin minimum olmasına dikkat edilmelidir (Everitt vd. 2011).

Hiyerarşik kümeleme yöntemleri varsayımlarında bahsedildiği üzere küme sayısı önceden bilinmediğinden, Ward yönteminde, analiz sonucunda oluşan dendogramlar yardımıyla küme sayısı belirlenebilir. Dendogram, benzer veri kümeleri arasındaki ilişkileri gösteren bir tür ağaç diyagramı olarak tanımlanabilir. Bu ağaç diyagramı, yapraklar ve dallardan oluşur. Her dalın bir veya daha fazla yaprağı vardır. Dallar benzerliklerine göre düzenlenir. Aynı yüksekliğe yakın olan dallar birbirine benzer olarak yorumlanır. Ayrıca yükseklik farkı ne kadar fazla ise benzerlik de o kadar fazladır. Küme sayısına karar verirken kabaca, en uzun daldan çizilen yatay çizginin küme sayısını verdiği söylenebilir.

#### 3.2.1.5 CURE (Clustering Using Representatives) algoritması

Kümeleme işlemi yapılırken, küme kalitesine en çok etki eden faktör, veri setinde yer alan diğer verilerin uzağında, sayısı az ve hiçbir kümeye dâhil olmaması gerektiği düşünülen uç verilerdir. CURE algoritması uç verilerin, oluşturulan kümelerin kalitesini etkilememesi düşüncesinden yola çıkarak Guha, Rastogi ve Shim tarafından 1998 yılında geliştirilmiş bir algoritmadır (Guha vd. 1998). Küme merkezleri ile veri noktaları arasında denge meydana getiren bir algoritmaya sahiptir. Dairesel olmayan kümeler oluşturan veri grupları için oldukça elverişli bir algoritmadır (Silahtaroğlu 2020). CURE algoritmasının çalışmasına ilişkin adımlar aşağıdaki gibidir.

Adım 1: Her bir girdi ayrı bir küme olarak değerlendirilir ve her adımda bu kümeler temsilcilerin birbirlerine yakınlıklarına göre veya belirlenen $k$ küme sayısına ulaşılıncaya kadar birleştirilir.
Adım 2: Her bir küme için $c$ adet iyi dağıtılmış temsilci nokta seçilir.
Adım 3: Dağıtılmış temsilci noktalar, bir $\alpha$ katsayısıyla kümenin merkezine kaydırılır. ($\alpha$= daraltma katsayısı)
Adım 4: Her bir kümeye ait olan en yakın temsilci çifti arasındaki uzaklık iki küme arasındaki uzaklığı verir. Uzaklık, veriye uygun uzaklık ölçüsü ile hesaplanarak elde edilir.
Adım 5: En yakın temsilci çiftine göre kümeler birleştirilir.

Adım 6: Oluşan yeni kümenin temsilci noktaları bulunur. Bu noktalar α katsayısı ile çarpılarak merkeze doğru kaydırılır ve algoritmanın adımları $k$ küme sayısına ulaşılana dek tekrar edilir (Silahtaroğlu 2020).

CURE, kümeden iyi dağılmış noktalar seçilerek oluşturulan her küme için birden fazla temsili nokta kullanır ve belirli bir oranda kümenin ortasına doğru bir küçülme ile yeni kümeler oluşturur. Bu, küresel olmayan şekillere ve farklı varyanslara sahip kümelerin geometrisine iyi uyum sağlamasına olanak tanır. Büyük veri tabanlarını işlemek için CURE algoritması, bu veri tabanlarının verimli bir şekilde işlemesine olanak tanıyan bir rastgele örnekleme ve bölümleme kombinasyonunu kullanır. Rastgele örnekleme aykırı değer işleme teknikleriyle birleştiğinde CURE algoritması, veri setinde yer alan aykırı değerlerin etkili bir şekilde filtrelemesini de mümkün kılar. Ayrıca, CURE algoritmasındaki etiketleme yöntemi, her bir kümedeki veri noktalarını atamak için birden çok rastgele temsili nokta kullanır. Bu, kümelerin şekilleri küresel olmadığında ve kümelerin boyutları değiştiğinde bile noktaları doğru şekilde etiketlemesini sağlar. Rastgele bir örnek boyutu $n$ için, CURE algoritmasının zaman karmaşıklığı düşük boyutlu veriler için $O(n^2)$'dir ve uzay karmaşıklığı $n$'de doğrusaldır. En kötü durumdaki zaman karmaşıklığı ise $O(n^2 \log n)$'dir. Büyük veri tabanlarında CURE algoritmasının etkinliğini incelemek için yapılan kapsamlı deneyler sonucunda mevcut algoritmalardan daha iyi performans göstermesinin yanı sıra kümeleme kalitesinden ödün vermeden büyük veri tabanları için iyi ölçeklendiğini göstermektedir. Ayrıca algoritmanın uygulama süresi oldukça düşüktür ve aykırı değerlerin değerlendirilmesi için de iyi bir yöntemdir (Guha vd. 2001).

### 3.2.2 Bölümlemeli yöntemler

Bölümlemeli yöntemlerde veri seti ilk önce $k$ sayıda kümelere ayrılır. Burada $k$ parametresi oluşturulacak küme sayısını ifade eder. Ardından nesneleri bir gruptan diğerine taşıyarak bölümlemeyi iyileştirmeye çalışan yinelemeli bir yer değiştirme tekniği kullanır. Bu kümeleme teknikleri, veri noktalarının tek seviyeli bölümlenmesini oluşturur. Bu teknikler, bir merkez noktanın bir kümeyi temsil edebileceği fikrine dayanmaktadır. $K$-Ortalamalar için bir nokta grubunun ortalama veya medyan noktası

olan ağırlık merkezi kavramı kullanılır. Ancak bir ağırlık merkezi neredeyse hiçbir zaman gerçek bir veri noktasına karşılık gelmez. Benzer şekilde $K$-Medois algoritması için, bir nokta grubunun en merkezi noktası anlamına gelen bir medoid kavramı kullanılır.

Hiyerarşik yöntemlerin tersine, araştırmacı tarafından belirlenen kriterlere uygun küme oluşturulurken küme sayısı da önceden belirlidir. Ayrıca araştırmacı tarafından kümeler arasındaki maksimum/minimum mesafe ve küme içi benzerliklerin de verilmesi gerekir (Silahtaroğlu 2020).

Bölümlemeli kümeleme yöntemleri birçok problem alanına uyarlanmış basit ve temel yöntemlerdir. Nispeten hızlı olup anlaşılması ve uygulaması kolaydır. Benzerlik/mesafe matrisi kullanmanın zorunlu olmaması büyük verilerin kümelenmesinde hiyerarşik yöntemlere göre avantaj sağlamaktadır. Ayrıca önceden belirlenmiş kriterlere uygun birden fazla sonuç elde etmek mümkün olabilir. Algoritmanın gerçekten en uygun çözümü bulup bulamadığını ise bilmek mümkün değildir. Bunun öğrenilebilmesi için verilerin dağıtılması, yerlerinin değiştirilmesi, algoritmanın tekrar koşulması ve elde edilen sonuçların karşılaştırılması gerekir. Bu da zaman ve maliyeti arttıracaktır (Silahtaroğlu 2020).

#### 3.2.2.1 $K$-Ortalamalar ($K$-Means) algoritması

$K$-Ortalamalar algoritması, bir veri kümesini maksimum küme içi benzerlikler ve minimum kümeler arası benzerlikler olacak şekilde kümelere ayırmak şeklinde tanımlanabilir. $K$-Ortalamalar, iyi bilinen kümeleme problemini çözen en basit bölümlemeli yöntemlerden biridir. Belirli sayıda küme aracılığıyla belirli bir veri kümesini sınıflandırmak için basit ve kolay bir yol izler.

$K$-Ortalamalar, mesafeye dayalı bir kümeleme algoritmasıdır. Bu nedenle buradaki varyans, mesafeye göre hesaplanır. $K$-Ortalamalar kümeleme analizindeki mesafe, iki gözlem veya örnek arasındaki gerçek mesafeyi ifade etmemektedir. Algoritmadaki iki

gözlem veya örnek arasındaki benzerliği ölçmek için kullanılır. Farklı mesafe seçimine bağlı olarak benzerlik ölçümü de farklıdır.

Uzaklık ölçüsü kümelemeye ilişkin performansı etkilediğinden $K$-Ortalamalar ile gerçekleştirilen kümeleme analizinde en iyi kümeleme performansını veren uzaklık ölçüsünü kullanmak önemlidir.

$K$-Ortalamalar algoritması bazı prensiplere göre çalışır. Küme içindeki noktaların ortalama değerine göre kümenin merkezi tanımlanır. İlk olarak küme ortalamasını veya merkezini temsil eden kümeden nesneler rastgele seçilir. Kalan nesnelerin her biri için nesne ve küme arasındaki uzaklık hesaplanarak en benzer olan kümeye atama yapılır. Her yeni nesne atamasından sonra küme merkezleri önceki adımda kümeye atanan nesneleri kullanarak yeni ortalamayı hesaplar. Sonraki nesneler bu merkezlere olan uzaklıklarına göre yeniden atanır. Daha sonra algoritma yinelemeli olarak devam eder. Tüm nesneler güncellenmiş küme ortalamalarına göre yeniden atanır. Yinelemeler küme merkezleri sabit olana kadar devam eder.

Farklı $k$ değerleri ile yürütülen $K$-Ortalamalar uygulamaları sonucunda hangisinin en uygun olduğunu belirlemek üzere hata kareler toplamı kullanılır. Hata kareler toplamı minimum olan model, küme merkezlerinin kümelerde bulunan nesneleri en iyi temsil eden noktalar olduğu kabul edilir (Rençber 2020).

Hata kareler toplamı gösterimi aşağıda belirtildiği gibidir.

$$SSE = \sum_{j=1}^{k} \sum_{x \in C_j} dist\,(Cj, x)^2 \qquad (3.14)$$

$C_j$: Küme merkezi

$k$: Küme sayısı

dist $(x, y)$: Öklid uzaklığı

Verilerin çeşitlenmesi ve bu verilerden çıkarım yapmak için geliştirilen farklı $K$-Ortalamalar algoritmaları vardır. Bunlardan ilki çekirdek ağırlıklı $K$-Ortalamalar algoritmasıdır. Çekirdek ağırlıklı $K$-Ortalamalar algoritması, çalıştığı uzay bakımından

klasik yöntemden ayrılmaktadır. Doğrusal olmayan veriler ve kümeler için kullanılmaktadır. Yani, çekirdek ağırlıklı $K$-Ortalamalar algoritması, $K$-Ortalamalar algoritmasının doğrusal olmayan bir versiyonudur (Rençber 2020).

Bir diğer algoritma ise bulanık $K$-Ortalamalar algoritmasıdır. Bulanık mantık ile klasik mantık arasındaki temel fark, ortaya atılan önermelerin 0 ve 1'den farklı olarak 0.5 değeri alabileceği görüşüdür. İlk defa Zadeh (1965) tarafından duyurulan bulanık mantık, insan düşünce yapısına uyumu ile sıklıkla kullanılan istatistiksel ve matematiksel yöntemlere uyarlanarak geliştirilmesine olanak sağlamıştır. Zadeh, nesnelerin kümelere dahil olmasının [0,1] arasında değişecek bir üyelik derecesi ile gösterilmesini önermiştir. Bu üyelik derecesine göre nesnenin atanacağı küme belirlenmektedir. Dunn (1973) tarafından ortaya atılarak Bezdek (1981) tarafından geliştirilen bulanık $K$-Ortalamalar algoritmasının $K$-Ortalamalar algoritmasından farkı, bir nesnenin birden fazla kümeye ait olmasına imkan vermesidir (Rençber 2020).

$K$-ortalamalar ve bulanık $K$-Ortalamalar algoritmalarında küme sayısı önceden belirlenmesi gereken parametrelerdendir. Bu durumda optimum küme sayısının belirlenmesi için bazı yöntemlerden yararlanılmaktadır. Kümeleme analizinde optimum küme sayısının tespit edilmesi önemli bir sorundur. Bu sorunun çözümü için çeşitli yöntemler geliştirilmiş olsa da optimum küme sayısını en iyi şekilde tespit eden bir yöntemin varlığından söz etmek mümkün değildir (Rençber 2020). Küme sayısının belirlenmesi doğrudan veya istatistiksel yöntemlere dayalı olarak gerçekleştirilebilir. Doğrudan yöntemlerde küme içi uzaklıklar toplamı veya ortalamasının optimizasyonu temel alınırken istatistiksel yöntemler hipotez testlerine dayanmaktadır. Optimum küme sayısının belirlenmesinde kullanılan yöntemler, Dirsek (Elbow), Siluet (Silhouette )ve GAP İstatistik Yöntemleridir.

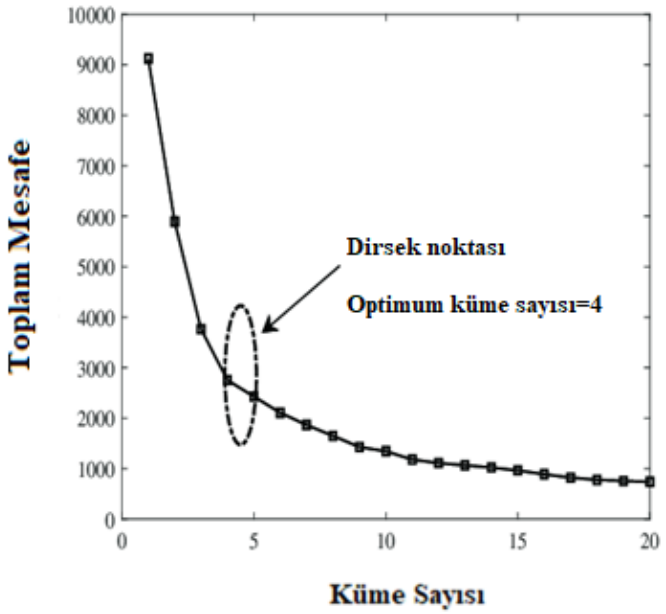
#### 3.2.1.1.1 Küme sayısını belirleme yöntemleri

***a) Dirsek (Elbow method) yöntemi***

Elbow yönteminde her bir noktanın küme merkezlerine olan uzaklıklarının karesi hesaplanarak ortaya çıkan değerler ile grafik çizilmektedir (Rençber 2020). Elbow

yöntemine göre optimum küme sayısı aşağıdaki işlemler ile tespit edilmektedir (Kassambara 2017).

Adım 1: Farklı $k$ değerleri (örneğin $k$ değeri 1 ile 5 arasında değiştirilerek) için kümeleme algoritması hesaplanır.
Adım 2: Her $k$ değeri için küme içi kareler toplamı (WSS) hesaplanır.
Adım 3: $k$ küme sayısına göre WSS eğrisi oluşturulur.
Adım 4: Grafikte oluşan dirsek noktası optimum küme sayısının göstergesi olarak kabul edilir.




Şekil 3.8 Elbow yöntemi örnek grafik (Zhang vd. 2020)

Şekil 3.8'de Elbow yöntemine örnek olarak gösterilen grafiğe göre optimum küme sayısı dirsek noktasına denk gelen 4 değeridir.

### *b) Ortalama siluet yöntemi (Average silhouette method)*

Rousseeuw (1987) tarafından oluşturulan ortalama siluet yöntemi her nesnenin kendi kümesine ne kadar uygun olduğunun grafiksel gösterimini sağlayan bir yöntemdir (Rençber, 2020). $-1$ ile 1 arasında değer alan siluet katsayısı, temelde her bir veri noktasının diğer verilere ne kadar benzer olduğunun bir ölçüsü olup $i$. veri için

$$S(i) = \frac{b(i)-a(i)}{\max\left(a(i),b(i)\right)} \tag{3.15}$$

biçiminde hesaplanır.

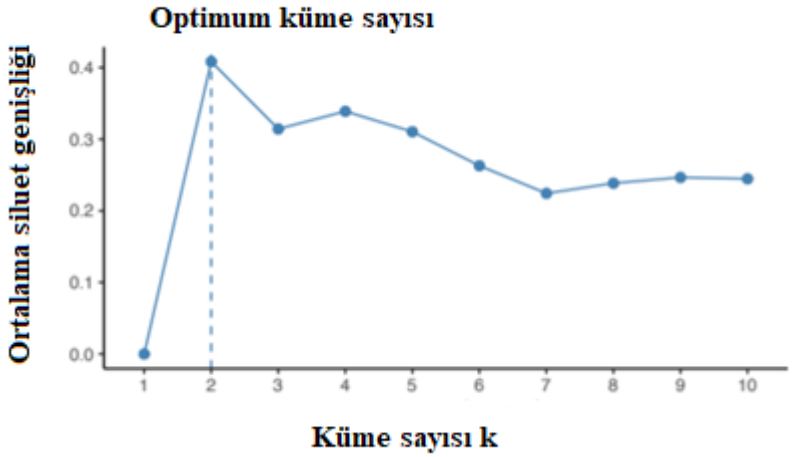
Adım 1: Belli bir $i$ noktasının kendi kümesinde bulunan tüm noktalara olan ortalama mesafesi hesaplanır. $[a(i)]$

Adım 2: Verilen $i$ noktasının en yakın komşu kümesindeki diğer tüm noktalara olan ortalama mesafesi hesaplanır. $[b(i)]$

Adım 3: $S(i)$ hesaplanır.

Adım 4: Veri kümesindeki her nokta için ilk üç adım tekrarlanır. Ardından o veri kümesi için genel siluet katsayısına ulaşmak için tüm siluet katsayılarının ortalaması alınır.

Şekil 3.9'da örnek olarak gösterilen grafiğe göre önerilen küme sayısı 2'dir. Siluet katsayısının 1'e en yakın olduğu değere göre karar verilmiştir. Verilerin iki kümede sınıflandırılması önerilmektedir.




Şekil 3.9 Ortalama siluet yöntemi örnek grafik (Kassambara 2017)

### *c) GAP istatistik yöntemi*

R. Tibshirani, G. Walther, ve T. Hastie (2001) tarafından oluşturulan GAP istatistik yöntemi optimum küme sayısının belirlenmesinde kullanılır. Herhangi bir kümeleme yöntemine uygulanabilir (Rençber 2020).

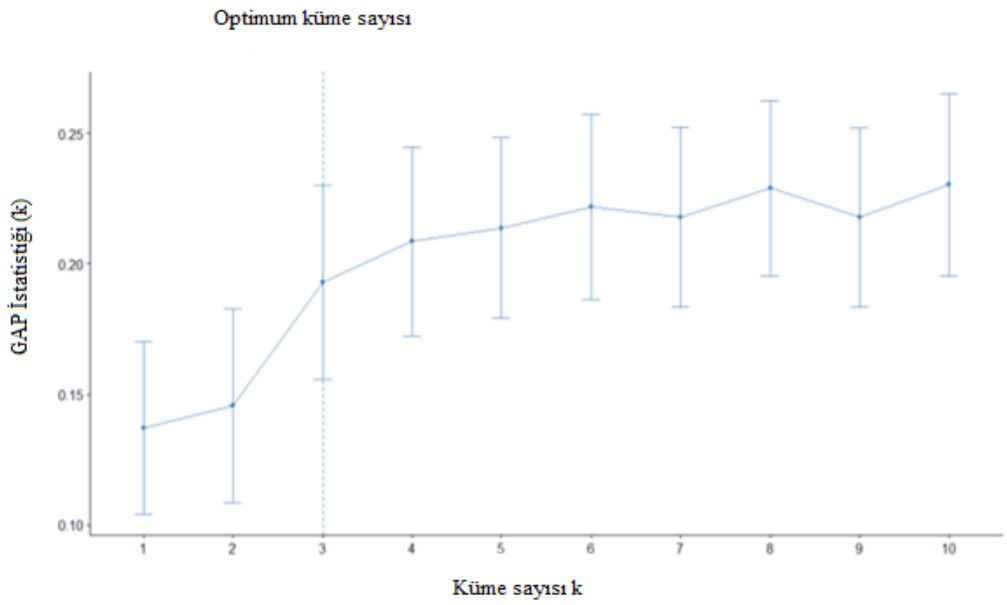
GAP istatistiği, kümeye ait gözlem değerleri ile merkezde yer alan noktalar arasındaki mesafenin ölçülmesi ile hesaplanır. Küme sayısının artmasıyla GAP istatistiği değerinde düşmenin gözlemlendiği nokta, optimum küme sayısı olarak değerlendirilmektedir. GAP istatistiği aşağıdaki gibi hesaplanmaktadır.

$$GAP_n(k) = E_n^*\{\log(w_k)\} - \log(w_k) \tag{3.16}$$

$k$: küme sayısı

$w_k$: küme içi kareler toplamı

$E_n^*$: n örneklem büyüklüğü altındaki beklenen durum



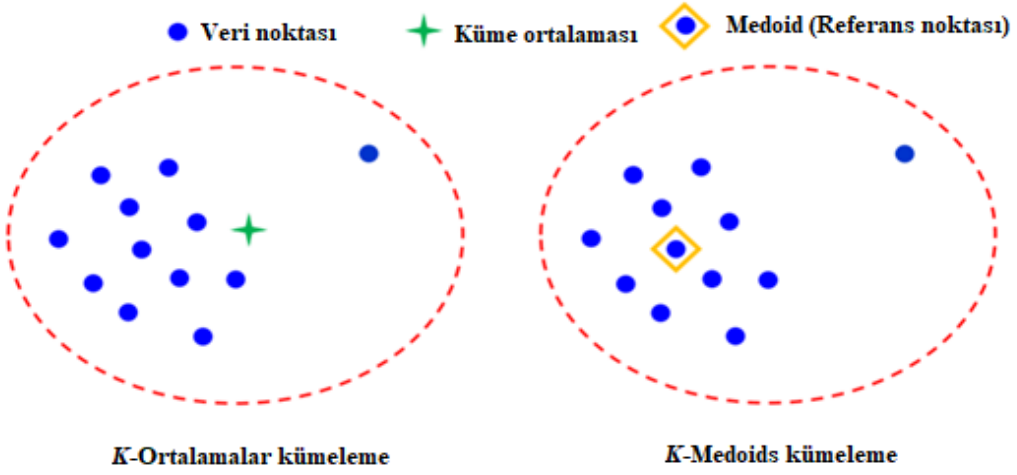

Şekil 3.10 GAP istatistik yöntemi örnek grafik (Melnykova vd. 2020)

Şekil 3.10'da örnek olarak gösterilen grafiğe göre GAP istatistik değerinin düşüş yaşadığı 3 değeri optimum küme sayısını ifade etmektedir.

