T.C.
KONYA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

K-MEANS KÜMELEME YÖNTEMİNDE EN UYGUN KÜME SAYISININ TEMATİK KARTOGRAFYA UYGULAMALARI AÇISINDAN İNCELENMESİ

Utku Can DEMİR

YÜKSEK LİSANS TEZİ

Harita Mühendisliği Anabilim Dalı

ARALIK-2021
KONYA

TEZ KABUL VE ONAYI

Utku Can DEMİR tarafından hazırlanan "K-MEANS KÜMELEME YÖNTEMİNDE EN UYGUN KÜME SAYISININ TEMATİK KARTOGRAFYA UYGULAMALARI AÇISINDAN İNCELENMESİ" adlı tez çalışması 02/12/2021 tarihinde aşağıdaki jüri tarafından oy birliği ile Konya Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Kartografya Anabilim Dalı'nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Danışman Prof. Dr. İbrahim Öztuğ BİLDİRİCİ	……..
Üye Prof. Dr. İsmail Bülent GÜNDOĞDU	……..
Üye Doç. Dr. Hüseyin Zahit SELVİ	……..

Yukarıdaki sonucu onaylarım.

Prof. Dr. Saadettin Erhan KESEN
Enstitü Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Utku Can DEMİR

Tarih:

ÖZET

YÜKSEK LİSANS TEZİ

K-MEANS KÜMELEME YÖNTEMİNDE EN UYGUN KÜME SAYISININ TEMATİK KARTOGRAFYA UYGULAMALARI AÇISINDAN İNCELENMESİ

Utku Can DEMİR

Konya Teknik Üniversitesi
Lisansüstü Eğitim Enstitüsü
Harita Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. İbrahim Öztuğ BİLDİRİCİ

2021, 97 Sayfa

Jüri
Prof. Dr. İbrahim Öztuğ BİLDİRİCİ
Prof. Dr. İsmail Bülent GÜNDOĞDU
Doç. Dr. Hüseyin Zahit SELVİ

Çok değişkenli harita yapımı için kullanılan kümeleme analizi ile iki veya daha fazla konunun birbirleri ile ortak özellikleri olan nesneler bulunur. Bu tez çalışmasında kümeleme analizi yöntemlerinden, hiyerarşik olmayan bir yöntem olan K-Means yöntemi ile iki uygulama yapılması, kullanılan yöntemin gereği olan küme sayısının kullanıcı tarafından girilmesi ile oluşan kümelerin uygunluğunun test edilmesi ve en uygun küme sayısının belirlenmesi amaçlanmıştır. En uygun küme sayısı, Davies-Bouldin İndeksi değeri kullanılarak belirlenmiştir. Burada en önemli kriterlerden birisi Davies-Bouldin İndeks değeridir. Hesaplanan indeks değeri küçüldükçe kümelerin birbirinden ayırt edilme gücü artmaktadır. Hesaplanan indeks değerlerine bakıldığında sıfıra yakın çıkan değer hangi küme sayısına aitse o kümeleme test sonucu olarak en uygundur. Ortaya çıkan sonuçlar yapılan uygulamalar ile gösterilmiştir. 2019 yılına ait km^2'ye düşen taşıt sayısı ve buna karşılık kullanılan akaryakıt miktarı (m^3) verileri ile yapılan ilk uygulamada en düşük indeks değeri 4 küme için hesaplanmıştır. İkinci uygulamada ise 2018 yılına ait ortalama kişi başı tüketilen su miktarı (m^3) ve kişi başı yıllık arz edilen su miktarı (m^3/kişi-yıl) verileri kullanılarak 6 küme için en düşük indeks değeri hesaplanmıştır. K-Means yöntemi ile kümeleme analizi uygulamaları yapılırken Davies-Bouldin İndeks değerinin hesaplanması yararlı olacaktır.

Anahtar Kelimeler: Akaryakıt kullanımı, Araç yoğunluğu, Çok Değişkenli Harita Yapımı, Kümeleme Analizi, K-Means Yöntemi, Su tüketimi, Tematik Kartografya.