### 3.2.2.2 *K*-Medoids algoritması

*K*-Ortalamalar algoritması aykırı değerlere karşı hassastır çünkü aşırı büyük değere sahip olan nesne, veriye ilişkin dağılımı önemli oranda değiştirebilir. Algoritmanın bu

duruma olan hassasiyetini azaltmak için referans nokta olarak bir kümenin en merkezi konumundaki nesne olan bir medoid kullanılması, bu kümede yer alan nesnelerin ortalamasına ilişkin değeri referans noktası olarak almak yerine kullanılabilir. $K$-Medoids algoritması, $K$-Ortalamalar algoritmasına benzemekle birlikte bu algoritmanın aykırı değerlere olan duyarlılığını azaltma amacı taşımaktadır (Rençber 2020). $K$-Medoids ve $K$-Ortalamalar algoritmalarının her ikisi de gözlem ile küme merkezi arasındaki hata kareler toplamını minimize ederek veri setini belirlenen $k$ adet kümeye bölmeyi amaçlar. Bu iki algoritma arasındaki fark ise, $K$-Ortalamalar algoritmasında genellikle referans noktası olarak ifade edilen merkez, gözlemlerin ortalaması iken $K$-Medoids algoritmasında referans noktası (medoid) kümenin en merkezi konumundaki temsili bir gözlem olarak seçilir. Başka bir deyişle, $K$-Medoids algoritmasındaki referans noktası küme içindeki diğer tüm gözlemlere olan toplam farklılıkları minimize eden gözlem olarak belirlenir (Şekil 3.11).




Şekil 3.11 $K$-Ortalamalar ve $K$-Medoids kümeleme algoritmaları arasındaki farkın grafiksel gösterimi (Entezami vd. 2020)

$K$-Medoids kümeleme algoritmasının temeli her bir kümenin gözlemlerine ait ortalamayı kullanmak yerine temsili gözlemleri referans olarak kullanmaya dayalı olduğundan gürültüye ve aykırı değerlere karşı dayanıklıdır. Şekil 3.11'de $K$-Medoids algoritmasının aykırı değere karşı dayanıklı olduğu açıkça görülmektedir. $K$-Medoids algoritmasının adımları aşağıdaki gibidir.

Adım 1: Uzaklık ölçüsü ve $k$ küme sayısı belirlenir.

Adım 2: Gözlemlerin veya noktaların $k$ kümeye ilk ataması oluşturulur.

Adım 3: Her bir küme için referans noktası (medoidler) bulunur. ( $k$. kümenin referans noktası o kümedeki diğer tüm nesnelere olan toplam mesafeyi en aza indiren nokta olacak şekilde belirlenir.)

Adım 4: $r$. küme ($r = 1,2, \dots k$) için $i$. gözlemin en yakın küme referans noktasına göre ataması amaç fonksiyonundaki azalma devam ettikçe sürer. (Amaç fonksiyonu, en yakın referans noktaları (medoidler) için bütün nesnelerin uzaklıkları toplamıdır.)

Adım 5: Başka bir gözlem ataması gerçekleşmeyinceye kadar üçüncü ve dördüncü adımlar tekrarlanır.

#### 3.2.2.3 CLARA (Clustering Large Applications) algoritması

CLARA algoritması büyük veri setlerinin oldukça kısa zamanda kümelenmesi için Kauffman ve Rousseeuw tarafından geliştirilmiştir. Algoritma, $K$-Medoids algoritmasında olduğu gibi tüm veri setini tarayarak referans noktalar belirlemek yerine rastgele bir örnek küme seçer (Nguyen vd. 2011). Bu örnek kümeye $K$-Medoids algoritması uygular ve uygulama sonucunda oluşacak olan kümelerin her biri için referans noktası belirlenir. Daha sonra veri setinden başka bir örnek küme seçilir. Bu aşamada referans noktalarını, ilk referans noktalarında olduğu gibi rastgele seçimle bulmak yerine bir önceki aşamada belirlenen noktalar kullanılır. Böylece referans noktası değişimi azaltılarak algoritmanın daha hızlı ve kaliteli sonuçlar vereceği ileri sürülmektedir. Kauffman ve Rousseeuw tarafından örneklemenin beş kez tekrar edilmesi ve her bir tekrarda 40+2$k$ adet örnek seçilmesinin en iyi sonucu verdiği raporlanmıştır (Silahtaroğlu 2020). $K$-Medoids ile karşılaştırıldığında, CLARA veri setinin nispeten küçük bir örneğini tarar. Seçilen örneğin referans noktaları, en iyi $K$-Medoids referans noktalarından uzaksa CLARA iyi bir kümeleme bulamaz. Bu nedenle optimal bir çözüm bulması pek olası değildir. Bununla birlikte, büyük veri kümeleri için makul bir hesaplama süresinde nispeten iyi kalitede çözümler ürettiği gösterilmiştir.

#### 3.2.2.4 CLARANS (Clustering Large Applications Based on Randomized Search) algoritması

CLARANS algoritması $K$-Medoids ve CLARA algoritmalarının geliştirilmiş halidir. Bu algoritma ile $n$ adet nesne referanslar yardımıyla $k$ adet kümeye ayrılırken şebeke diyagramından faydalanılır. Algoritmanın işleyişine ilişkin adımlar aşağıdaki gibidir.

Adım1: *Maxneighbor* ve *numlocal* olmak üzere iki adet parametresi vardır. (*Maxneighbor* parametresi incelenecek komşu sayısının üst limiti anlamına gelirken, *numlocal* parametresi ise yerel minimum nokta sayısının alt sınırı anlamına gelmektedir.)
Adım 2: Herhangi bir *k* nesnesi rastgele olarak veri tabanından seçilir.
Adım 3: Bu *k* nesnesi $S_i$ ve seçilmemiş $S_i$ olarak işaretlenerek bölünür.
Adım 4: Seçilen $S_i$'ler için T maliyeti hesaplanır.
Adım 5: Pozitif ve negatif olmak üzere iki ayrı T maliyeti hesaplanır. Eğer T negatifse merkez nokta güncellenir. Aksi takdirde, merkez nokta yerel optimum olarak seçilir.
Adım 6: En iyi sonuç alınana kadar devam edilir.

CLARANS algoritmasının CLARA algoritması ile benzerliği bütün veri tabanını yani bütün komşuları taramamasıdır. Bu iki algoritma arasındaki fark ise CLARA algoritmasının örnek düğümler seçme işlemini tarama işleminin başında gerçekleştirmesi, CLARANS algoritmasının her adımda örnek komşular seçerek devam etmesidir (Silahtaroğlu 2020). Ayrıca aykırı değerlere karşı dayanıklı bir algoritmadır ve yapılan örnekleme dinamiktir.

### 3.2.3 Yoğunluğa dayalı yöntemler

Yoğunluğa dayalı kümeleme yöntemlerinin genel yapısı, komşu nesnelerin sayısı yani yoğunluk belli bir eşiği aştığında kümenin büyümeye devam etmesini sağlamaktır. Bir kümenin her nesnesi, belli bir aralıkta minimum sayıda komşu nesne içermeli ve bu minimum komşu nesne miktarı önceden bildirilmelidir. Yöntem, keyfi bir nesne ile çalışmayı başlatır ve bu nesnenin komşuluğu, önceden bildirilen minimum yoğunluğu

karşılarsa, bu komşular gruba dahil edilir. Bu yöntemin en önemli avantajı silindirik veya eliptik gibi rastgele şekillerdeki grupları bulabilmeleri ve oluşturulacak küme sayısı hakkında herhangi bir başlangıç bilgisi gerektirmemeleridir (Rezende vd. 2010). $K$-Ortalamalar gibi yöntemler nesneler arasındaki mesafe ölçüsünü baz alarak kümeleme işlemi uyguladığından genellikle üretilen kümeler daireseldir ve keyfi şekiller oluşturamazlar. Buna karşılık, yoğunluğa dayalı kümeleme algoritmaları, gürültüyü etkili bir şekilde ayırırken, rastgele boyut ve şekillerde kümeleri bulabilir. Böyle durumlarda kümeleme işlemi yoğunluğa dayalı olarak gerçekleştirilebilir. Başka bir ifadeyle, birlikte yoğunluk oluşturan noktalar ayrı kümeler olarak değerlendirilir (Silahtaroğlu 2020). Bu tür kümelenmeler daha fazla araştırma ve geliştirmeyi cezbetmektedir.

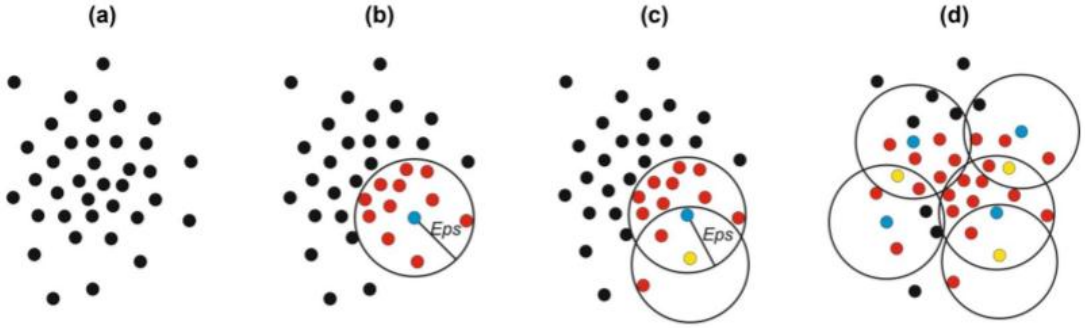
Yoğunluğa dayalı yöntemler arasında en bilinenler; DBSCAN (Density Based Spatial Clustering of Applications with Noise) algoritması, OPTICS (Ordering Points To Identify the Clustering Structure) algoritması ve DENCLUE (Densitybased Clustering) algoritmasıdır.

#### 3.2.3.1 DBSCAN (Density Based Spatial Clustering of Applications with Noise) algoritması

DBSCAN, büyük veri tabanlarının kümelenmesi amacıyla kullanılan yoğunluğa dayalı bir algoritmadır. Bir kümedeki iki nokta arasındaki maksimum uzaklık Eps, bir kümede olması gereken en az (minimum) nokta sayısı MinPts olarak tanımlanır. Eps ve MinPts değerleri algoritmanın giriş parametreleridir. Aşağıda verilen denklemlerdeki koşullar sağlandığı takdirde herhangi bir $p$ noktası başka bir $q$ noktası yardımıyla yoğunluğa erişebilir demektir (Silahtaroğlu 2020).

$$p \in N_{Eps}(q) \tag{3.17}$$

$$|N_{Eps}(q)| \geq \text{MinPts} \tag{3.18}$$




Şekil 3.12 DBSCAN algoritmasının temel bileşenlerinin grafiksel gösterimi ( (a) bir küme, (b) merkez (mavi nokta), (c) sınır (sarı nokta), (d) yoğunluğa erişebilir noktalar) (Entezami vd. 2020)

DBSCAN algoritmasının adımları aşağıdaki gibidir.

Adım 1: Bir kümeyi belirlemek için daha önce işlem yapılmamış rastgele bir *p* noktası seçilir.

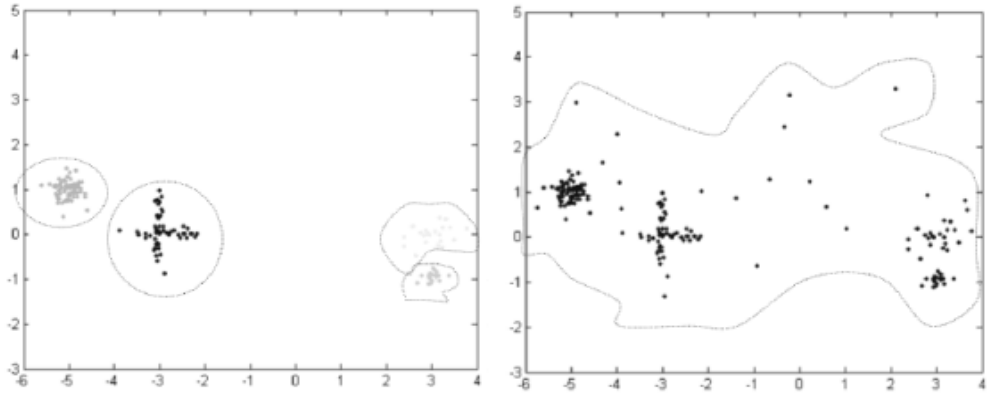
Adım 2: MinPts ve Eps koşulları sağlanacak şekilde *p* noktası üzerinden yoğunluğa erişebilecek tüm noktalar toplanır. (Eps komşuluğundaki komşuları tespit edilir.)

Adım 3: Tespit edilen komşu sayısı MinPts sayısına eşit veya fazla ise bu nokta ve komşuları bir küme oluşturur. Eğer komşu sayısı MinPts sayısından az ise bu nokta gürültü olarak belirlenir. Gürültü olarak belirlenen noktalar daha sonra başka kümelere dahil edilebilir.

Adım 4: Bir *p* noktası kümeye dahil edilirse *p* noktası üzerinden yoğunluğa erişebilecek noktaları yani komşuları da bu kümeye dahil edilir. Bu işlem ekleme yapılacak nokta kalmayana kadar devam eder.

Adım 5: Daha önce işlem yapılmamış başka bir nokta rastgele seçilir ve adımlar tekrarlanır.

Şekil 3.13'te farklı Eps ve MinPts değerleri ile kümelemedeki değişim görülmektedir. Birinci şekilde ideale yakın bir kümeleme gözlemlenirken, ikinci şekilde büyük değerde Eps seçilmesi ile tek bir kümenin ortaya çıktığı görülmektedir.

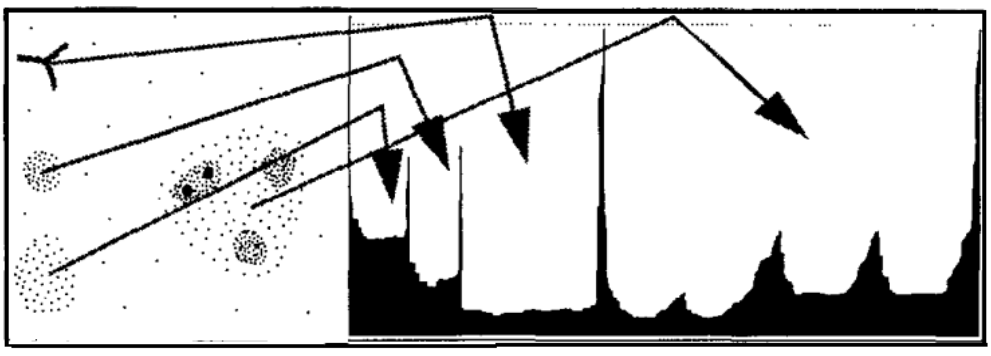


Eps=0.4 ve MinPts=4 Eps=3 ve MinPts=2

Şekil 3.13 Eps ve MinPts değerlerindeki değişimin küme oluşumuna etkisi (Bilgin vd. 2005)

### 3.2.3.2 OPTICS (Ordering Points To Identify the Clustering Structure) algoritması

OPTICS algoritması herhangi bir veri setindeki tüm nesnelerin sıralamasını hesaplar ve veri setindeki her nesne için çekirdek uzaklığı ve erişilebilir uzaklığı muhafaza eder. Çıktı sıralamasını oluşturmak için bir liste oluşturur ve ilgili en yakın çekirdek nesnelerin mesafesine göre sıralama yapar. Daha sonra algoritma, veri setinden rastgele bir nesne ile çalışmaya başlar. Eps'nin komşuluğunda çekirdek uzaklığı belirler ve erişilebilir uzaklığı tanımsız olarak ayarlar ve geçerli nesne, $p$, çıktıya yazılır (Han vd. 2012).



Şekil 3.14 Farklı boyut, şekil ve yoğunluklu kümeler için erişilebilirlik grafikleri (Ankerst vd. 1999)

OPTICS algoritması giriş parametresi olarak MinPts değerini alırken Eps değerini kendisi üretir ve bir kümeleme sıralaması geliştirir. Bu işlem için yararlanılan parametreler çekirdek uzaklık ve erişilebilir uzaklıktır. Veri tabanındaki tüm noktalar için bu parametreler hesaplanırken aynı zamanda veri tabanı da sıralamaya tabi tutulur. Bir $p$ nesnesinin çekirdek uzaklığı, $p$'nin Eps komşuluğunda en az MinPts nesnesine sahip olabileceği en küçük Eps değeridir. Yani Eps, $p$'yi çekirdek nesne yapan minimum mesafe eşiğidir. $p$, Eps ve MinPts'lere göre bir çekirdek nesne değilse, $p$'nin çekirdek uzaklığı tanımsızdır. $p$ nesnesine $q$'dan erişilebilir uzaklık, $p$ yoğunluğunu $q$'dan erişilebilir yapan minimum yarıçap değeridir. Erişilebilirlik için $p$, $q$'nun komşuluğunda olmalıdır. Bu nedenle $q$'dan $p$'ye erişilebilir uzaklık, $\max\{core-distance(q), dist(p,q)\}$ şeklindedir (Han vd. 2012). MinPts' ye göre bir çekirdek nesne değilse, $q$'dan $p$'ye erişilebilirlik mesafesi tanımsızdır.

DBSCAN algoritması Eps ve MinPts değerleri ile kümeleme işlemi gerçekleştirmektedir. Ancak bu değerler kümeleme işlemi yapılmadan belirleneceği için doğrudan bir müdahale söz konusu olmaktadır. Eps ve MinPts değerleri de bilindiği üzere küme sayısını ve niteliğini oldukça etkilemektedir. DBSCAN algoritmasında karşılaşılan bu sorunun çözümü için OPTICS algoritması geliştirilmiştir. Algoritma daha çok veri görselleştirme olarak nitelendirilebilir.

#### 3.2.3.3 DENCLUE (Densitybased Clustering) algoritması

Büyük veri tabanlarına çeşitli kümeleme algoritmaları uygulanabilir. Bu veri tabanları yüksek boyut özelliğine sahip vektörlerin kümelenmesini gerektirdiğinden genellikle büyük miktarda gürültü içerirler. Bununla birlikte, etkinlik ve verimlilik açısından mevcut algoritmaların sayısı da biraz sınırlıdır.

DENCLUE algoritmasının avantajları aşağıdaki gibidir.

- Kümeleri belirlemek için yoğunluk dağılım fonksiyonlarından faydalandığından sağlam matematiksel temellere dayanır.
- Gürültülü büyük veri kümelerinde etkin kümeleme özelliğine sahiptir.

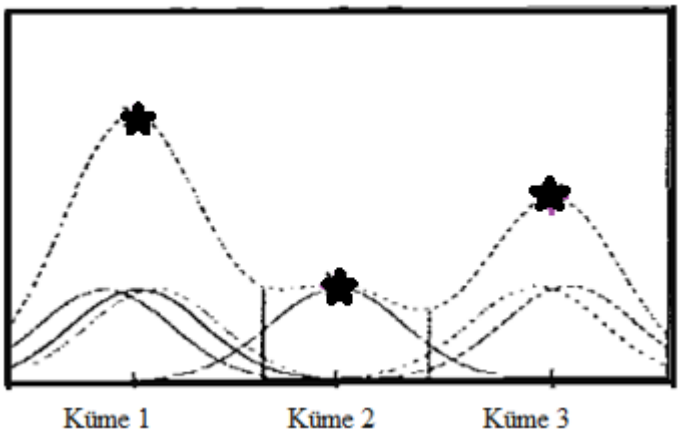
- Yüksek boyutlu veri setlerinde rastgele şekillendirilmiş kümelerin kompakt matematiksel tanımına izin verir.
- Mevcut algoritmalardan önemli ölçüde daha hızlıdır. DBSCAN algoritmasından 45 kata kadar hızlı olduğunu test eden deneysel çalışmalar yapılmıştır. (Hinneburg ve Keim 1998)

DENCLUE algoritmasında, genellikle kare dalga etki fonksiyonu (Square Wave Influence Function) (3.19) veya Gauss etki fonsiyonu (3.20) kullanılır.

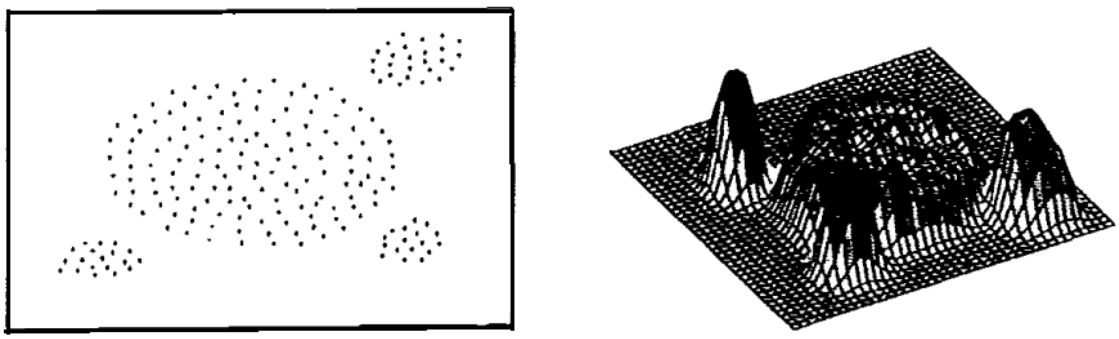
$$f_{kare}(x,y) = \begin{cases} 0, & eğer\ d(x,y) > \sigma \\ 1, & diğer\ durumlarda \end{cases} \quad (3.19)$$

$$f_{Gauss}(x,y) = e^{-\frac{d(x,y)^2}{2\sigma^2}} \quad (3.20)$$

$\sigma$ :standart sapma




Şekil 3.15 Genel yoğunluk fonksiyonu (Hinneburg ve Keim 1998)



Şekil 3.16 İki boyutlu uzayda Gauss yoğunluk fonksiyonu (Hinneburg ve Keim, 1998)

Şekil 3.15'te genel yoğunluk fonksiyonu verilmiş olup, yerel minimum noktaları küme merkezlerini göstermektedir.

DENCLUE algoritması çok boyutlu veri kümelerinde gürültüye rağmen başarılı sonuçlar vermektedir. Bunun yanı sıra matematiksel temellere dayalı bir algoritma olması sebebiyle modelleme konusunda oldukça iyi sonuçlar verir (Zhu vd. 2022).

### 3.2.4 Grid temelli (Izgara Tabanlı) yöntemler

Grid temelli veya ızgara tabanlı olarak adlandırılan yöntemler, kümeleme işlemlerini ızgara yapısı üzerinde yani nicel hale getirilmiş sonlu uzayda gerçekleştirir. Bu yaklaşımın en önemli avantajı olarak veri nesnelerinin boyutundan bağımsız gerçekleştirilen hızlı işlem süresine sahip olması söylenebilir. Kümeleme işlemleri veri nesnelerinin boyutuna bağlı olduğundan bu yöntem ile güçlü ve etkili sonuçlar üretilmektedir. Izgaraları kullanmak, kümelemenin haricinde birçok veri madenciliği sorunu için de etkili bir yöntemdir. Bu nedenle ızgara tabanlı yöntemler, yoğunluğa dayalı yöntemler ve hiyerarşik yöntemler gibi diğer kümeleme yöntemleriyle bütünleşik olarak kullanılabilir (Han vd. 2012).

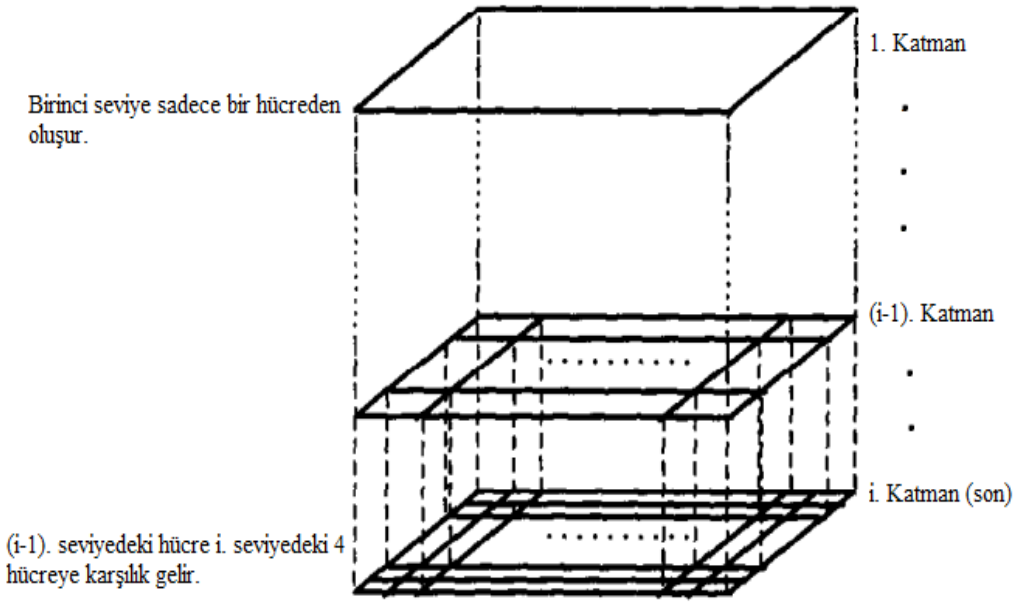
Izgara tabanlı kümeleme algoritmasının adımları aşağıdaki gibidir.

Adım 1: Sonlu uzayda ızgara yapısının meydana getirilmesi yani veriye ilişkin alanın sınırlı sayıda hücreye ayrılması,
Adım 2: Her hücre için yoğunluğun hesaplanması ve bu yoğunluklara uygun olarak hücrelerin sıralanması,
Adım 3: Kümeye ilişkin merkezlerin belirlenmesi,
Adım 4: Komşu hücrelere geçilmesi (Yousefi vd. 2020).

En sık kullanılan ızgara tabanlı yöntemler; STING (Statistical Information Grid Approach) algoritması, WaveCluster algoritması ve CLIQUE algoritmasıdır.

### 3.2.4.1 STING (Statistical Information Grid Approach) algoritması

STING, nesnelerin uzamsal alanının dikdörtgen hücrelere bölündüğü ızgara tabanlı çok çözünürlüklü bir algoritmadır. Bu tür dikdörtgen hücrelerin birkaç seviyesi, farklı çözünürlük seviyelerine karşılık gelir ve verileri bölümlere ayırmak için hiyerarşik çok çözünürlüklü bir yapı oluşturur. Bu veri yapısının faydası birçok yaygın bölge odaklı sorguyu, bir dizi noktada verimli bir şekilde işlemesidir. Bir ızgara hücresine bir kaydın yerleştirilmesi tamamen fiziksel konumu tarafından belirlenir. Yüksek seviyedeki hücrelerin her biri kendinden sonraki düşük seviyede yer alan hücreleri oluşturmak için bölümlere ayrılır (Şekil 3.17).




Şekil 3.17 Hiyerarşik yapı (Wang vd. 1997)

STING algoritmasının adımları aşağıdaki gibidir.

Adım 1: Başlangıç için bir katman belirlenir.
Adım 2: Belirlenen katmandaki her bir hücre için, bu hücrelerin gerekirse ilgili sorguya yanıt vermesi için gereken istatistiksel bilgileri (güven aralığı, tahmin,..) hesaplanır.
Adım 3: Hesaplamalara göre hücreler ilgili veya ilgisiz olarak etiketlenir. İlgisiz hücreler değerlendirmeden çıkarılır.

Adım 4: En alttaki katmana ulaşılıncaya kadar işlem devam eder.
Adım 5: Sorguya yanıt verilirse, sorguyu karşılayan ilgili hücrelerin bölgeleri döndürülür. Aksi halde ilgili hücrelere düşen veriler alınır ve sorgunun gereksinimleri karşılanana kadar işlem devam eder (Dat vd. 2017).

STING algoritmasında düşük seviyelere doğru (ayrıntı düzeyi '0'a yaklaştığında) DBSCAN algoritmasının sonuçlarına yaklaşıldığı gözlemlenir. Başka bir deyişle, hesap ve hücre boyutu bilgileri kullanıldığında, yoğun kümeler yaklaşık olarak STING algoritması kullanılarak tanımlanabilir. Bu nedenle STING algoritmasından bazı durumlarda yoğunluğa dayalı yöntem olarak bahsedilebilir (Han vd. 2012).

Algoritmada her hücrede depolanan istatistiksel bilgiler, ızgara hücresindeki verilerin özet bilgilerini temsil eder.

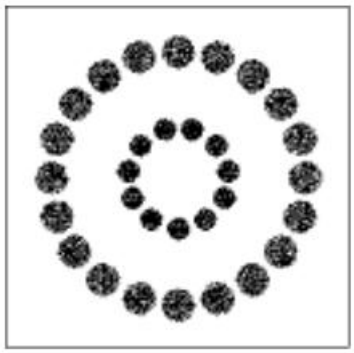
#### 3.2.4.2 WaveCluster algoritması

1998 yılında Sheikholeslami ve arkadaşları tarafından önerilen bu kümeleme tekniği veriler üzerinde tek tip iki boyutlu bir ızgara tanımlar ve her hücredeki verileri nokta sayısıyla temsil eder. Böylece her bir veri, bir görüntü olarak ele alınan bir dizi gri tonlamalı nokta haline gelir. Daha sonra küme arama problemi, dalgacıkların çoklu ölçekleme ve gürültü azaltma özelliklerinden yararlanmak için kullanıldığı bir görüntü ayırma problemine dönüştürülür (Murtagh vd. 2012).

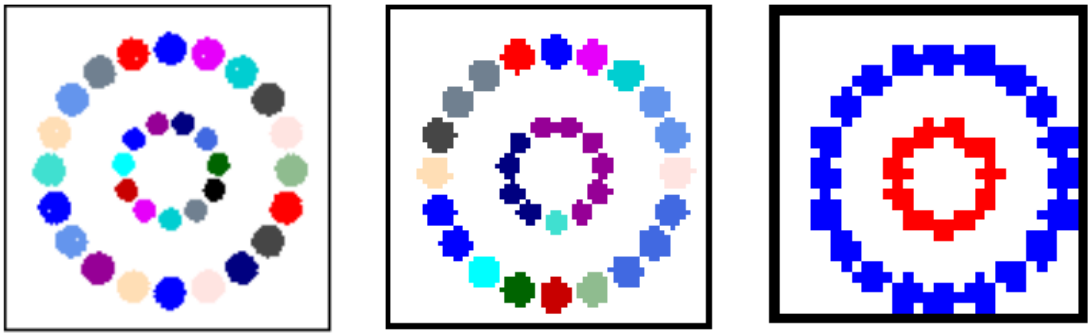
WaveCluster algoritmasının adımları aşağıdaki gibidir.

Adım 1: Bir veri ızgarası oluşturulur ve her bir veri, ızgaradaki bir hücreye atanır.
Adım 2: Verilere dalgacık dönüşümü uygulanır. Dalgacık dönüşümü, işaretlerin zaman frekans dönüşümünü elde etmeye yarayan bir işaret işleme tekniğidir (Silahtaroğlu 2020).
Adım 3: Bağlantılı kümeleri bulmak için ortalama alt görüntü kullanılır.
Adım 4: Elde edilen kümeler, orijinal uzaydaki noktalara geri eşlenir.

Bu yöntem ile kümeler belirlenirken orijinal uzaydaki yoğun bölgeler, yakın noktaları çekerken uzaktaki noktaları engeller. Böylece veri seti içindeki kümeler ortaya belirgin bir şekilde çıkar.



Şekil 3.18 Veri setinin asıl hali (Sheikholeslami vd., 1998)




Şekil 3.19 Veri setine uygulanmış üç farklı ölçekte dalga dönüşümü (Sheikholeslami vd. 1998)

WaveCluster, kullanıcının ihtiyacına göre farklı taneciklerde kümeleme yapmak için kullanılabilen önemli bir özelliğe sahiptir. Şekil 3.19, şekil 3.18'de gösterilen veri setine ait dalgacık dönüşümünün sonuçlarını göstermektedir. Burada, WaveCluster algoritmasının kümeleri farklı ölçeklerde nasıl bulduğu gösterilmektedir. Ayrıca WaveCluster algoritmasının bu özelliği, kullanıcıya ilk sonuçlara göre sorguları değiştirme esnekliği sağlamaktadır (Sheikholeslami vd. 1998).

### 3.2.4.3 CLIQUE (Clustering in Quest) algoritması

CLIQUE algoritması alt uzaylarda yoğunluğa dayalı kümeleri bulmaya yarayan ızgara tabanlı bir yöntemdir. Boyutları ızgara yapısı gibi bölerken hücre içerisindeki nokta sayısına göre yoğunluğunu belirler. Yüksek boyutlu veri kümelerinde doğru sonuçları bulduğu görülür (Agrawal vd. 1998)

CLIQUE algoritmasının adımları aşağıdaki gibidir.

Adım 1: Kümeleri içeren alt uzaylar belirlenir.
Adım 2: Kümeler tanımlanır.
Adım 3: Küme için minimum tanımlamalar üretilir.

# 4. ÜLKELERİN COVID-19 VERİLERİNE GÖRE KÜMELENMESİ

2019 yılı sonunda Çin'de başlayan ve 2020 yılının başlarından itibaren dünyayı etki altına alan ve Dünya Sağlık Örgütü tarafından 20 Mart 2020'de Pandemi olarak ilan edilen Covid-19 hastalığı ile ilgili bilimsel birçok araştırma ve analiz yapılmıştır. Bu çalışmada dünyayı etkisi altına alan Covid-19 verileri ile bir kümeleme analizi gerçekleştirilmiştir. Kullanılan veri seti, https://github.com/owid/covid-19-data/tree/master/public/data sitesi üzerinden 28 Mayıs 2022 tarihinde elde edilmiştir. Bu analiz çalışmasında 52 ülke ve ülkelere ilişkin 5 farklı değişken kullanılmıştır (Çizelge 4.1-4.2). Veri seti, 1.06.2021-28.05.2022 arasındaki bir yıllık dönemi kapsamaktadır. Çalışmada veri seti Türkiye'deki mevsimler göz önüne alınarak 1.06.2021-31.08.2021 (yaz), 1.09.2021-30.11.2021 (sonbahar), 1.12.2021-28.02.2022 (kış) ve 1.03.2022-27.05.2022 (ilkbahar) biçiminde 4 gruba ayrılarak incelenmiştir. Böylece mevsimlere göre kümelerde oluşacak farklılıklar da gözlemlenebilecektir. Analiz Python programı kullanılarak gerçekleştirilmiştir. Covid-19 ile ilgili sağlıklı verilere ulaşmanın zorluğundan dolayı 52 ülke, veri kaybının en az olduğu ülkeler olarak seçilip çalışmaya dahil edilmiştir. Ülkeler konum ve gelişmişlik düzeyi olarak heterojen bir yapıya sahiptir. Eksik veriler ortalama değer ile doldurularak veri, analize hazır hale getirilmiştir. Kümeleme analizi öncesi değişkenler arası ölçek farklılıklarını ortadan kaldırmak için standartlaştırma işlemi yapılmıştır. Küme sayısının belirlenmesinde Elbow yönteminden faydalanılmıştır. Kümeler, Ward yöntemi ve $K$-Ortalamalar yöntemi kullanılarak elde edilmiş ve ortaya çıkan sonuçlar karşılaştırılmıştır.

Çizelge 4.1 Kümeleme analizinde yer verilen ülkelerin listesi

| | | | |
|---|---|---|---|
| Amerika | Estonya | Kanada | Panama |
| Arjantin | Etiyopya | Kolombiya | Polonya |
| Avustralya | Finlandiya | Küba | Portekiz |
| Belçika | Fransa | Letonya | Slovenya |
| Birleşik Arap Emirlikleri | Guatemala | Litvanya | Sri Lanka |
| Bolivya | Güney Kore | Lüksemburg | Suudi Arabistan |
| Brezilya | Hindistan | Macaristan | Şili |

Çizelge 4.1 Kümeleme analizinde yer verilen ülkelerin listesi (devam)

| | | | |
|---|---|---|---|
| Bulgaristan | İngiltere | Maldivler | Tayland |
| Çekya | İrlanda | Malezya | Türkiye |
| Çin | İsrail | Malta | Ukrayna |
| Danimarka | İsviçre | Meksika | Uruguay |
| Ekvator | İtalya | Moldova | Yunanistan |
| Endonezya | Japonya | Norveç | Zimbabve |

Çizelge 4.2 Kümeleme analizinde yer verilen değişkenlerin listesi

| | |
|---|---|
| **Vaka Oranı** | Her ülke için belirlenen tarih aralığındaki toplam vaka sayısının ülke nüfusuna oranıdır. Her yüz kişide gerçekleşen değerler baz alınmıştır. |
| **Ölüm Oranı** | Her ülke için belirlenen tarih aralığındaki toplam ölüm sayısının ülke nüfusuna oranıdır. Her yüz kişide gerçekleşen değerler baz alınmıştır. |
| **Test Oranı** | Her ülke için belirlenen tarih aralığındaki toplam test sayısının ülke nüfusuna oranıdır. Her yüz kişide gerçekleşen değerler baz alınmıştır. |
| **Pozitiflik Oranı** | Her ülke için belirlenen tarih aralığındaki toplam vaka sayısının toplam test sayısına oranıdır. |
| **Aşı Oranı** | Her ülke için belirlenen tarih aralığında uygulanan toplam Covid-19 aşı dozlarının ülke nüfusuna oranıdır. Her yüz kişide gerçekleşen değerler baz alınmıştır. |

Yaz dönemi olarak 1.06.2021-31.08.2021 tarihleri arası göz önüne alınmıştır. Bu tarih aralığında ilk olarak hiyerarşik yöntemlerden Ward yöntemi uygulanmıştır. Bilindiği üzere, hiyerarşik yöntemlerde küme sayısını önceden belirlemeye gerek yoktur. Bu çalışmada hiyerarşik ve hiyerarşik olmayan iki yöntem karşılaştırılmak istenmiştir. Bu nedenle, daha önceki bölümlerde ifade edilen küme sayısını belirlemek için kullanılan yöntemlerden en çok tercih edilen Elbow yöntemi kullanılarak küme sayısı belirlenmiş ve doğru karşılaştırma yapabilmek için küme sayısı her iki yöntemde de aynı olarak alınmıştır. Elbow yöntemi ile yaz dönemi için küme sayısı $k$=4 olarak belirlenmiştir (Şekil 4.1).


Şekil 4.1 Elbow yöntemine göre elde edilen yaz dönemine ait grafik

Çizelge 4.3 ile Ward yöntemi kullanılarak elde edilen 4 kümede yer alan ülkeler verilmiştir.

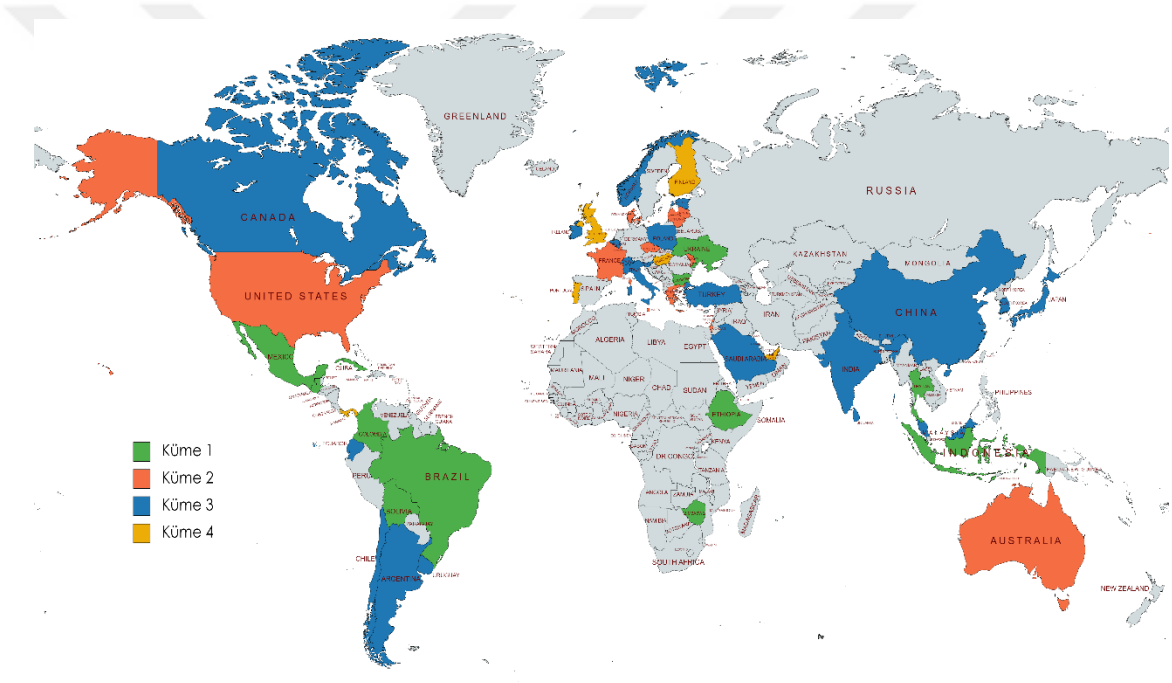
Çizelge 4.3 Ward yöntemine göre elde edilen yaz dönemi kümeleri

| **Küme 1** | **Küme 2** | **Küme 3** | **Küme 4** |
|---|---|---|---|
| Bolivya | Amerika | Arjantin | Birleşik Arap Emirlikleri |
| Brezilya | Avustralya | Belçika | Finlandiya |
| Bulgaristan | Çekya | Çin | İngiltere |
| Endonezya | Danimarka | Ekvador | Macaristan |
| Etiyopya | Fransa | Estonya | Maldivler |
| Guatemala | İsrail | Güney Kore | Panama |
| Kolombiya | Letonya | Hindistan | Portekiz |
| Küba | Litvanya | İrlanda | |
| Meksika | Lüksemburg | İsviçre | |
| Tayland | Malta | İtalya | |
| Ukrayna | Moldova | Japonya | |
| Zimbabve | Yunanistan | Kanada | |
| | | Malezya | |
| | | Norveç | |
| | | Polonya | |
| | | Slovenya | |

Çizelge 4.3 Ward yöntemine göre elde edilen yaz dönemi kümeleri (devam)

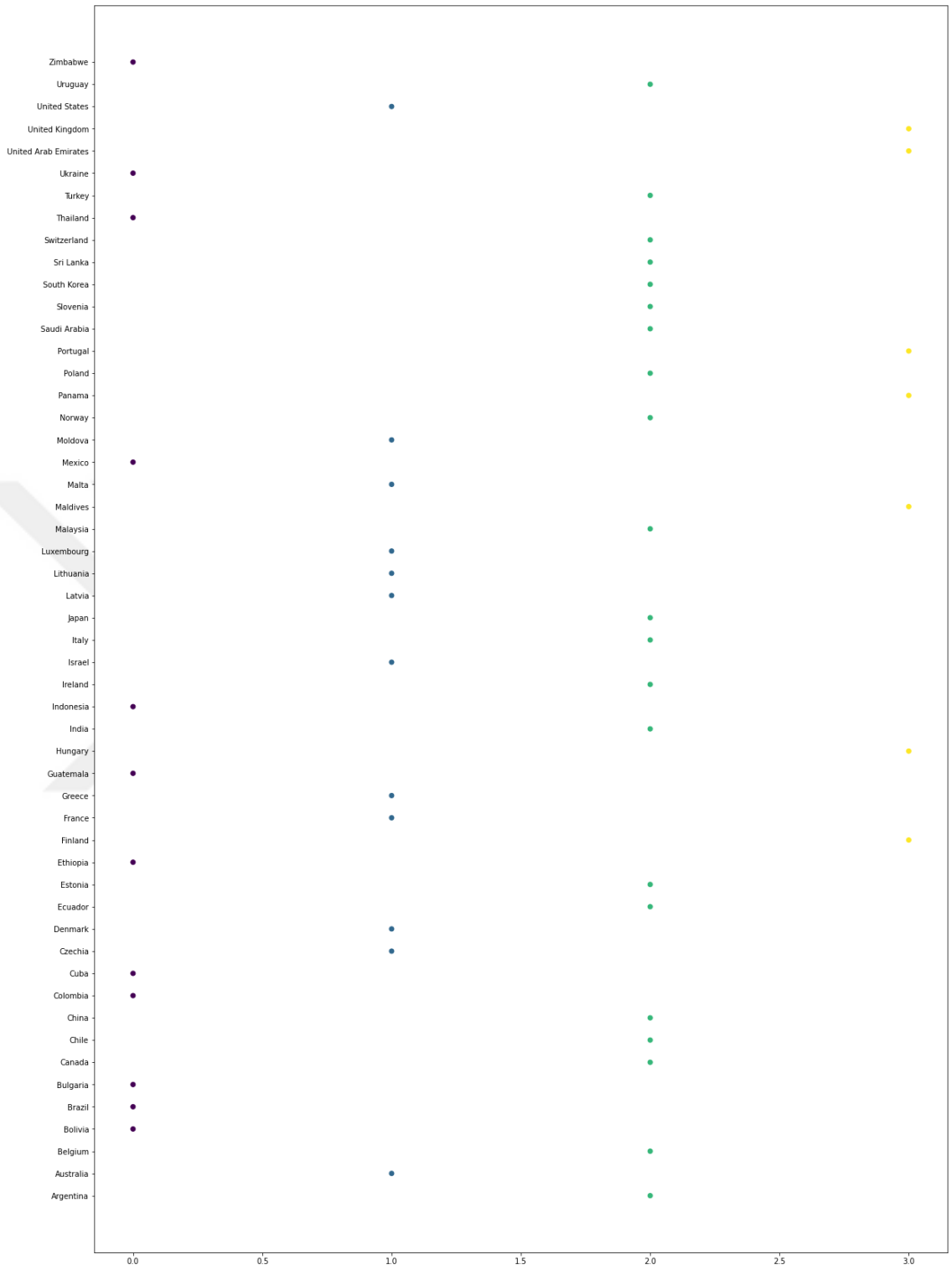
| | | | |
|---|---|---|---|
| | | Sri Lanka | |
| | | Suudi Arabistan | |
| | | Şili | |
| | | Türkiye | |
| | | Uruguay | |

Bu kümelendirmeye göre, yaz dönemi için ülkemiz İtalya, İsviçre, Belçika gibi Avrupa ülkeleri yanında Hindistan, Suudi Arabistan gibi ülkelerle aynı kümede yer almıştır. Elde edilen bu kümeler farklı renkler ile renklendirilerek her kümeye düşen ülkeler dünya haritası üzerinde görsel olarak aşağıdaki gibi verilmiştir.



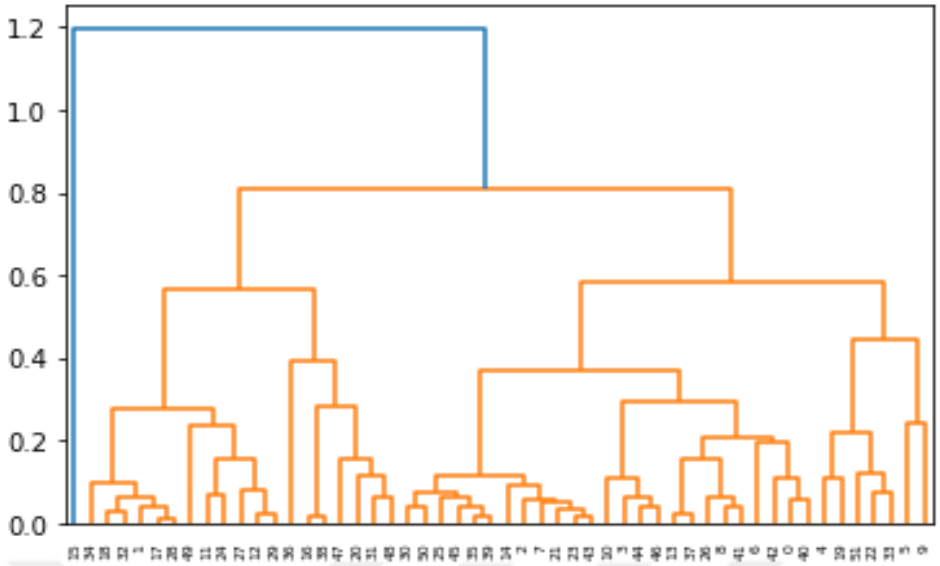

Şekil 4.2 Ward yöntemine göre yaz dönemi kümelerinin dünya haritasında gösterimi

Şekil 4.3 ile ülkelerin hangi kümeye düştüğünü gösteren Python çıktısı verilmiştir. Şekil 4.3'e bakıldığında, Python program çıktısında küme indisinin 0'dan başladığına dikkat edilmelidir. Çıktıya göre, analizin sonucunda 0'dan 3'e kadar 4 küme yer almaktadır. Yorumlamaların kolay olması için kümeler yazılırken 0-3 indisleri yerine indisler, 1'den 4'e kadar ifade edilmiştir.




Şekil 4.3 Ward yöntemine göre yaz dönemi kümelerinin Python programı çıktısı

Şekil 4.4 ile Ward yöntemine ilişkin dendogram gözükmektedir.




Şekil 4.4 Ward yöntemine göre yaz dönemine ait dendogram

Yaz dönemine ait ülkelerin kümelenmesi için bir de hiyerarşik olmayan yöntemlerden *K*-Ortalamalar yöntemi kullanılmıştır. Buna göre, yaz dönemine ait gözlemler kullanılarak *K*-Ortalamalar yöntemi ile elden edilen kümeler çizelge 4.4 ile verilmiştir.

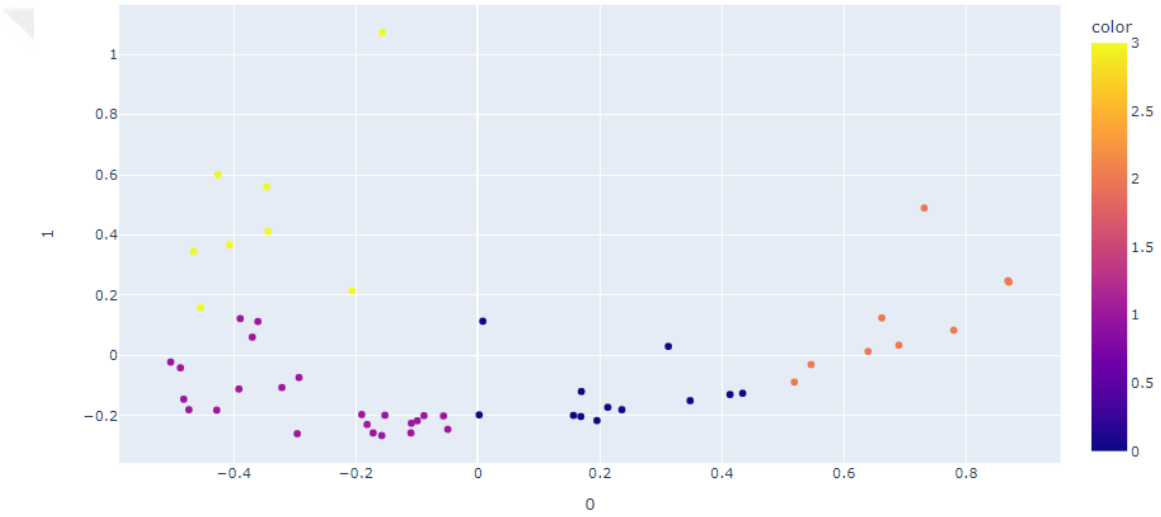
Çizelge 4.4 *K*-Ortalamalar yöntemine göre elde edilen yaz dönemi kümeleri

| **Küme 1** | **Küme 2** | **Küme 3** | **Küme 4** |
|---|---|---|---|
| Amerika | Arjantin | Birleşik Arap Emirlikleri | Brezilya |
| Avustralya | Belçika | Çekya | Endonezya |
| Bulgaristan | Bolivya | Finlandiya | Etiyopya |
| Danimarka | Çin | İngiltere | Guatemala |
| Estonya | Ekvador | İsrail | Kolombiya |
| Fransa | Güney Kore | Macaristan | Küba |
| Letonya | Hindistan | Maldivler | Meksika |
| Litvanya | İrlanda | Panama | Zimbabve |
| Lüksemburg | İsviçre | Portekiz | |
| Malta | İtalya | | |
| Moldova | Japonya | | |
| Yunanistan | Kanada | | |
| | Malezya | | |
| | Norveç | | |
| | Polonya | | |
| | Slovenya | | |
| | Sri Lanka | | |
| | Suudi Arabistan | | |
| | Şili | | |

Çizelge 4.4 *K*-Ortalamalar yöntemine göre elde edilen yaz dönemi kümeleri (devam)

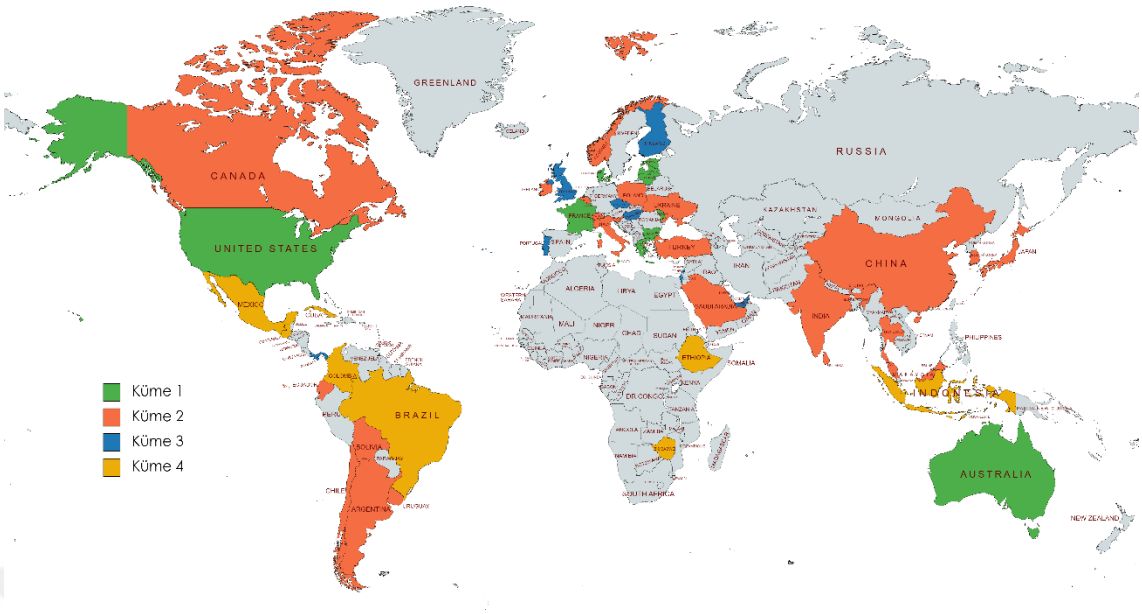
| | | | |
|---|---|---|---|
| | Tayland | | |
| | Türkiye | | |
| | Ukrayna | | |
| | Uruguay | | |

Bu kümelendirmeye göre yine ülkemiz İtalya, İsviçre, Belçika gibi Avrupa ülkeleri yanında Hindistan, Suudi Arabistan gibi ülkelerin bulunduğu ikinci kümede yer almıştır. Şekil 4.5'te *K*-Ortalamalar yöntemi kullanılarak yapılan kümeleme analizi sonucu ülkelerin hangi kümeye düştüğünü gösteren Python çıktısı yer almaktadır.



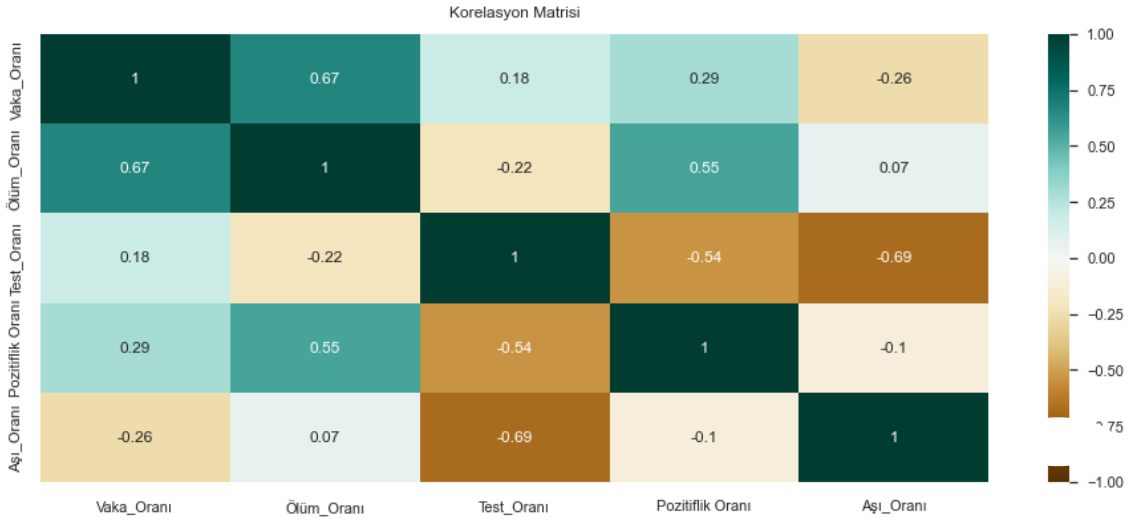

Şekil 4.5 *K*-Ortalamalar yöntemine göre yaz dönemi kümelerinin Python programı çıktısı

Yine Ward yöntemine benzer şekilde, *K*-Ortalamalar yöntemine göre elde edilen bu kümeler farklı renkler ile renklendirilerek her kümeye düşen ülkeler dünya haritası üzerinde görsel olarak aşağıdaki gibi verilebilir.




Şekil 4.6 *K*-Ortalamalar yöntemine göre yaz dönemi kümelerinin dünya haritasında gösterimi

Şekil 4.7 ile yaz dönemine ait korelasyon matrisi verilmiştir.



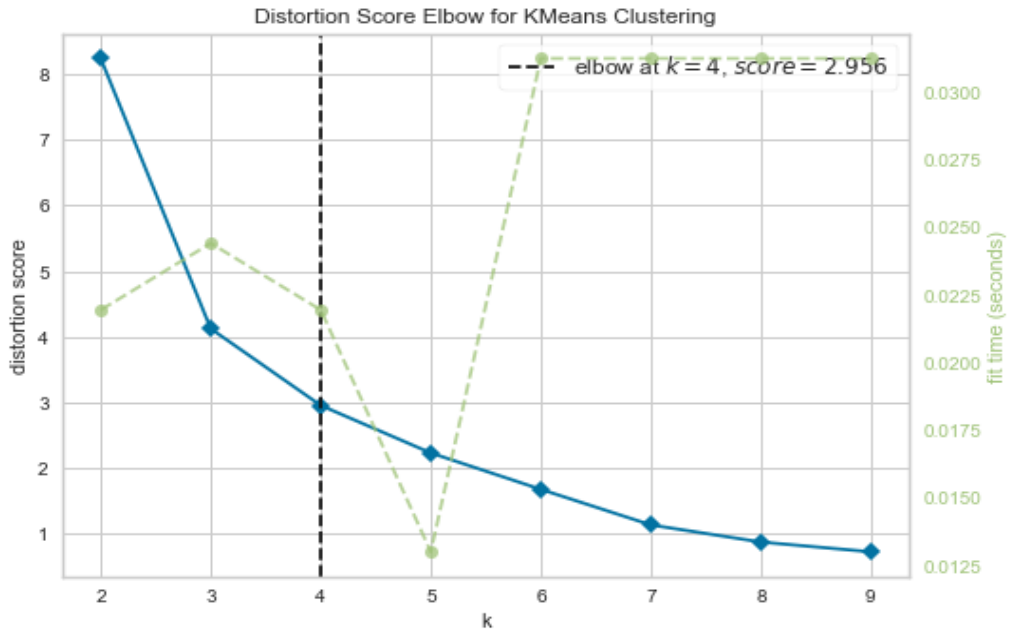

Şekil 4.7 Yaz dönemi için korelasyon matrisi

Şekil 4.7' de yer alan korelasyon matrisine göre yaz döneminde aşı oranı ile test oranı arasındaki korelasyon katsayısı -0.69 olarak elde edilmiştir. Buna göre, iki değişken arasında ters yönlü bir ilişki olduğu, yani aşı yaptıranların oranı arttıkça test

yaptıranların oranının azaldığı söylenebilir. Benzer şekilde vaka oranı ile ölüm oranı arasında pozitif bir ilişki olduğu görülmektedir.

Yaz döneminde; Ward yöntemine göre Türkiye, Arjantin, Belçika, Çin, Ekvador, Estonya, Güney Kore, Hindistan, İrlanda, İsviçre, İtalya, Japonya, Kanada, Malezya, Norveç, Polonya, Slovenya, Sri Lanka, Suudi Arabistan, Şili ve Uruguay ile aynı kümede yer almaktadır. *K*-Ortalamalar yöntemine göre ise, Türkiye, Arjantin, Belçika, Bolivya, Çin, Ekvador, Güney Kore, Hindistan, İrlanda, İsviçre, İtalya, Japonya, Kanada, Malezya, Norveç, Polonya, Slovenya, Sri Lanka, Suudi Arabistan, Şili, Tayland, Ukrayna ve Uruguay ile aynı kümede yer almaktadır. İki yöntem ile elde edilen kümelerin geneline bakıldığında Bulgaristan, Estonya, Bolivya, Tayland, Ukrayna, Çekya ve İsrail farklı kümelerde yer almaktadır.

Sonbahar dönemi olarak 1.09.2021-30.11.2021 tarihleri arası göz önüne alınmıştır. Bu tarih aralığında Ward yöntemi kullanılarak elde edilen sonuçlara göre ülkelere ait kümeler çizelge 4.5 ile verilmiştir. Bir önceki dönem için elde edildiği gibi küme sayısı Elbow yöntemi ile *k*=4 olarak belirlenmiştir (Şekil 4.8).

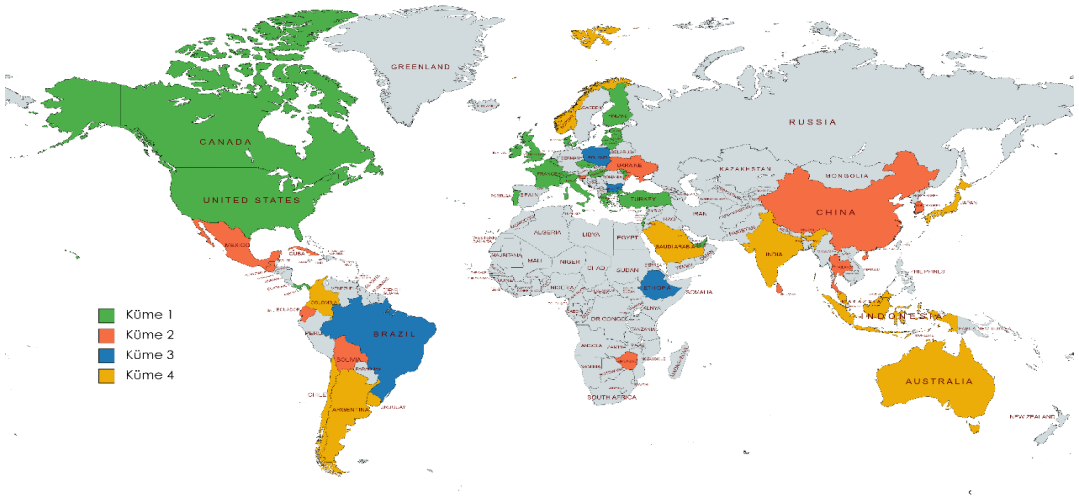



Şekil 4.8 Elbow yöntemine göre elde edilen sonbahar dönemine ait grafik

Çizelge 4.5 Ward yöntemine göre elde edilen sonbahar dönemi kümeleri

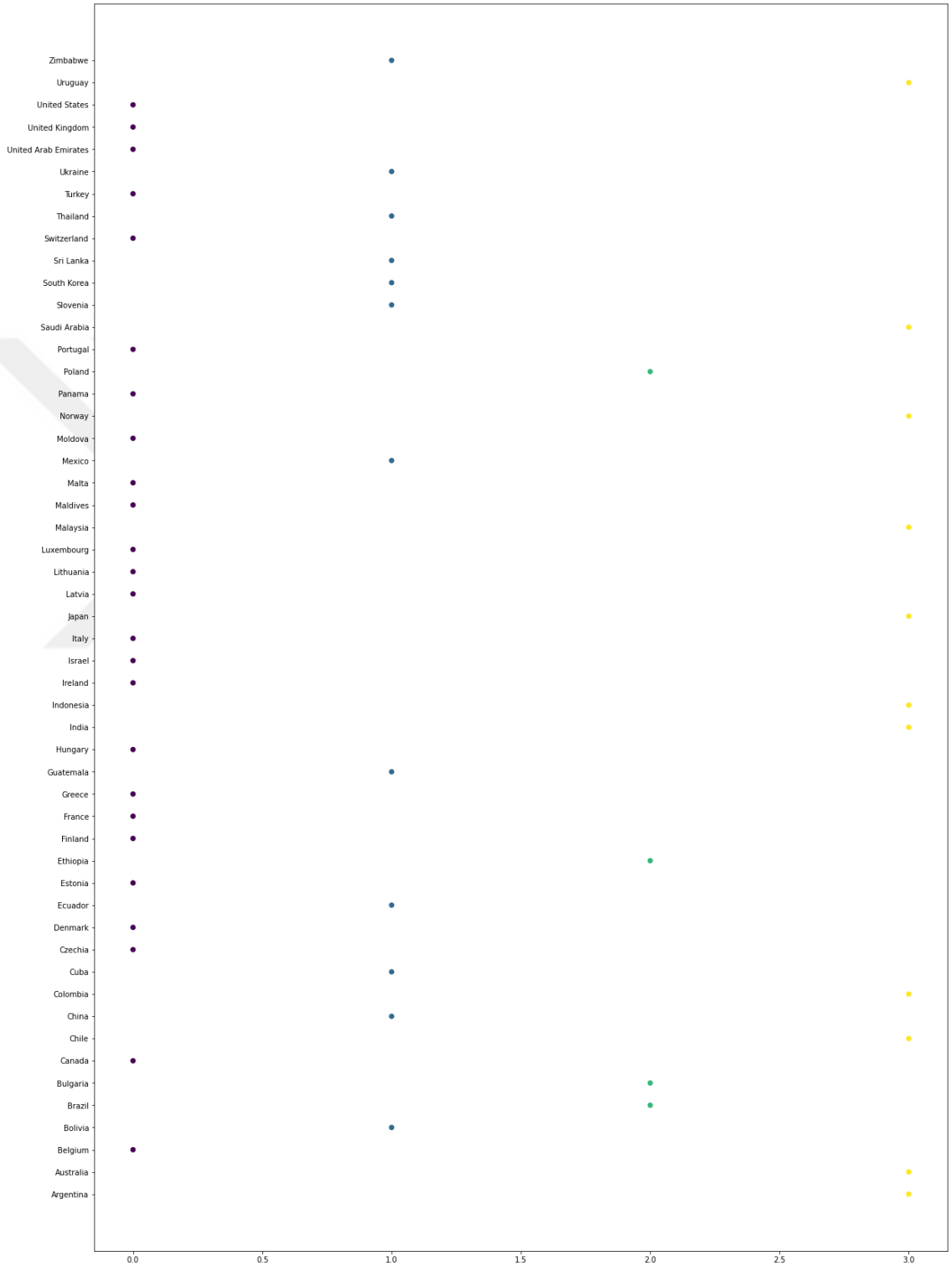
| **Küme 1** | **Küme 2** | **Küme 3** | **Küme 4** |
|---|---|---|---|
| Amerika | Bolivya | Brezilya | Arjantin |
| Belçika | Çin | Bulgaristan | Avustralya |
| Birleşik Arap Emirlikleri | Ekvador | Etiyopya | Endonezya |
| Çekya | Guatemala | Polonya | Hindistan |
| Danimarka | Güney Kore | | Japonya |
| Estonya | Küba | | Kolombiya |
| Finlandiya | Meksika | | Malezya |
| Fransa | Slovenya | | Norveç |
| İngiltere | Sri Lanka | | Suudi Arabistan |
| İrlanda | Tayland | | Şili |
| İsrail | Ukrayna | | Uruguay |
| İsviçre | Zimbabve | | |
| İtalya | | | |
| Kanada | | | |
| Letonya | | | |
| Litvanya | | | |
| Lüksemburg | | | |
| Macaristan | | | |
| Maldivler | | | |
| Malta | | | |
| Moldova | | | |
| Panama | | | |
| Portekiz | | | |
| Türkiye | | | |
| Yunanistan | | | |

Kümeler renklendirilerek dünya haritası üzerindeki gösterimi Şekil 4.9 ile verilmiştir.



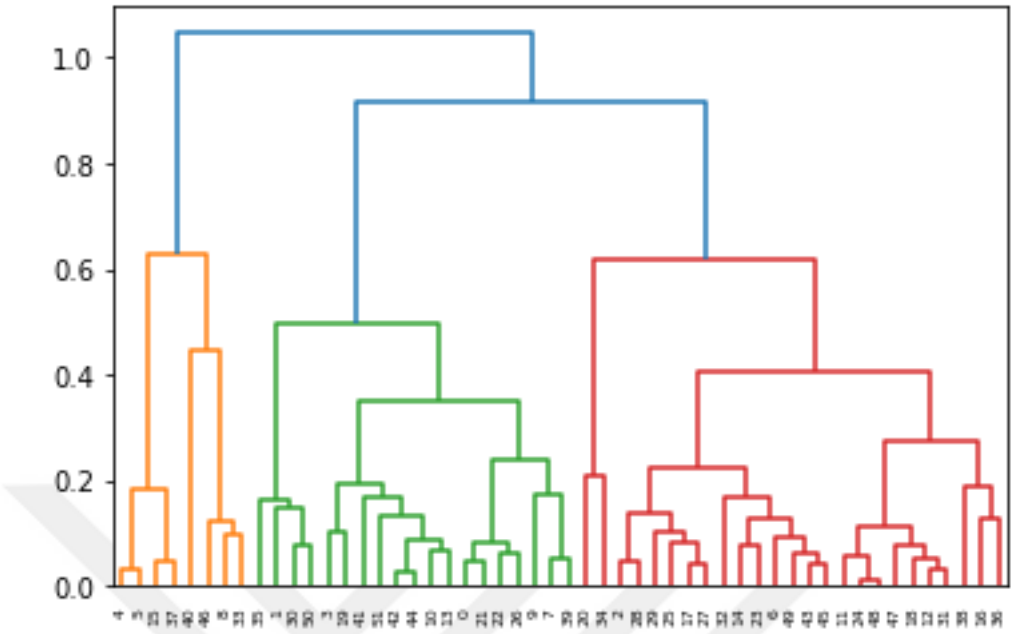

Şekil 4.9 Ward yöntemine göre sonbahar dönemi kümelerinin dünya haritasında gösterimi

Ward yöntemine göre elde edilen sonbahar dönemi kümelerine ait Python programı çıktısı Şekil 4.10 ile verilmiştir.



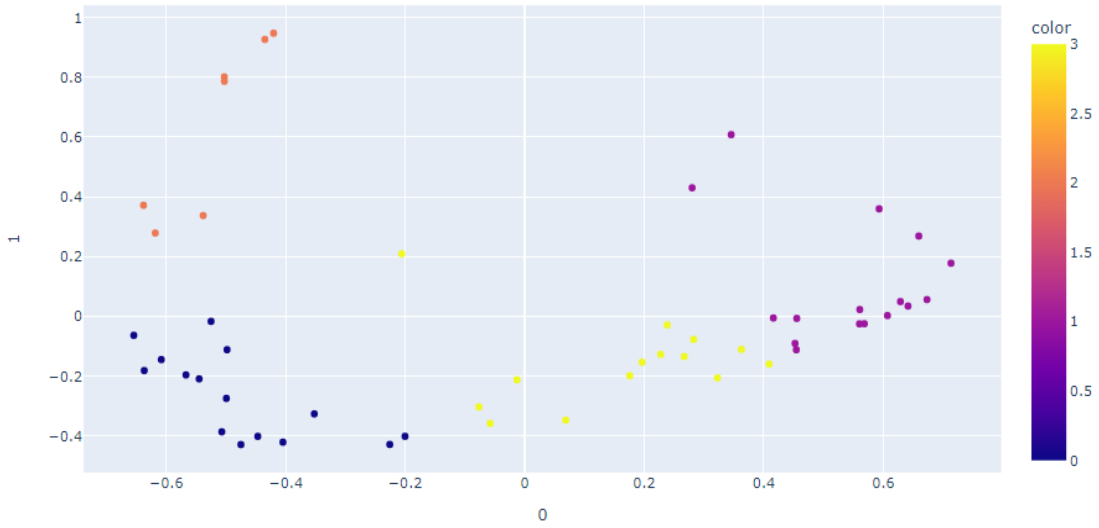

Şekil 4.10 Ward yöntemine göre sonbahar dönemi kümelerinin Python programı çıktısı

Şekil 4.11 ile Ward yöntemine ilişkin dendogram gözükmektedir.




Şekil 4.11 Ward yöntemine göre sonbahar dönemine ait dendogram

Şekil 4.12'de sonbahar dönemine ait $K-$Ortalamalar yöntemine göre elde edilen kümelerin gösterildiği Python program çıktısına yer verilmiştir.



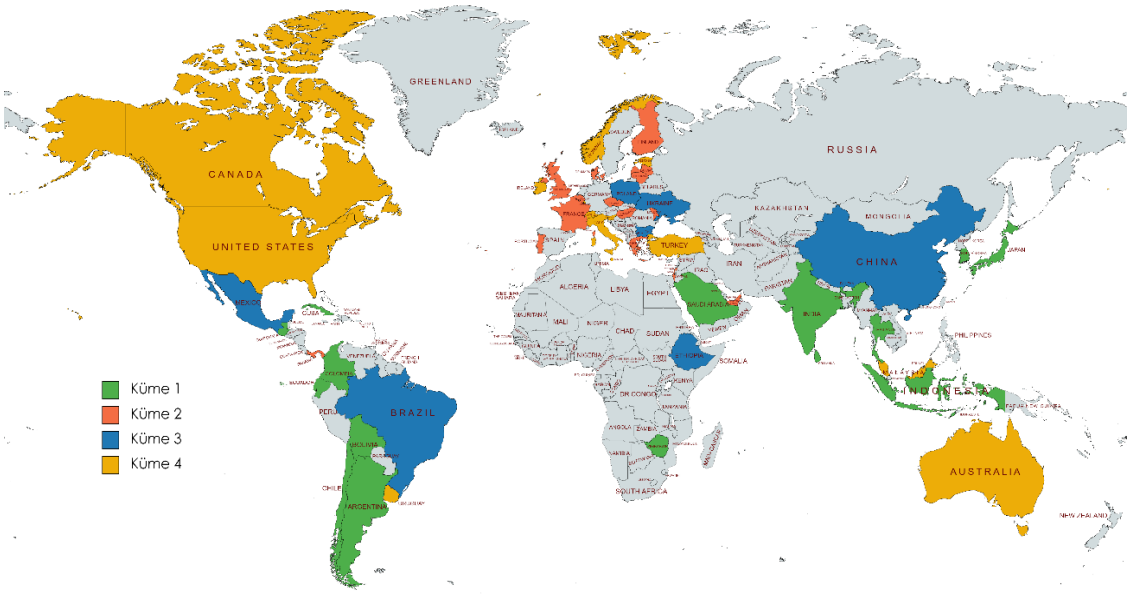

Şekil 4.12 $K$-Ortalamalar yöntemine göre elde edilen sonbahar dönemi kümelerinin Python programı çıktısı

Sonbahar dönemine ait *K*-Ortalamalar yöntemi ile elden edilen kümeler çizelge 4.6'da yer almaktadır.

Çizelge 4.6 *K*-Ortalamalar yöntemine göre elde edilen sonbahar dönemi kümeleri

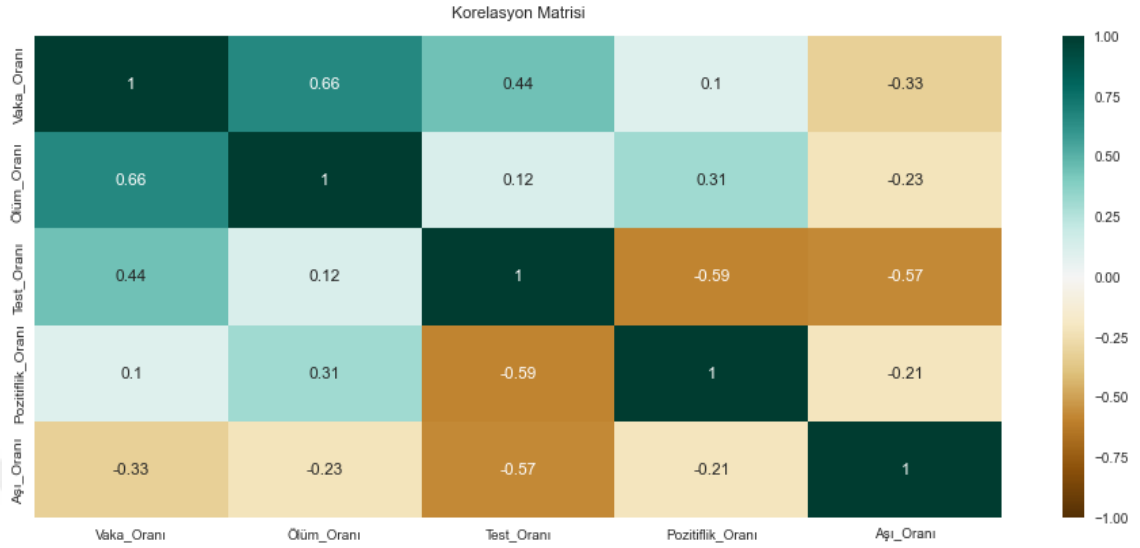
| **Küme 1** | **Küme 2** | **Küme 3** | **Küme 4** |
|---|---|---|---|
| Arjantin | Belçika | Brezilya | Amerika |
| Bolivya | Birleşik Arap Emirlikleri | Bulgaristan | Avustralya |
| Ekvador | Çekya | Çin | Estonya |
| Endonezya | Danimarka | Etiyopya | İrlanda |
| Guatemala | Finlandiya | Meksika | İsviçre |
| Güney Kore | Fransa | Polonya | İtalya |
| Hindistan | İngiltere | Ukrayna | Kanada |
| Japonya | İsrail | | Lüksemburg |
| Kolombiya | Letonya | | Malezya |
| Küba | Litvanya | | Malta |
| Sri Lanka | Macaristan | | Norveç |
| Suudi Arabistan | Maldivler | | Slovenya |
| Şili | Moldova | | Türkiye |
| Tayland | Panama | | Uruguay |
| Zimbabve | Portekiz | | |
| | Yunanistan | | |

Kümelerin renklendirilerek dünya haritası üzerindeki gösterimi Şekil 4.13 ile verilmiştir.




Şekil 4.13 *K*-Ortalamalar yöntemine göre sonbahar dönemi kümelerinin dünya haritasında gösterimi

Şekil 4.14 ile sonbahar dönemine ait korelasyon matrisi verilmiştir.

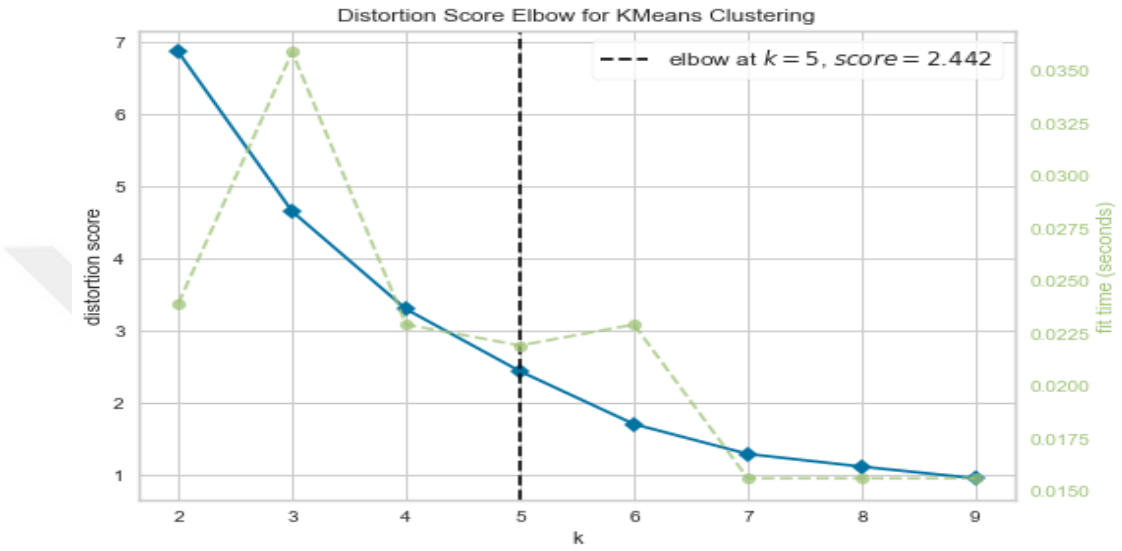



Şekil 4.14 Sonbahar dönemi korelasyon matrisi

Şekil 4.14'te yer alan korelasyon matrisine göre sonbahar döneminde aşı oranı ile test oranı arasındaki korelasyon katsayısı -0.57 olarak elde edilmiştir. Bu iki değişken arasında ters yönlü bir ilişki bulunmaktadır. Yani aşı yaptıranların oranı arttıkça test yaptıranların oranının azaldığı söylenebilir. Benzer şekilde vaka oranı ile ölüm oranı arasında pozitif bir ilişki olduğu da görülmektedir.

Sonbahar döneminde; Ward yöntemine göre Türkiye, Amerika, Belçika, Birleşik Arap Emirlikleri, Çekya, Danimarka, Estonya, Finlandiya, Fransa, İngiltere, İrlanda, İsrail İsviçre, İtalya, Kanada, Letonya, Litvanya, Lüksemburg, Macaristan, Maldivler, Malta, Moldova, Panama, Portekiz ve Yunanistan ile aynı kümede yer almaktadır. *K*-Ortalamalar yöntemine göre ise, Türkiye, Amerika, Avustralya, Estonya, İrlanda, İsviçre, İtalya, Kanada, Lüksemburg, Malezya, Malta, Norveç, Slovenya ve Uruguay ülkeleri ile aynı kümede yer almaktadır. Ward ve *K*-Ortalamalar yöntemlerine göre oluşturulmuş kümeler incelendiğinde Amerika, Estonya, İrlanda, İsviçre, İtalya, Kanada, Lüksemburg, Malta, Türkiye, Çin, Meksika, Slovenya, Ukrayna, Arjantin, Endonezya, Hindistan, Japonya, Kolombiya, Suudi Arabistan ve Şili ülkelerinin farklı kümelerde yer aldığı görülmektedir.

Kış dönemi, 1.12.2021-28.02.2022 arasındaki tarihleri kapsamaktadır. Bu tarih aralığında Ward yöntemi kullanılarak elde edilen sonuçlara göre ülkelerin yer aldığı kümeler çizelge 4.7 ile verilmiştir. Bu dönem için Elbow yöntemi kullanılarak küme sayısı $k$=5 olarak belirlenmiştir (Şekil 4.15). Yaz ve sonbahar dönemi için 4 küme yeterli iken, kış dönemi için ülkeler 5 kümeye ayrılmıştır.




Şekil 4.15 Elbow yöntemine göre elde edilen kış dönemine ait grafik

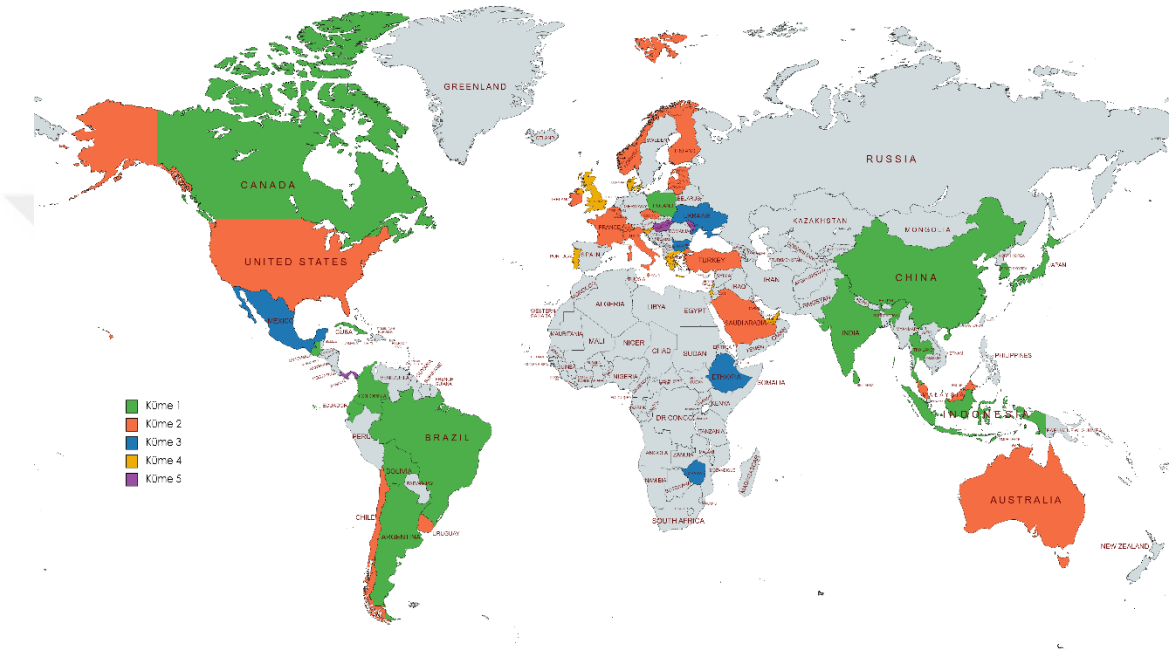
Çizelge 4.7 Ward yöntemine göre elde edilen kış dönemi kümeleri

| **Küme 1** | **Küme 2** | **Küme 3** | **Küme 4** | **Küme 5** |
|---|---|---|---|---|
| Arjantin | Amerika | Bulgaristan | Birleşik Arap Emirlikleri | Macaristan |
| Bolivya | Avustralya | Etiyopya | Danimarka | Moldova |
| Brezilya | Belçika | Meksika | İngiltere | Panama |
| Çin | Çekya | Ukrayna | İsrail | |
| Ekvador | Estonya | Zimbabve | Maldivler | |
| Endonezya | Finlandiya | | Portekiz | |
| Guatemala | Fransa | | Slovenya | |
| Güney Kore | İrlanda | | Yunanistan | |
| Hindistan | İsviçre | | | |
| Japonya | İtalya | | | |
| Kanada | Letonya | | | |
| Kolombiya | Litvanya | | | |
| Küba | Lüksemburg | | | |
| Polonya | Malezya | | | |
| Sri Lanka | Malta | | | |
| Tayland | Norveç | | | |
| | Suudi Arabistan | | | |

Çizelge 4.7 Ward yöntemine göre elde edilen kış dönemi kümeleri (devam)

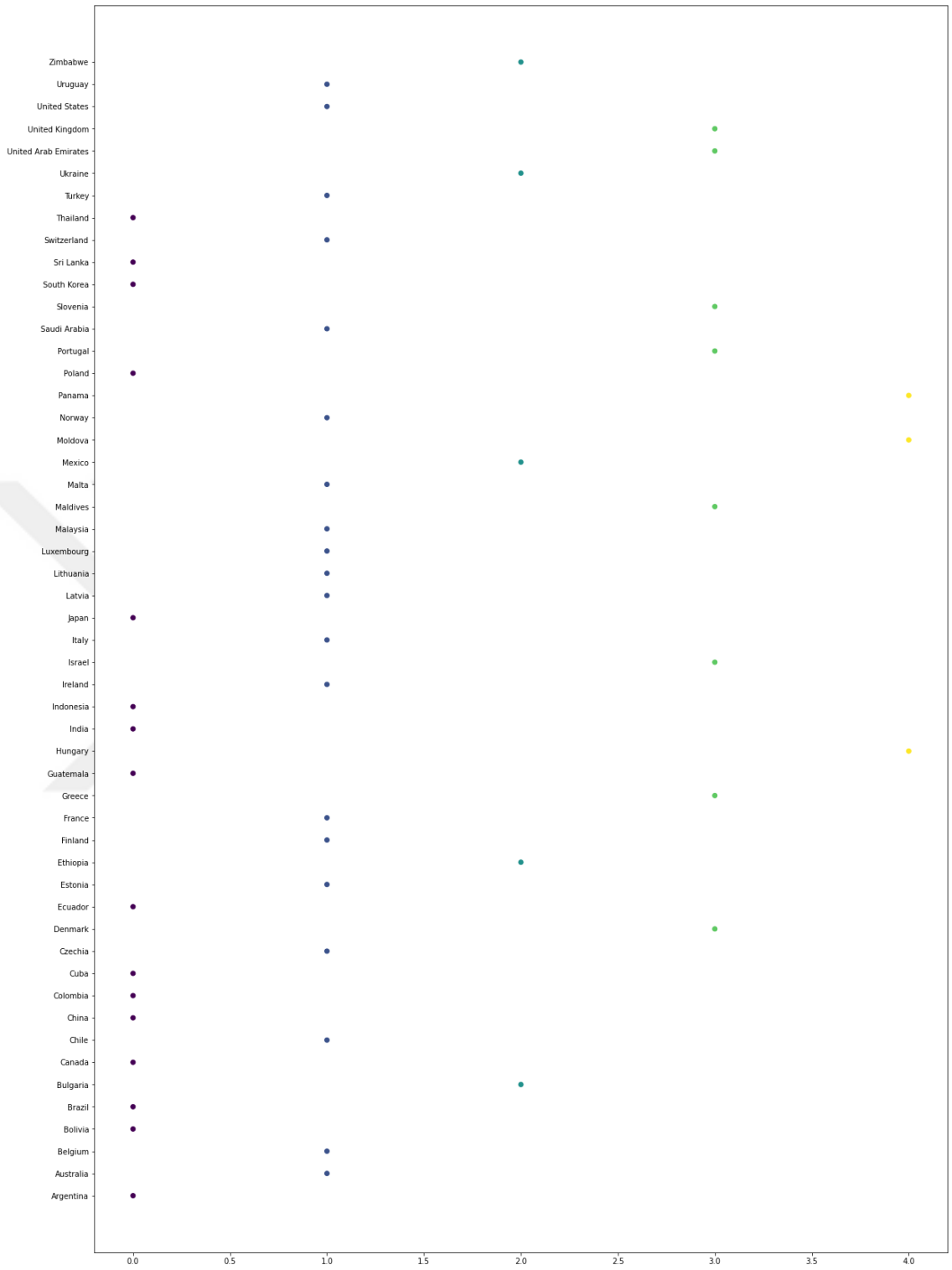
| | | | | |
|---|---|---|---|---|
| | Şili | | | |
| | Türkiye | | | |
| | Uruguay | | | |

Kümeler renklendirilerek dünya haritası üzerindeki gösterimi Şekil 4.16 ile verilmiştir.



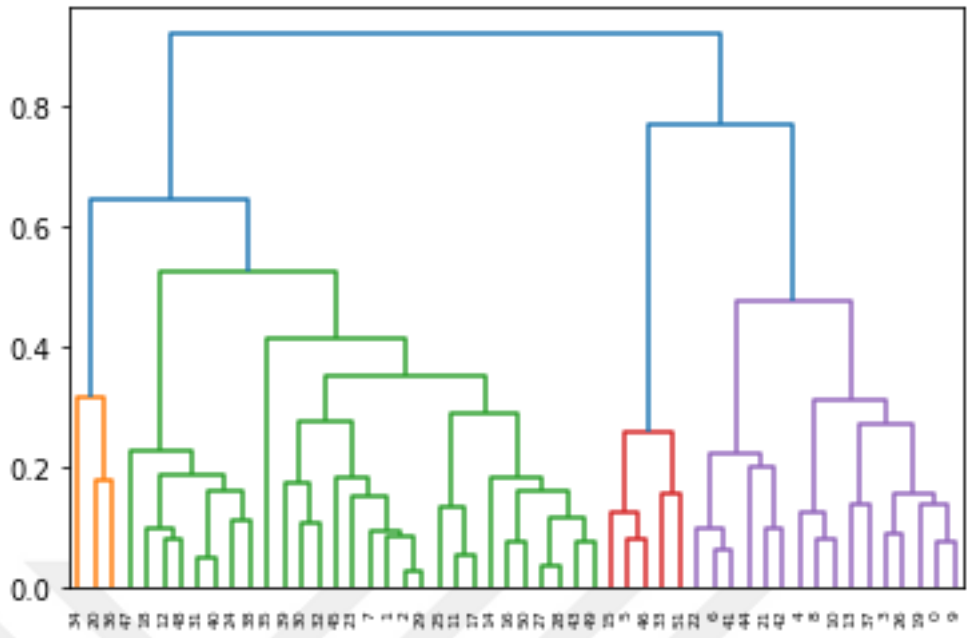

Şekil 4.16 Ward yöntemine göre kış dönemi kümelerinin dünya haritasında gösterimi

Ward yöntemine göre elde edilen kış dönemi kümelerine ait Python programı çıktısı Şekil 4.17 ile verilmiştir.



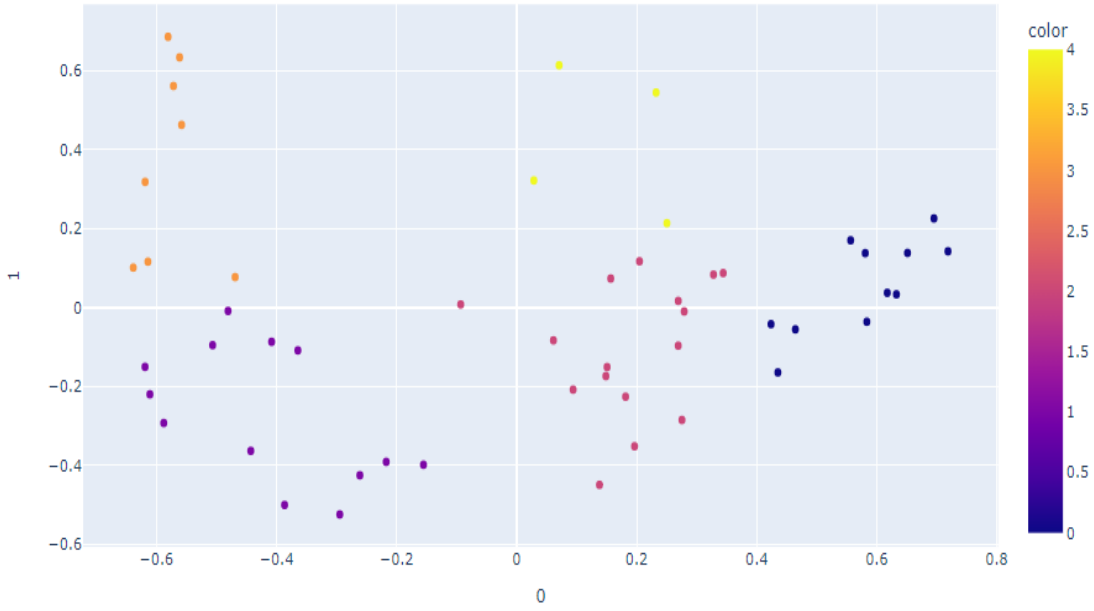

Şekil 4.17 Ward yöntemine göre kış dönemi kümelerinin Python programı çıktısı

Şekil 4.18 ile Ward yöntemine ilişkin dendogram gözükmektedir.




Şekil 4.18 Ward yöntemine göre kış dönemine ait dendogram

Şekil 4.19'da kümelerin gösterildiği Python programı çıktısı verilmiştir.



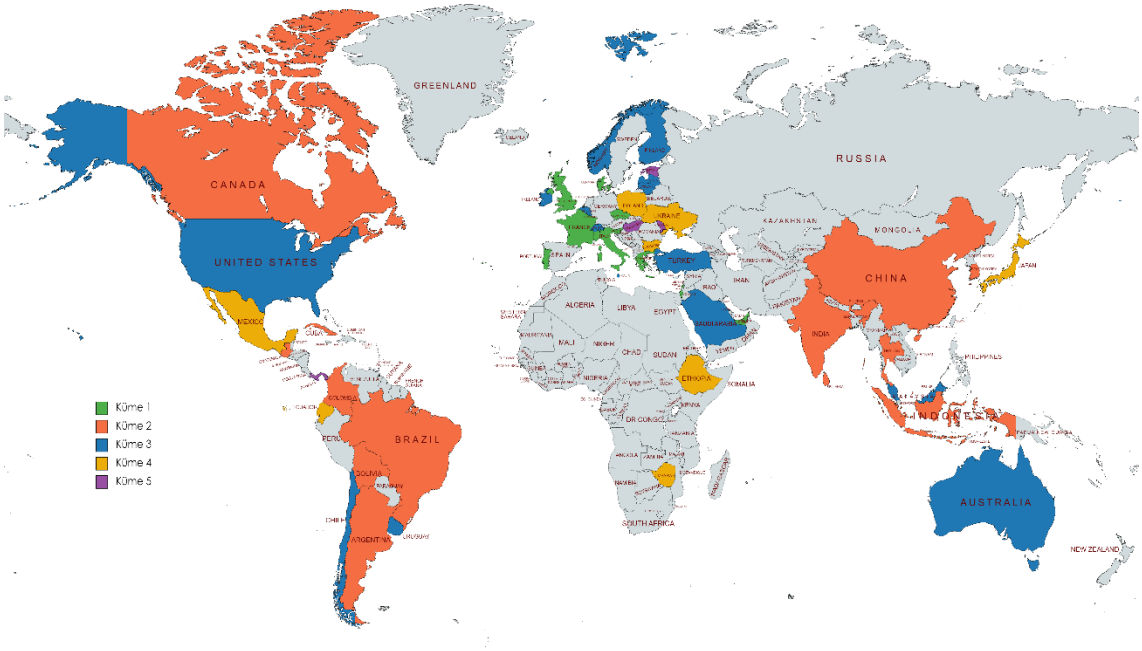

Şekil 4.19 *K*-Ortalamalar yöntemine göre kış dönemi kümelerinin python programı çıktısı

Kış dönemine ait *K*-Ortalamalar yöntemi ile elden edilen kümeler çizelge 4.8'de yer almaktadır.

Çizelge 4.8 *K*-Ortalamalar yöntemine göre elde edilen kış dönemi kümeleri

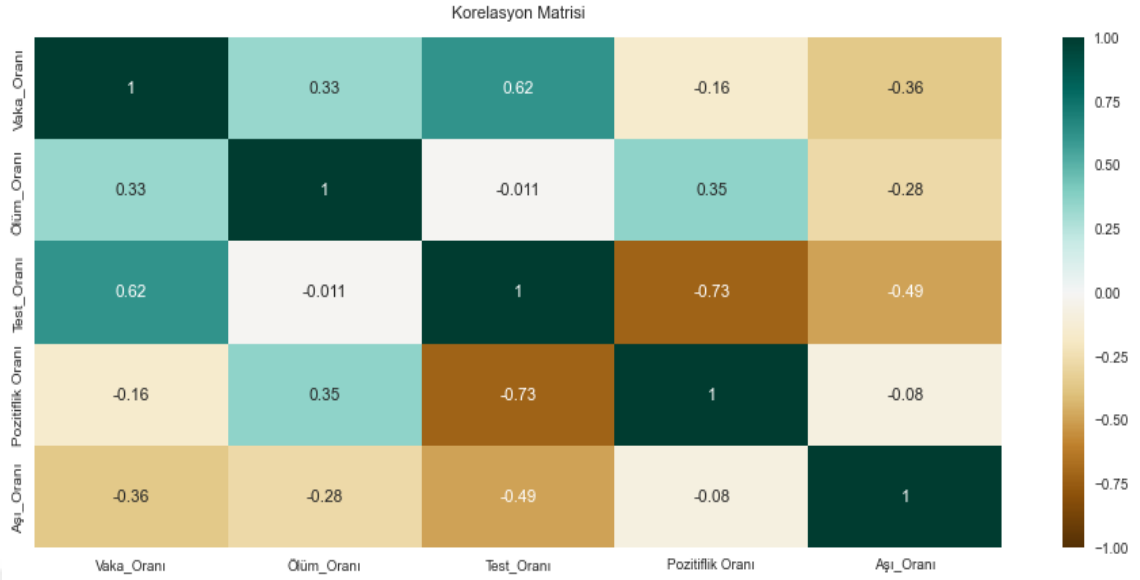
| **Küme 1** | **Küme 2** | **Küme 3** | **Küme 4** | **Küme 5** |
|---|---|---|---|---|
| Birleşik Arap Emirlikleri | Arjantin | Amerika | Bulgaristan | Estonya |
| Çekya | Bolivya | Avustralya | Ekvador | Macaristan |
| Danimarka | Brezilya | Belçika | Etiyopya | Moldova |
| Fransa | Çin | Finlandiya | Japonya | Panama |
| İngiltere | Endonezya | İrlanda | Meksika | |
| İsrail | Guatemala | İsviçre | Polonya | |
| İtalya | Güney Kore | Letonya | Ukrayna | |
| Maldivler | Hindistan | Litvanya | Zimbabve | |
| Portekiz | Kanada | Lüksemburg | | |
| Slovenya | Kolombiya | Malezya | | |
| Yunanistan | Küba | Malta | | |
| | Sri Lanka | Norveç | | |
| | Tayland | Suudi Arabistan | | |
| | | Şili | | |
| | | Türkiye | | |
| | | Uruguay | | |

Kümeler renklendirilerek dünya haritası üzerindeki gösterimi Şekil 4.20 ile verilmiştir.




Şekil 4.20 *K*-Ortalamalar yöntemine göre kış dönemi kümelerinin dünya haritasında gösterimi

Şekil 4.21 ile kış dönemine ait korelasyon matrisi verilmiştir.

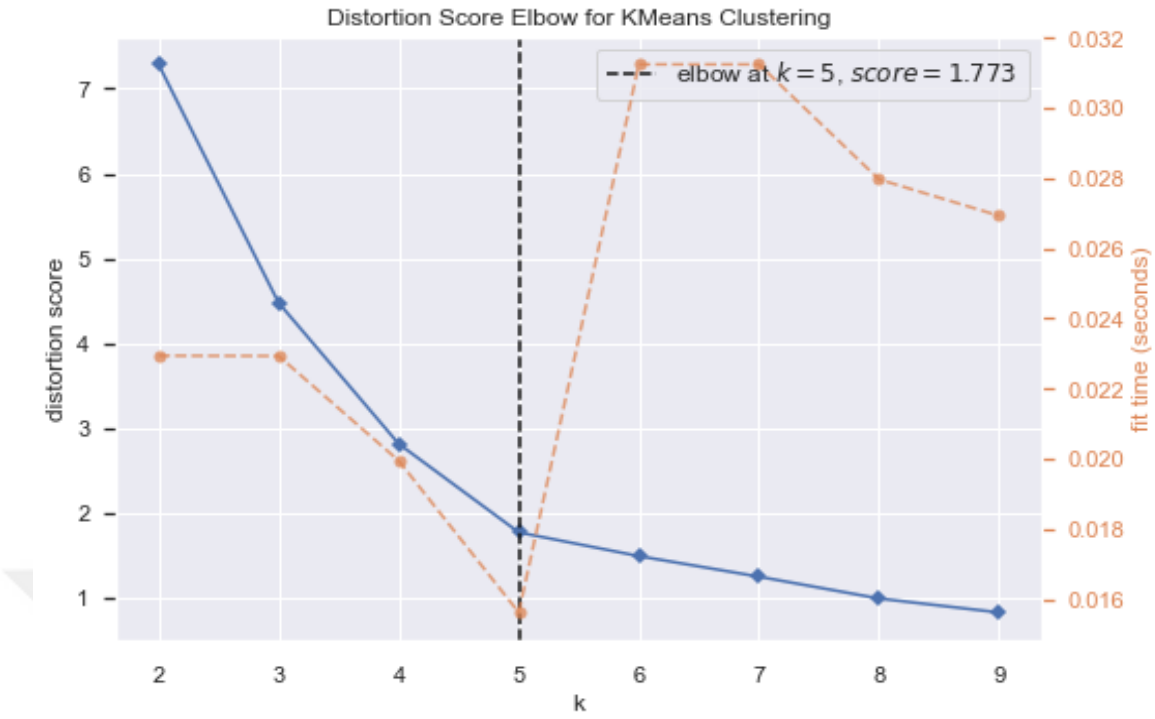



Şekil 4.21 Kış dönemi korelasyon matrisi

Şekil 4.21'de yer alan korelasyon matrisine göre kış döneminde aşı oranı ile test oranı arasındaki korelasyon katsayısı -0.49 olarak elde edilmiştir. Buna göre, iki değişken arasında ters yönlü bir ilişki bulunmaktadır. Yani aşı yaptıranların oranı arttıkça test yaptıranların oranının azaldığı söylenebilir. Benzer şekilde vaka oranı ile ölüm oranı ve vaka oranı ile test oranı arasında pozitif bir ilişki olduğu da görülmektedir.

Kış döneminde; Ward yöntemine göre Türkiye, Amerika, Avustralya, Belçika, Çekya, Estonya, Finlandiya, Fransa, İrlanda, İsviçre, İtalya, Letonya, Litvanya, Lüksemburg, Malezya, Malta, Norveç, Suudi Arabistan, Şili ve Uruguay ile aynı kümede yer almaktadır. *K*-Ortalamalar yöntemine göre ise, Türkiye, Amerika, Avustralya, Belçika, Finlandiya, İrlanda, İsviçre, Letonya, Litvanya, Lüksemburg, Malezya, Malta, Norveç, Suudi Arabistan, Şili ve Uruguay ile aynı kümede yer almaktadır. İki farklı yöntem ile elde edilen kümeler incelendiğinde Çekya, Fransa, İtalya, Ekvador, Estonya, Japonya ve Polonya ülkelerinin farklı kümelerde yer aldığı görülmektedir.

İlkbahar dönemi, 1.03.2022-27.05.2022 arasındaki tarihleri kapsamaktadır. Bu tarih aralığında Ward yöntemi kullanılarak elde edilen sonuçlara göre kümeler çizelge 4.9 ile verilmiştir. Bu dönem için küme sayısı $k$=5 olarak belirlenmiştir. Bu değerin belirlenmesinde Elbow yönteminden faydalanılmıştır (Şekil 4.22).



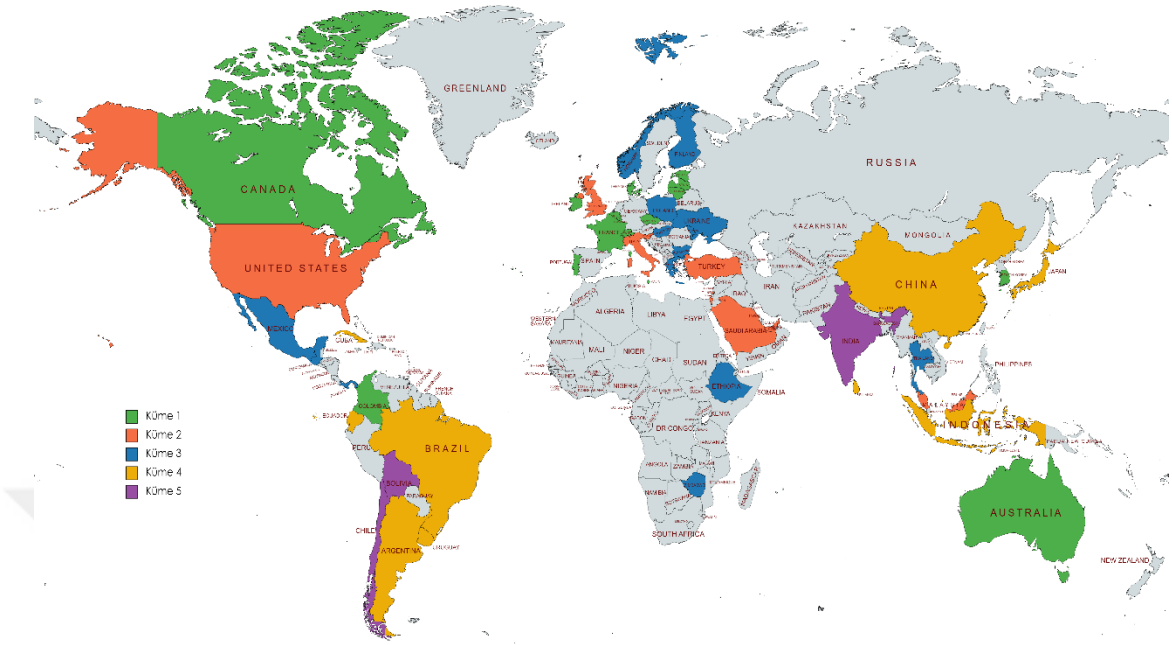

Şekil 4.22 Elbow yöntemine göre elde edilen ilkbahar dönemine ait grafik

Ward yöntemine göre ilkbahar döneminde elde edilen kümelerin ülkelere göre dağılımı çizelge 4.9 ile verilmiştir.

Çizelge 4.9 Ward yöntemine göre elde edilen ilkbahar dönemi kümeleri

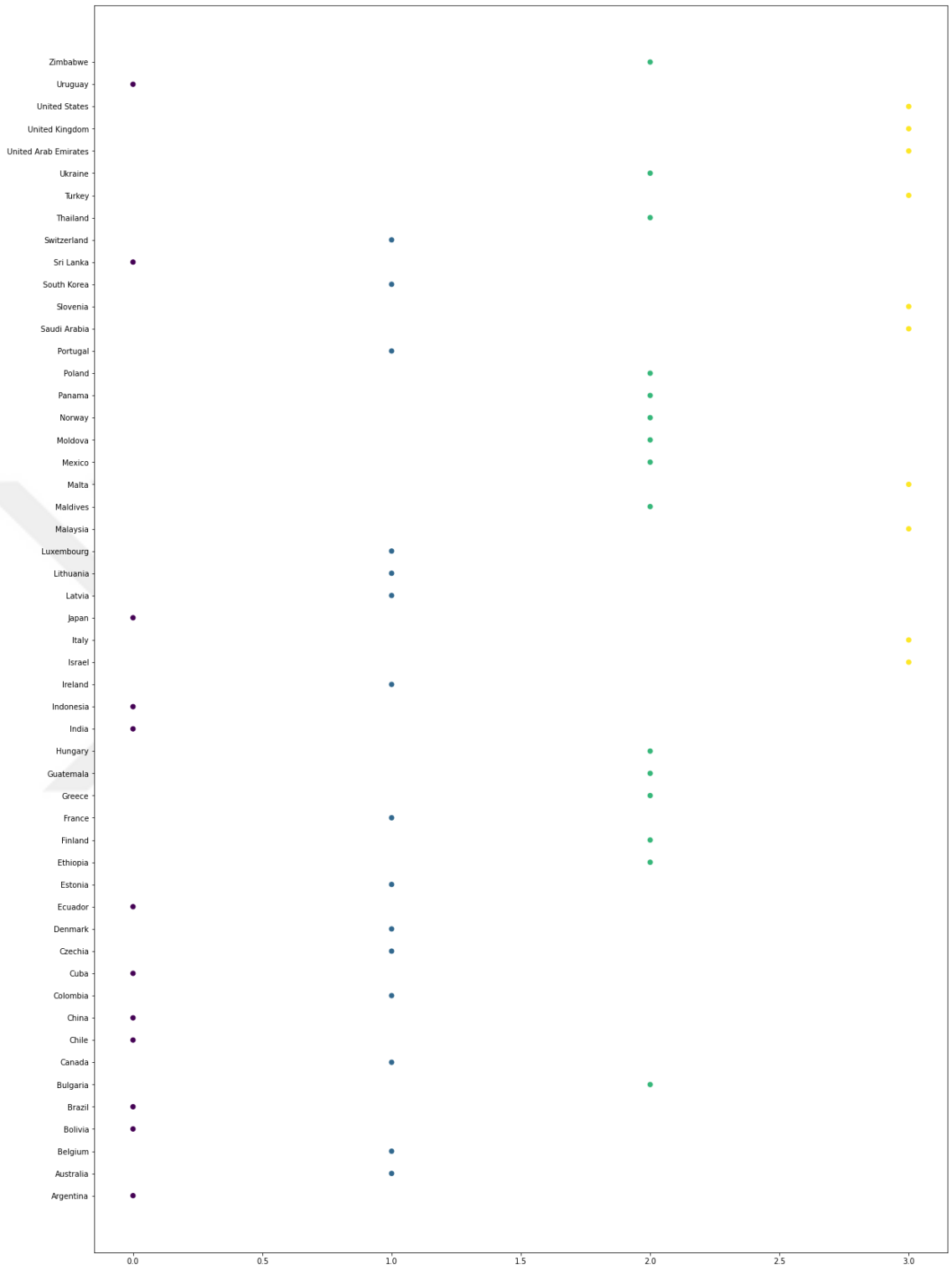
| **Küme 1** | **Küme 2** | **Küme 3** | **Küme 4** | **Küme 5** |
|---|---|---|---|---|
| Avustralya | Amerika | Bulgaristan | Arjantin | Bolivya |
| Belçika | Birleşik Arap Emirlikleri | Etiyopya | Brezilya | Hindistan |
| Çekya | İngiltere | Finlandiya | Çin | Şili |
| Danimarka | İsrail | Guatemala | Ekvador | |
| Estonya | İtalya | Macaristan | Endonezya | |
| Fransa | Malezya | Maldivler | Japonya | |
| Güney Kore | Slovenya | Meksika | Küba | |
| İrlanda | Suudi Arabistan | Moldova | Sri Lanka | |
| İsviçre | Türkiye | Norveç | Uruguay | |
| Kanada | | Panama | | |
| Kolombiya | | Polonya | | |
| Letonya | | Tayland | | |
| Litvanya | | Ukrayna | | |
| Lüksemburg | | Yunanistan | | |
| Malta | | Zimbabve | | |
| Portekiz | | | | |

Kümeler renklendirilerek dünya haritası üzerindeki gösterimi Şekil 4.23 ile verilmiştir.



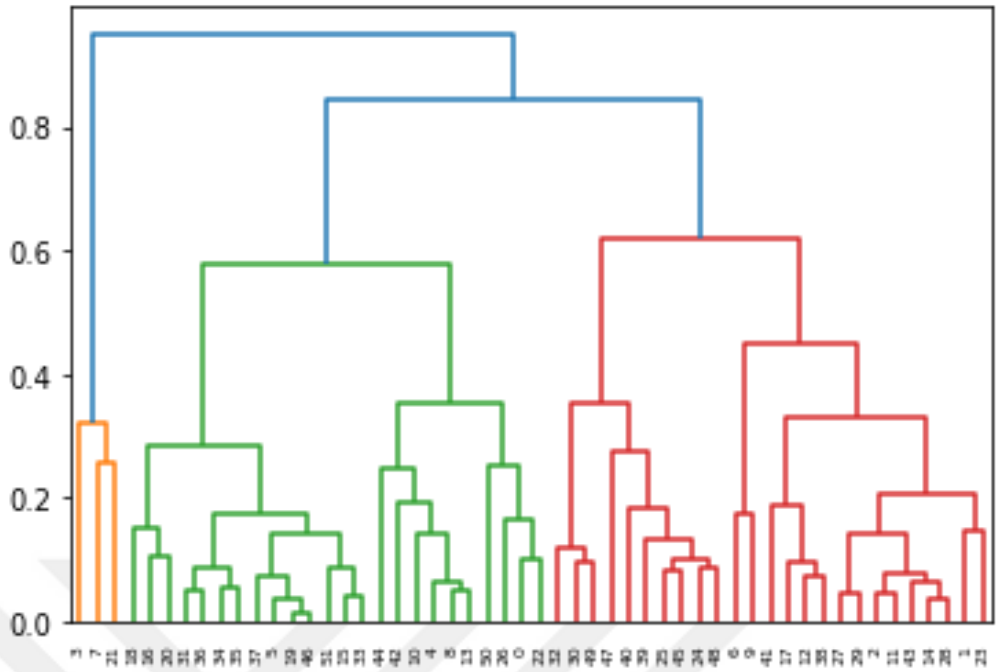

Şekil 4.23 Ward yöntemine göre ilkbahar dönemi kümelerinin dünya haritasında gösterimi

Ward yöntemine göre elde edilen ilkbahar dönemi kümelerine ait Python programı çıktısı Şekil 4.24 ile verilmiştir.



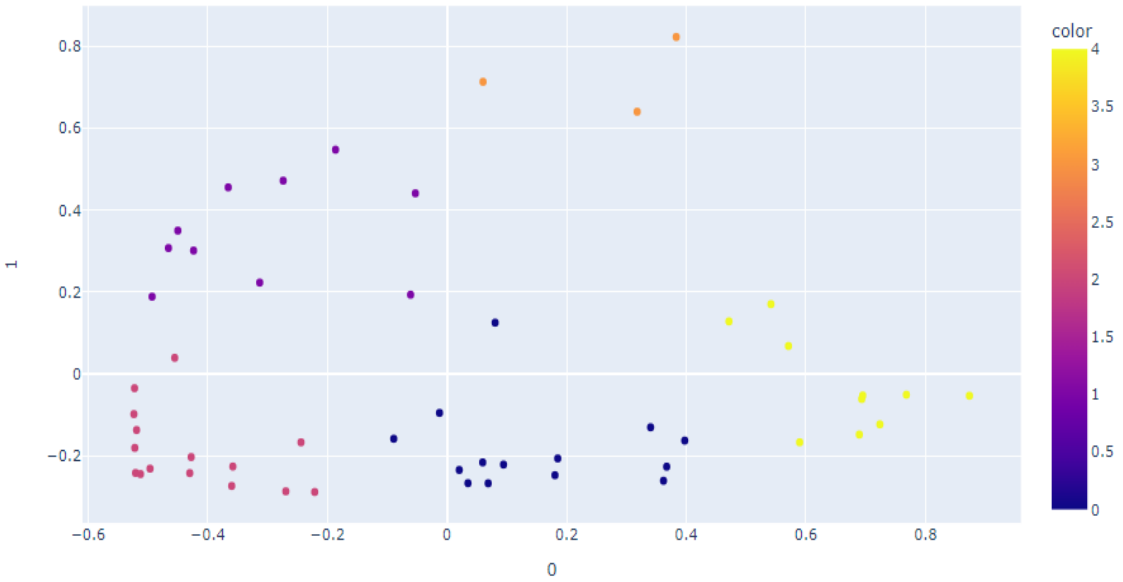

Şekil 4.24 Ward yöntemi ilkbahar dönemi kümelerinin python programı çıktısı

Şekil 4.25 ile Ward yöntemine ilişkin elde edilen dendogram gözükmektedir.




Şekil 4.25 Ward yöntemi ilkbahar dönemine ait dendogram

Şekil 4.26 ile kümelere ilişkin program çıktısı verilmiştir.



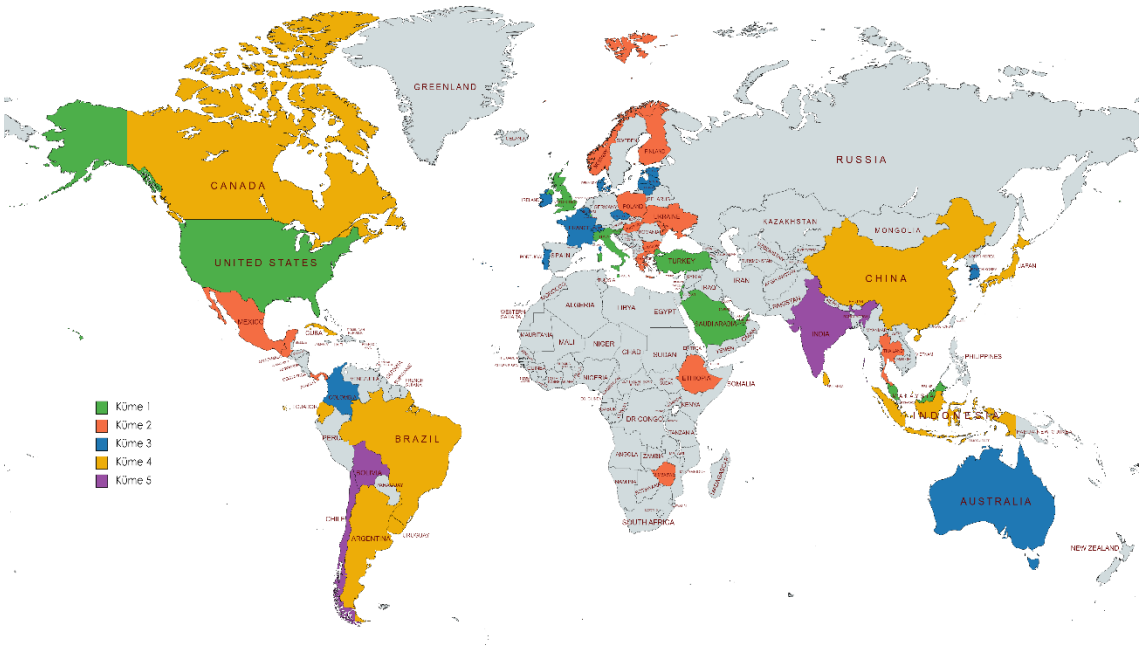

Şekil 4.26 *K*-Ortalamalar yöntemine göre ilkbahar dönemi kümelerinin Python programı çıktısı

İlkbahar dönemine ait *K*-Ortalamalar yöntemi ile elden edilen kümeler çizelge 4.10'da yer almaktadır.

Çizelge 4.10 *K*-Ortalamalar yöntemine göre elde edilen ilkbahar dönemi kümeleri

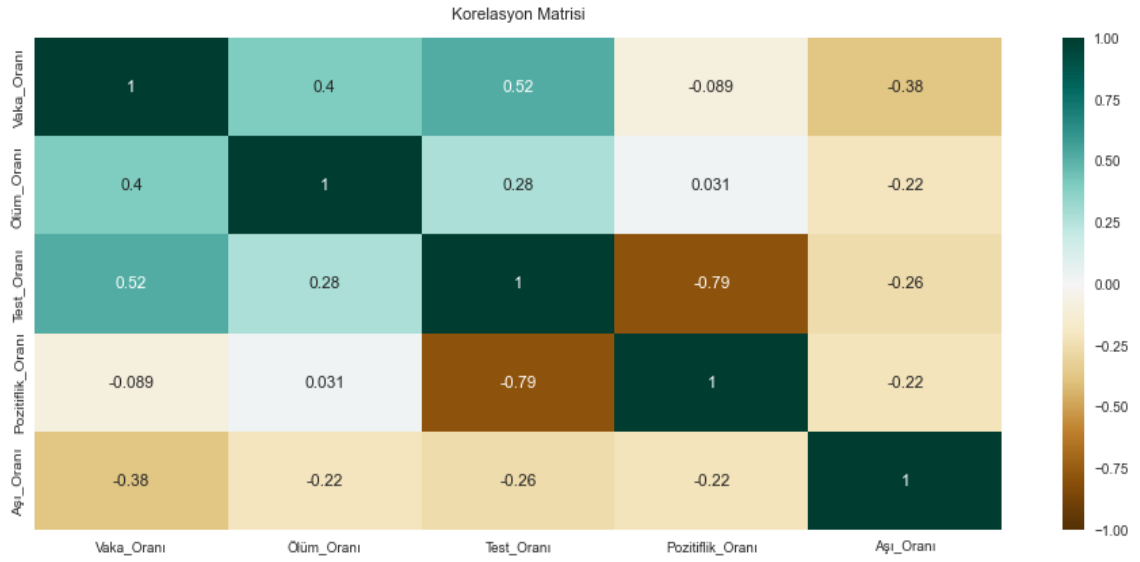
| **Küme 1** | **Küme 2** | **Küme 3** | **Küme 4** | **Küme 5** |
|---|---|---|---|---|
| Amerika | Bulgaristan | Avustralya | Arjantin | Bolivya |
| Birleşik Arap Emirlikleri | Etiyopya | Belçika | Brezilya | Hindistan |
| İngiltere | Finlandiya | Çekya | Çin | Şili |
| İsrail | Guatemala | Danimarka | Ekvador | |
| İtalya | Macaristan | Estonya | Endonezya | |
| Malezya | Maldivler | Fransa | Japonya | |
| Malta | Meksika | Güney Kore | Kanada | |
| Slovenya | Moldova | İrlanda | Küba | |
| Suudi Arabistan | Norveç | İsviçre | Sri Lanka | |
| Türkiye | Panama | Kolombiya | Uruguay | |
| | Polonya | Letonya | | |
| | Tayland | Litvanya | | |
| | Ukrayna | Lüksemburg | | |
| | Yunanistan | Portekiz | | |
| | Zimbabve | | | |

Kümeler renklendirilerek dünya haritası üzerindeki gösterimi Şekil 4.27 ile verilmiştir.




Şekil 4.27 *K*-Ortalamalar yöntemine göre ilkbahar dönemi kümelerinin dünya haritasında gösterimi

Şekil 4.28 ile ilkbahar dönemine ait korelasyon matrisi verilmiştir.




Şekil 4.28 İlkbahar dönemi korelasyon matrisi

Şekil 4.28'de yer alan korelasyon matrisine göre ilkbahar döneminde aşı oranı ile test oranı arasındaki korelasyon katsayısı -0.26 olarak elde edilmiştir. Buna göre, diğer dönemlere kıyasla ilkbahar döneminde bu iki değişken arasında ters yönlü zayıf bir ilişki olduğu söylenebilir. Benzer şekilde vaka oranı ile ölüm oranı ve vaka oranı ile test oranı arasında pozitif bir ilişki olduğu görülmektedir.

İlkbahar döneminde; Ward yöntemine göre Türkiye, Amerika, Birleşik Arap Emirlikleri, İngiltere, İsrail, İtalya, Malezya, Slovenya ve Suudi Arabistan aynı kümede yer almaktadır. $K$-Ortalamalar yöntemine göre ise, Türkiye, Amerika, Birleşik Arap Emirlikleri, İngiltere, İsrail, İtalya, Malezya, Malta, Slovenya ve Suudi Arabistan ile aynı kümede yer almaktadır. $K$-Ortalamalar yöntemine göre oluşturulmuş kümelerde, Ward yöntemine göre oluşturulmuş kümelerden farklı olarak Malta ve Kanada ülkelerinin farklı kümelerde yer aldığı görülmektedir.

Kümeleme analizi sonucunda elde edilen kümelerin değişkenler bakımından aralarında anlamlı bir fark olup olmadığını test etmek için ilk olarak verilerin dağılımının çok değişkenli Normal dağılıma uygunluğu test edilmiştir. Literatürde çok değişkenli Normal dağılıma uygunluk için Mardia, Henze-Zirkler ve Royston testleri kullanılmaktadır. Tez kapsamında kullanılan Python programı Henze-Zirkler testinin

sonuçlarını vermektedir. Yapılan test sonucu dört döneme ait sonuçlar Çizelge 4.11 ile verilmiştir.

$H_0$: Veri setinin dağılımı çok değişkenli Normal dağılıma uygundur.

$H_1$: Veri setinin dağılımı çok değişkenli Normal dağılıma uygun değildir.

Çizelge 4.11 Henze-Zirkler çok değişkenli Normallik testi sonuçları

| | **İlkbahar** | **Yaz** | **Sonbahar** | **Kış** |
|---|---|---|---|---|
| **Test İstatistiği** | 1.826 | 2.429 | 2.368 | 1.886 |
| **p-değeri** | 1.806324408932264e-18 | 1.2115689119887032e-32 | 3.4453468289509105e-31 | 7.223263402903194e-20 |
| **Hipotezler** | $H_0$ red | $H_0$ red | $H_0$ red | $H_0$ red |

Elde edilen sonuçlara göre ilkbahar, yaz, sonbahar ve kış dönemlerinde veri setinin çok değişkenli Normal dağılıma uygunluk göstermediği söylenebilir. Bu nedenle kümeler arasında değişkenler bakımından farklılık olup olmadığının belirlenmesi için tek yönlü MANOVA'nın (Çok Değişkenli Varyans Analizi-Multivariate Analysis of Variance) parametrik olmayan alternatifi kullanılmalıdır. Literatür taramasında bunun için çok değişkenli Kruskal Wallis ve Permutasyon MANOVA (PERMANOVA) yöntemlerinin kullanıldığı görülmüştür. PERMANOVA yöntemi için R programında mevcut paket bulunduğundan bu test R programı kullanılarak gerçekleştirilmiştir. Veri setinin dört dönemi için, $K$-Ortalamalar ve Ward yöntemi ile gerçekleştirilen analizlerin sonuçları Çizelge 4.12-Çizelge 4.19 ile verilmiştir.

$H_0$: Kümler arasında değişkenler bakımından istatistiksel olarak anlamlı bir farklılık yoktur.

$H_1$: Kümler arasında değişkenler bakımından istatistiksel olarak anlamlı bir farklılık vardır.

Çizelge 4.12 $K$-Ortalamalar ilkbahar PERMANOVA testi sonuçları

| | **Serbestlik Derecesi** | **Toplam Kareler** | **Ortalama Kareler** | **F (Pr>F)** | **p-değeri** |
|---|---|---|---|---|---|
| **Kümeler Arası** | 4 | 0,0280 | 0,0070 | 6,4181 | 0,001 |
| **Artık** | 47 | 0,0513 | 0,0011 | | |

Çizelge 4.13 Ward ilkbahar PERMANOVA testi sonuçları

| | **Serbestlik Derecesi** | **Toplam Kareler** | **Ortalama Kareler** | **F (Pr>F)** | **p-değeri** |
|---|---|---|---|---|---|
| **Kümeler Arası** | 4 | 0,028 | 0,007 | 6,4181 | 0,001 |
| **Artık** | 47 | 0,0513 | 0,0011 | | |

Çizelge 4.14 *K*-Ortalamalar yaz PERMANOVA testi sonuçları

| | **Serbestlik Derecesi** | **Toplam Kareler** | **Ortalama Kareler** | **F (Pr>F)** | **p-değeri** |
|---|---|---|---|---|---|
| **Kümeler Arası** | 3 | 0,0964 | 0,0321 | 6,8998 | 0,001 |
| **Artık** | 48 | 0,2236 | 0,0047 | | |

Çizelge 4.15 Ward yaz PERMANOVA testi sonuçları

| | **Serbestlik Derecesi** | **Toplam Kareler** | **Ortalama Kareler** | **F (Pr>F)** | **p-değeri** |
|---|---|---|---|---|---|
| **Kümeler Arası** | 3 | 0,0877 | 0,0292 | 7,9221 | 0,001 |
| **Artık** | 48 | 0,1771 | 0,0037 | | |

Çizelge 4.16 *K*-Ortalamalar sonbahar PERMANOVA testi sonuçları

| | **Serbestlik Derecesi** | **Toplam Kareler** | **Ortalama Kareler** | **F (Pr>F)** | **p-değeri** |
|---|---|---|---|---|---|
| **Kümeler Arası** | 3 | 0,0763 | 0,0254475 | 7,9357 | 0,001 |
| **Artık** | 48 | 0,1539 | 0,0032 | | |

Çizelge 4.17 Ward sonbahar PERMANOVA testi sonuçları

| | **Serbestlik Derecesi** | **Toplam Kareler** | **Ortalama Kareler** | **F (Pr>F)** | **p-değeri** |
|---|---|---|---|---|---|
| **Kümeler Arası** | 3 | 0,1054 | 0,0351 | 9,9135 | 0,001 |
| **Artık** | 48 | 0,1702 | 0,0035 | | |

Çizelge 4.18 *K*-Ortalamalar kış PERMANOVA testi sonuçları

| | **Serbestlik Derecesi** | **Toplam Kareler** | **Ortalama Kareler** | **F (Pr>F)** | **p-değeri** |
|---|---|---|---|---|---|
| **Kümeler Arası** | 4 | 0,0267 | 0,0067 | 4,8805 | 0,003 |
| **Artık** | 47 | 0,0643 | 0,0014 | | |

Çizelge 4.19 Ward kış PERMANOVA testi sonuçları

| | **Serbestlik Derecesi** | **Toplam Kareler** | **Ortalama Kareler** | **F (Pr>F)** | **p-değeri** |
|---|---|---|---|---|---|
| **Kümeler Arası** | 4 | 0,1065 | 0,0266 | 17,917 | 0,001 |
| **Artık** | 47 | 0,0699 | 0,0015 | | |

Buna göre elde edilen sonuçlardan, her iki yöntem için de elde edilen kümeler arasında değişkenler bakımından istatistiksel olarak anlamlı bir fark olduğu söylenebilir.

Hangi değişkenler bakımından kümeler arasında fark olduğunun gözlemlenebilmesi için değişken bazında Kruskal Wallis-H testi uygulanarak sonuca ulaşılabilir. Buna göre, test edilecek hipotezler aşağıdaki gibidir.(i=İ,Y,S,K, İ-İlkbahar, Y-Yaz, S-Sonbahar, K-Kış)

$H_{01i}$ : Kümeler arasında *Vaka_Oranı* kriteri bakımından istatistiksel olarak anlamlı bir farklılık yoktur.

$H_{11i}$: Kümeler arasında *Vaka_Oranı* kriteri bakımından istatistiksel olarak anlamlı bir farklılık vardır.

$H_{02i}$ : Kümeler arasında *Ölüm_Oranı* kriteri bakımından istatistiksel olarak anlamlı bir farklılık yoktur.

$H_{12i}$: Kümeler arasında *Ölüm_Oranı* kriteri bakımından istatistiksel olarak anlamlı bir farklılık vardır.

$H_{03i}$ : Kümeler arasında *Test_Oranı* kriteri bakımından istatistiksel olarak anlamlı bir farklılık yoktur.

$H_{13i}$: Kümeler arasında *Test_Oranı* kriteri bakımından istatistiksel olarak anlamlı bir farklılık vardır.

$H_{04i}$ : Kümeler arasında *Pozitiflik_Oranı* kriteri bakımından istatistiksel olarak anlamlı bir farklılık yoktur.

$H_{14i}$: Kümeler arasında *Pozitiflik_Oranı* kriteri bakımından istatistiksel olarak anlamlı bir farklılık vardır.

$H_{05i}$ : Kümeler arasında *Aşı_Oranı* kriteri bakımından istatistiksel olarak anlamlı bir farklılık yoktur.

$H_{15i}$: Kümeler arasında *Aşı_Oranı* kriteri bakımından istatistiksel olarak anlamlı bir farklılık vardır.

İlkbahar dönemine ilişkin elde edilen sonuçlar, *K*-Ortalamalar yöntemine göre ve Ward yöntemine göre aşağıdaki gibi tablolaştırılarak verilmiştir.

Çizelge 4.20 *K*-Ortalamalar ilkbahar Kruskal Wallis-H testi sonuçları

| | **Vaka_Oranı** | **Ölüm_Oranı** | **Test_Oranı** | **Pozitiflik_Oranı** | **Aşı_Oranı** |
|---|---|---|---|---|---|
| **H istatistiği** | 26,195 | 9,344 | 43,463 | 44,282 | 29,982 |
| **p-değeri** | 0,0000 | 0,0530 | 0,0000 | 0,0000 | 0,0000 |
| **Hipotezler** | $H_{01i}$ red | $H_{02i}$ reddedilemedi | $H_{03i}$ red | $H_{04i}$ red | $H_{05i}$ red |

Çizelge 4.21 Ward ilkbahar Kruskal Wallis-H testi sonuçları

| | **Vaka_Oranı** | **Ölüm_Oranı** | **Test_Oranı** | **Pozitiflik_Oranı** | **Aşı_Oranı** |
|---|---|---|---|---|---|
| **H istatistiği** | 26,565 | 12,035 | 43,844 | 44,308 | 28,551 |
| **p-değeri** | 0,0000 | 0,0171 | 0,0000 | 0,0000 | 0,0000 |
| **Hipotezler** | $H_{01i}$ red | $H_{02i}$ red | $H_{03i}$ red | $H_{04i}$ red | $H_{05i}$ red |

Elde edilen sonuçlara göre, K-ortalamalar yönteminde kümeler arasında ölüm oranı değişkeni bakımından fark görülmezken, diğer tüm değişkenler bakımından kümeler arasında fark olduğu söylenebilir. Ward yöntemine göre ise kümeler arasında tüm değişkenler bakımından fark vardır.

Yaz dönemine ilişkin elde edilen sonuçlar, *K*-Ortalamalar yöntemine göre ve Ward yöntemine göre aşağıdaki gibi tablolaştırılarak verilmiştir.

Çizelge 4.22 $K$-Ortalamalar yaz Kruskal Wallis-H testi sonuçları

| | **Vaka_Oranı** | **Ölüm_Oranı** | **Test_Oranı** | **Pozitiflik_Oranı** | **Aşı_Oranı** |
|---|---|---|---|---|---|
| **H istatistiği** | 8,206 | 9,597 | 40,422 | 26,809 | 42,636 |
| **p-değeri** | 0,0419 | 0,0223 | 0,0000 | 0,0000 | 0,0000 |
| **Hipotezler** | $H_{01i}$ red | $H_{02i}$ red | $H_{03i}$ red | $H_{04i}$ red | $H_{05i}$ red |

Çizelge 4.23 Ward yaz Kruskal Wallis-H testi sonuçları

| | **Vaka_Oranı** | **Ölüm_Oranı** | **Test_Oranı** | **Pozitiflik_Oranı** | **Aşı_Oranı** |
|---|---|---|---|---|---|
| **H istatistiği** | 9,558 | 15,426 | 38,147 | 31,115 | 41,499 |
| **p-değeri** | 0,0227 | 0,0015 | 0,0000 | 0,0000 | 0,0000 |
| **Hipotezler** | $H_{01i}$ red | $H_{02i}$ red | $H_{03i}$ red | $H_{04i}$ red | $H_{05i}$ red |

Yaz dönemi için her iki yöntemde de kümeler arasında tüm değişkenler bakımından istatistiksel olarak anlamlı farklılığın olduğu söylenebilir.

Sonbahar dönemine ilişkin elde edilen sonuçlar, $K$-Ortalamalar yöntemine göre ve Ward yöntemine göre aşağıdaki gibi çizelge 4.24 ve çizelge 4.25 ile tablolaştırılarak gösterilmiştir.

Çizelge 4.24 $K$-Ortalamalar sonbahar Kruskal Wallis-H testi sonuçları

| | **Vaka_Oranı** | **Ölüm_Oranı** | **Test_Oranı** | **Pozitiflik_Oranı** | **Aşı_Oranı** |
|---|---|---|---|---|---|
| **H istatistiği** | 17,692 | 2,305 | 43,345 | 21,479 | 41,126 |
| **p-değeri** | 0,0005 | 0,5116 | 0,0000 | 0,0000 | 0,0000 |
| **Hipotezler** | $H_{01i}$ red | $H_{02i}$ reddedilemedi | $H_{03i}$ red | $H_{04i}$ red | $H_{05i}$ red |

Çizelge 4.25 Ward sonbahar Kruskal Wallis-H testi sonuçları

| | **Vaka_Oranı** | **Ölüm_Oranı** | **Test_Oranı** | **Pozitiflik_Oranı** | **Aşı_Oranı** |
|---|---|---|---|---|---|
| **H istatistiği** | 18,109 | 6,673 | 41,969 | 30,266 | 38,196 |
| **p-değeri** | 0,0004 | 0,0831 | 0,0000 | 0,0000 | 0,0000 |
| **Hipotezler** | $H_{01i}$ red | $H_{02i}$ reddedilemedi | $H_{03i}$ red | $H_{04i}$ red | $H_{05i}$ red |

Her iki yöntem için sonbahar döneminde, ölüm oranı değişkeni bakımından anlamlı fark olmadığı, diğer tüm değişkenler bakımından istatistiksel olarak anlamlı farklılığın olduğu söylenebilir.

Kış dönemine ilişkin elde edilen sonuçlar, $K$-Ortalamalar yöntemine göre ve Ward yöntemine göre aşağıdaki gibi çizelge 4.26 ve çizelge 4.27 ile tablolaştırılarak gösterilmiştir.

Çizelge 4.26 $K$-Ortalamalar kış Kruskal Wallis-H testi sonuçları

| | **Vaka_Oranı** | **Ölüm_Oranı** | **Test_Oranı** | **Pozitiflik_Oranı** | **Aşı_Oranı** |
|---|---|---|---|---|---|
| **H istatistiği** | 25,882 | 12,300 | 44,282 | 37,742 | 40,100 |
| **p-değeri** | 0,0000 | 0,0153 | 0,0000 | 0,0000 | 0,0000 |
| **Hipotezler** | $H_{01i}$ red | $H_{02i}$ red | $H_{03i}$ red | $H_{04i}$ red | $H_{05i}$ red |

Çizelge 4.27 Ward kış Kruskal Wallis-H testi sonuçları

| | **Vaka_Oranı** | **Ölüm_Oranı** | **Test_Oranı** | **Pozitiflik_Oranı** | **Aşı_Oranı** |
|---|---|---|---|---|---|
| **H istatistiği** | 26,034 | 11,384 | 43,917 | 34,274 | 40,498 |
| **p-değeri** | 0,0000 | 0,0226 | 0,0000 | 0,0000 | 0,0000 |
| **Hipotezler** | $H_{01i}$ red | $H_{02i}$ red | $H_{03i}$ red | $H_{04i}$ red | $H_{05i}$ red |

Her iki yöntem için de kış döneminde kümeler arasında tüm değişkenler bakımından anlamlı farklılığın olduğu söylenebilir.

# 5. SONUÇ VE TARTIŞMA

Bu çalışmada, çok değişkenli istatistiğin bir konusu olan ve veri madenciliğinde sıklıkla kullanılan kümeleme analizi yöntemleri araştırılmıştır. Kümeleme analizi yöntemleri, kabaca hiyerarşik ve hiyerarşik olmayan yöntemler olmak üzere iki ana başlıkta düşünülebilir. Yöntemler detaylandırıldığında ise, hiyerarşik yöntemler, bölümlemeli yöntemler, yoğunluğa dayalı yöntemler ve ızgara tabanlı yöntemler olarak sınıflandırılır. Çizelge 5.1'de yöntemlerin detaylarına ilişkin genel özellikler verilmiştir. Bu özellikler yapılacak çalışmalarda hangi yöntemin kullanılmasının uygun olacağına karar verilmesinde yol göstericidir. Veriye ilişkin özellikler, kullanılması düşünülen algoritmalar, oluşturulacak küme sayıları ve aykırı değerler gibi birçok etken kümeleme analizinin sonuçlarını etkilemektedir.

Çizelge 5.1 Kümeleme yöntemlerinin genel özellikleri

| **YÖNTEM** | **GENEL ÖZELLİKLER** |
|---|---|
| **Hiyerarşik Yöntemler** | - Hatalı birleştirmeler veya bölünmeler düzeltilemez.<br>-Kümeleme, çoklu aşamalar ile hiyerarşik bir şekildedir.<br>-Mikro kümeleme gibi diğer teknikleri içerebilir veya nesne bağlantılarını değerlendirebilir. |
| **Bölümlemeli Yöntemler** | -Mesafeye dayalıdır.<br>-Küme merkezini belirlemede ortalama veya orta nokta (vb.) kullanılabilir.<br>-Küçük ve orta ölçekli veri setlerinde etkili sonuçlar verir. |
| **Yoğunluğa Dayalı Yöntemler** | -Farklı şekillere sahip kümeleri bulabilir.<br>-Aykırı değerleri filtreleyebilir.<br>-Kümeler, uzayda düşük yoğunluklu bölgelerle ayrılmış yoğun nesne bölgeleridir. |
| **Izgara Tabanlı Yöntemler** | -Çoklu çözünürlüklü ızgara veri yapısı kullanır.<br>-Hızlı işlem süresine sahiptir. (Veri nesneleri, işlem sayısından bağımsızdır ancak ızgara boyutuna bağlıdır.) |

Bu çalışmada Covid-19 salgınına ilişkin 01.06.2021 ve 27.05.2022 tarihleri arasındaki veriler göz önüne alınmıştır. Bu tarihler dört döneme ayrılarak ülkelere ait vaka oranı, ölüm oranı, test oranı, pozitiflik oranı ve aşı oranı değişkenleri kullanılarak ülkeler benzerliklerine göre kümelere ayrılmıştır. Analizler için Python programından faydalanılmıştır. Bu analiz çalışmasında hiyerarşik yöntemlerden biri olan Ward yöntemi, hiyerarşik olmayan yöntemlerden ise $K$-Ortalamalar yöntemi kullanılmıştır. Her bir döneme de iki farklı yöntem uygulanmış olup elde edilen sonuçlar karşılaştırılmıştır. Elde edilen sonuçları karşılaştırabilmek için her iki yöntemde de küme sayıları Elbow yöntemi ile eşit olarak belirlenmiştir. Bu sonuçlarda ülkemizin hangi ülkelerle aynı kümede yer aldığı yani benzerlik gösterdiği incelenmiştir. Yapılan çok değişkenli analiz ile kümeler arasında değişkenler bakımından istatistiksel olarak anlamlı farklılık olduğu tespit edilmiştir. Farklılığı oluşturan değişkenlerin belirlenmesi için ayrıca değişken bazında Kruskal Wallis-H Testi uygulanmıştır. Kruskal Wallis- H testi sonucunda $K$-Ortalamalar yöntemi için ilkbahar döneminde ölüm oranı değişkeninin etkisiz olduğu, diğer tüm değişkenlerin etkili olduğu görülmüştür. Sonbahar dönemi için ise her iki kümeleme yönteminde de ölüm oranı değişkeninin etkisiz olduğu sonucuna ulaşılmıştır. Bu sonuçlara göre, kümeleme yöntemlerinde ele alınan tüm değişkenlerin kümelemede etkili oldukları söylenebilir (ilkbahar döneminde $K$-Ortalamalar ve sonbahar döneminde her iki kümeleme yöntemi için ölüm oranı değişkeni dışında).

Dönemsel olarak kümeler incelendiğinde, her iki yöntem için oluşturulan kümeler birbirinden farklılık göstermekle birlikte, dönemsel olarak yöntemlerin kendi içinde de farklılıklar gösterdiği söylenebilir. Elde edilen kümeler incelendiğinde, Covid-19 salgını için ülkelerin kümelendirilmesinde ülkelerin gelişmişlik düzeylerinin, coğrafi konumlarının ve nüfuslarının etkisinin olmadığı söylenebilir. Yani, aynı kümede çok farklı konum, farklı gelişmişlik düzeyi ya da nüfusa sahip ülkelerin birlikte yer aldığı görülmektedir. Türkiye göz önüne alındığında, örneğin $K$-Ortalamalar yönteminde sonbahar dönemi için Amerika, Avustralya, Estonya, İrlanda gibi ülkelerle aynı kümede yer alırken yaz dönemi için Japonya, Şili, Ukrayna, Arjantin gibi ülkelerle aynı kümede yer almaktadır. Diğer ülkeler için de dönemsel olarak benzer farklılıklar göze çarpmaktadır.

Genel olarak, Covid-19 salgını ile ilgili bahsedilen veriler kullanılarak yapılan analizler sonucunda, $K$-Ortalamalar ve Ward olmak üzere iki farklı kümeleme yöntemi kullanılarak elde edilen kümelerin birbirinden farklılıklar gösterdiği sonucuna ulaşılmıştır. Bununla birlikte, yapılan analizler sonucunda, kümeleri oluşturmakta (ölüm oranı dışında) tüm değişkenlerin etkili olduğu söylenebilir.

## KAYNAKLAR


Agrawal, R., Gehrke, J., Gunopulos, D. And Raghavan, P. 1998. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. ACM-SIGMOND Management of Data Conference, 94-105, Seattle/The USA.

Akdamar, E. 2019. OECD Ülkelerinin Bazı İş Gücü Piyasası Göstergeleri Kullanılarak Kümeleme Analizi ve Çok Boyutlu Ölçekleme Analizi İle İrdelenmesi. Akademik Araştırmalar ve Çalışmalar Dergisi, 11(20); 50-65.

Akküçük, U. 2011. Veri Madenciliği Kümeleme ve Sınıflama Algoritmaları. Yalın Yayıncılık, 133, İstanbul.

Andritsos, P. 2002. Data Clustering Techniques. Qualifying Oral Examination Paper, 11 March 2002, Toronto/Canada.

Ankerst, M., Breunig, M. M., Kriegel, H.P. and Sander, J. 1999. OPTICS: Ordering Points to Identify the Clustering Structure. Proceedings. International Conference on Management of Data, 1-3 June, ACM SIGMOD Record, 28(2), 49-60, Philadelphia, Pennsylvania/The USA.

Anonim 2022. Web Sitesi: https://github.com/owid/covid-19-data/tree/master/public/data. Erişim tarihi: 28.05.2022.

Balasubramanian, T. ve Umarani, R. 2012. An Analysis on the Impact of Fluoride in Human Health (Dental) Using Clustering Data Mining Technique Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, March 21-23.

Bezdek, J.C. 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York/The USA.

Bilgin, T. T. ve Çamurcu, Y. 2005. DBSCAN, OPTICS ve K-Means Kümeleme Algoritmalarının Uygulamalı Karşılaştırılması. Politeknik Dergisi, 8(2); 139-145.

Bulum, A. Z., Ersöz, F. ve Ersöz, T. 2013. Dünya Ticaret Örgütü (WTO) Üyesi Ülkelerin Uluslararası Ticaret Hacimleri Üzerine Bir Çalışma. Nevşehir Bilim ve Teknoloji Dergisi, 2(2); 153-165.

Çelik, Ş. 2013. Kümeleme Analizi İle Sağlık Göstergelerine Göre Türkiye'deki İllerin Sınıflandırılması. Doğuş Üniversitesi Dergisi, 14(2); 175-194.

Çetintürk, İ. ve Gençtürk, M. 2020. OECD Ülkelerinin Sağlık Harcama Göstergelerinin Kümeleme Analizi İle Sınıflandırılması. Süleyman Demirel Üniversitesi Vizyoner Dergisi, 11(26); 228-244.

Çınaroğlu, S. ve Avcı, K. 2018. Sağlıkta Bölgesel Planlama Çalışmalarında Verimliliğin Artırılması İçin Alternatif Bir Yaklaşım: İki Aşamalı Kümeleme Uygulaması. Verimlilik Dergisi, 2019 (2); 7-25.

Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. 2014. Sosyal Bilimler İçin Çok Değişkenli İstatistik. Pegem Akademi, 416, Ankara.

Dat, N., Phu, V. N., Tran, V. T. N., Chau, V. T. N. and Nguyen, T. A. 2017. STING Algorithm Used English Sentiment Classification in a Parallel Environment. International Journal of Pattern Recognition and Artificial Intelligence, 31(7); 1750021.

Değirmenci, N. ve Yakıcı Ayan, T. 2020. OECD Ülkelerinin Sağlık Göstergeleri Açısından Bulanık Kümeleme Analizi ve TOPSIS Yöntemine Göre Değerlendirilmesi. Hacettepe Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 38(2); 229-241.

Demircioğlu, M. ve Eşiyok, S. 2019. COVID-19 Salgını ile Mücadelede Kümeleme Analizi ile Ülkelerin Sınıflandırılması. İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi, 19(37); 369-389.

Demirkale, Ö. ve Özarı, Ç. 2020. K-Ortalamalar Kümeleme Yöntemi ile Temel Makroekonomik ve Finansal Göstergeler ile Değerlendirilmesi: Kırılgan Beşli Ülkelerinin Örneği. Finans, Ekonomi ve Sosyal Araştırmalar Dergisi, 5(1); 22-32.

Dunn, J.C. 1973. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics, 3(3); 32-57.

Emre, İ.E. ve Selçukcan Erol, Ç. 2017. Veri Analizinde İstatistik mi Veri Madenciliği mi? Bilişim Teknolojileri Dergisi, 10(2); 161-167.

Entezami, A., Sarmadi, H., and Saeedi Razavi, B. 2020. An İnnovative Hybrid Strategy For Structural Health Monitoring by Modal Flexibility and Clustering Methods. Journal Of Civil Structural Health Monitoring.

Erkekoğlu H. 2007. AB'ye Tam Üyelik Sürecinde Türkiye'nin Üye Ülkeler Karşısındaki Göreli Gelişme Düzeyi: Çok Değişkenli İstatistiksel Bir Analiz, 14(2); 28-50.

Ersöz, F. 2009. OECD'ye Üye Ülkelerin Seçilmiş Sağlık Göstergelerinin Kümeleme ve Ayırma Analizi İle Karşılaştırılması, Turkiye Klinikleri Journal of Medical Sciences, 29(6); 1650-1659.

Everitt, B. S., S. Landau, M. Leese and D. Stahl. 2011. Cluster Analysis. John Wiley & Sons, 330, United Kingdom.

Guha, S., Rastogi, R. and Shim, K. 1998. CURE: An Efficient Clustering Algorithm For Large Databases. ACM SIGMOD Record, 27(2); 73-84.

Guha, S., Rastogi, R. and Shim, K. 2001. CURE: An Efficient Clustering Algorithm For Large Databases. Information Systems, 26(1); 35–58.

Han, J., Kamber, M and Pei, J. 2012. Data Mining: Concepts and Techniques. Morgan Kaufmann, 703, USA.

Hemant, P., ve Pushpavathi, T. 2012. A Novel Approach to Predict Diabetes by Cascading Clustering and Classification. 2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12).

Hinneburg, A. and Keim, D. A. 1998. "An Efficient Approach to Clustering in Large Multimedia Databases with Noise", Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD'98), Agustos 1998, 58-65, New York/The USA.

İşler, Y. ve Narin, A. 2012. WEKA Yazılımında K-Ortalama Algoritması Kullanılarak Konjestif Kalp Yetmezliği Hastalarının Teşhisi. SDU Teknik Bilimler Dergisi, 2(4); 21-29.

Kangallı, S. G., Uyar, U. ve Buyrukoğlu, S. 2014. OECD Ülkelerinde Ekonomik Özgürlük: Bir Kümeleme Analizi. Uluslararası Alanya İşletme Fakültesi Dergisi, 6(3); 95-109.

Karypis, G., Han, E.H. and Kumar, V. 1999. Chameleon: Hierarchical Clustering Using Dynamic Modeling. Computer, 32(8); 68–75.

Kassambara, A. 2017. Practical Guide To Cluster Analysis in R: Unsupervised Machine Learning. STHDA, 187.

Kauffman, L. and Rousseeuw, P. J. 1986. Clustering Large Data Sets. Elsevier Science Publishers B. V. North Holland, 425-434.

Kılıç, İ., Emir, O. ve Kılıç, G. 2011. Bulanık Kümeleme Analizi İle Ülkelerin Turizm İstatistikleri Bakımından Sınıflandırılması. İstatistikçiler Dergisi, 4(2011), 31-38.

Kolay, N. ve Erdoğmuş, P. 2016. The Classification of Breast Cancer with Machine Learning Techniques. Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT).

Lorbeer, B., Kosareva, A., Deva, B., Softic, D., Ruppel, P. and Küpper, A. 2018. Variations on the Clustering Algorithm BIRCH. Big Data Research, 11(2018); 44-53.

Mccuddy, M. K. 2010. Economic Freedoms and Public Corruption: A Global Analysis. ASBBS Annual Conference: Las Vegas, 11 (1); 404-418.

Melnykova, N., Shakhovska, N., Gregus, M., Melnykov, V., Zakharchuk, M., and Vovk, O. 2020. Data-Driven Analytics for Personalized Medical Decision Making. Mathematics, 8(8); 1211.

Murtagh, F. and Contreras, P. 2012. Algorithms for hierarchical clustering: An overview. Wires Data Mining Knowl Discov 2(1); 86–97.

Mut, S. ve Akyürek, Ç.E. 2017. Classifying OECD Countries According to Health Indicators Using Clustering Analysis. International Journal of Academic Value Studies, 3(12); 411-422.

Nguyen, Q. and Rayward-Smith, V. J. 2011. CLAM: Clustering Large Applications Using Metaheuristics. Springer Science Business Media B.V. 2011(10); 57-78.

Özdamar, K. 1999. Paket Programlar ile İstatistiksel Veri Analizi 2 (Çok Değişkenli Analizler). Kaan kitabevi, 502, Eskişehir.

Rençber, Ö.F. 2020. Veri Madenciliğinde Kullanılan Kümeleme Algoritmaları ve R ile Uygulamalı Örnekler. Nobel Bilimsel Eserler yayınevi, 186, Ankara.

Rezende, H.R. and Esmin, A.A.A. 2010. Proposed Application of Data Mining Techniques for Clustering Software Projects. INFOCOMP, Special Edition, 43-48.

Rousseeuw, P.J. 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

Sheikholeslami, G. Chatterjee, S. and Zhang, A. 1998. Wavecluster: A Multiresolution Clustering Approach for Very Large Spatial Databases. International Very Large Databases (VLDB), 428-439, New York.

Sığırlı, D., Ediz, B., Cangür, Ş., Ercan, İ. ve Kan, İ. 2006. Türkiye ve Avrupa Birliği'ne Üye Ülkelerin Sağlık Düzeyi Ölçütlerinin Çok Boyutlu Ölçekleme Analizi İle İncelenmesi. İnönü Üniversitesi Tıp Fakültesi Dergisi, 13(2); 81-85.

Silahtaroğlu G. 2020. Veri Madenciliği Kavram ve Algoritmaları. Papatya Bilim yayınevi, 310, İstanbul.

Silitonga, P.D.P. 2017. Clustering of Patient Disease Data by Using K-Means Clustering. International Journal of Computer Science and Information Security (IJCSIS), 15(7); 219-221.

Şahin, D. 2017. Kümeleme Analizi İle Doğu Avrupa Ülkelerinin Ekonomik Özgürlükler Açısından Değerlendirilmesi. Hitit Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10(2); 1299-1314.

Tatlıdil, H. 1996. Uygulamalı Çok Değişkenli İstatistiksel Analiz. Cem Web Ofset, Ankara.

Tekin, B. 2020. COVID-19 Pandemisi Döneminde Ülkelerin COVID-19, Sağlık ve Finansal Göstergeler Bağlamında Sınıflandırılması: Hiyerarşik Kümeleme Analizi. Finans Ekonomi ve Sosyal Araştırmalar Dergisi, 5(2); 336-349.

Tibshirani, R., Walther, G. and Hastie, T. 2001. Estimating the Number of Clusters in a Data Set via the Gap Statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411-423.

Veena, G. S., Sneha, D., Basavaraju, D. and Tanvi, T. 2018. Effective Analysis and Diagnosis of Liver Disorder. International Conference on Communication and Signal Processing (ICCSP), April 3-5, India

Wang, W., Yang, J. and Muntz, R. 1997. STING : A Statistical Information Grid Approach to Spatial Data Mining. Proceedings of the 23rd VLDB Conference Athens/Greece.

Ward, J. H. 1963. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association, 58(301); 236-244.

Yousefi, T., Odabaş, M. S. ve Oktaş, R. 2020. Kümeleme Algoritmalarında Kullanılan Farklı Yöntemlere Genel Bakış. Black Sea Journal of Engineering and Science, 3(4); 173-189.

Zadeh, L.A. 1965. Fuzzy Sets. Information and Control, 8(3), 338-353.

Zhang, J., Liu, L., Fan, Y., Zhuang, L., Zhou, T., and Piao, Z. 2020. Wireless Channel Propagation Scenarios Identification: A Perspective of Machine Learning. IEEE Access, 8(2020), 47797–47806.

Zhu, Y., Ting, K.M., Jin, Y.ve Angelova, M. Hierarchical Clustering That Takes Advantage of Both Density-Peak and Density-Connectivity, Information Systems 103 (2022); 1-16.