ABSTRACT

MASTER THESIS

EXAMINATION OF THE MOST SUITABLE NUMBER OF CLUSTERS IN K-MEANS CLUSTERING IN THE CASE OF THEMATIC CARTOGRAPHY APPLICATIONS

Utku Can DEMİR

Konya Technical University
Institute of Graduate Studies
Department of Geomatics Engineering

Advisor: Prof. Dr. İbrahim Öztuğ BİLDİRİCİ

2021, 97 Pages

Jury
Prof. Dr. İbrahim Öztuğ BİLDİRİCİ
Prof. Dr. İsmail Bülent GÜNDOĞDU
Assoc. Prof. Dr. Hüseyin Zahit SELVİ

With the clustering analysis used for multivariate mapping, objects in which two or more subjects have similar properties are found. It was aimed in this thesis study to perform two applications with K-Means method, which is a non-hierarchical method among clustering analysis methods, to test the suitability of clusters formed through number of clusters being entered by the user, which is by the nature of the method used, and to determine the most suitable number of cluster. The most suitable number of clusters was determined using the Davies-Bouldin Index value. One of the most important criteria in this process is the Davies-Bouldin Index value. Distinctiveness of clusters increases with decreasing index value. Considering the calculated index values, the number of clusters that possesses the value close to zero is the most suitable one. The results were shown with the applications performed. In the first application performed with the data including number of vehicles per km^2 and the corresponding amount of fuel used (m^3) in 2019, the lowest index value was calculated for 4 clusters. In the second application, the lowest index value was calculated for 6 clusters using the data including the average amount of water used (m^3) per capita and the amount of water drawn annually (m^3/person-year). It would be useful to calculate the Davies-Boulding Index value when conducting cluster analysis with the K-Means method.

Keywords: Fuel consumption, vehicle density, multivariate mapping, clustering analysis, K-Means method, water consumption, thematic cartography.

ÖNSÖZ

Tez çalışmam boyunca danışmanım olarak çalışmalarımı yönlendiren, fikirleri ile vizyonumu geliştiren, desteğini hiçbir zaman esirgemeyen, saygıdeğer hocam Prof. Dr. İbrahim Öztuğ BİLDİRİCİ'ye, yüksek lisans eğitimim boyunca verdikleri eğitim ile bilimsel bakış açımı güçlendiren değerli hocalarıma ve tez çalışmam boyunca manevi desteklerini her zaman yanımda hissettiğim, bana ilham veren annem Rabia DEMİR, babam Naci DEMİR ve ablam Güliz ÖZKAYA'ya en içten duygularımla teşekkürü borç bilirim.

Utku Can DEMİR
KONYA-2021

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

Kısaltmalar

AGNES: Birleştirici Hiyerarşik Kümeleme
CBS : Coğrafi Bilgi Sistemleri
DB : Davies-Bouldin İndeks
GSMH : Gayri Safi Milli Hasıla
OECD : İktisadi İşbirliği ve Kalkınma Teşkilatı

1. GİRİŞ

Harita, yeryüzü ile ilişkili olan bilgilerin aktarılması için tasarlanmış olan ve bilgi iletişimini sağlayan bir araçtır. Geleneksel kartografya kaynaklarında haritanın bilgi aktarımı konusundaki etkinliğine değinilmeden sadece görünüşü baz alınarak tanımlamalar yapılmıştır. Bu düşünce ile yapılan tanımlamalarda haritanın iki boyutlu yapıda olduğu, yeryüzünün belirli bir ölçeğe göre küçültülmesi ile elde edildiği ve kuşbakışı özelliğinde olduğundan bahsedilir. Imhof (2007) tarafından yapılmış olan tanım bu düşünceye örnek verilebilir. "Harita, yeryüzünün ya da yeryüzünün bir kısmının küçültülmüş, basitleştirilmiş ve açıklamalarla tamamlanmış iki boyutlu resmidir". Uçar ve Uluğtekin (2006) ise haritanın bilgi iletişim aracı olma özelliğine değinmiştir. Aynı zamanda yeryüzü dışındaki gök cisimlerinin de haritalarının yapılabileceğini vurgulamıştır: "Harita, yer ya da diğer büyük gök cisimlerinin yüzeylerine veya bu yüzeylerin bir bölgesine ait konulara ilişkin obje ve bilgilerin doğadaki konumlarını çizim altlığı üzerinde belli matematik kurallara göre yansıtan, kartografik işaretlerle gösteren ve gerektiğinde yazılı sözcüklerle tamamlayarak aktaran bir bilgi iletişim aracıdır." Kartografya ise harita yapmak ve kullanmak için gerekli bilim sanat ve teknik olarak tanımlanır. Bu tanım günümüzde geçerliliğini korumakta olup Uluslararası Kartografya Birliği özgörü özgörev belgesinde yer almaktadır (Bildirici, 2018).

Haritalar çeşitli bakış açılarına göre sınıflandırılmaktadır (Bildirici, 2018). Bunlardan bazıları:

- Haritada işlenilen konulara ilişkin bilgilerin elde ediliş biçimi
- Ölçek
- Haritada işlenilen konunun içeriği

Haritada işlenilen konulara ait olan bilgilerin elde ediliş şekli açısından haritalar,

- Temel haritalar
- Türetme haritalar

olmak üzere ikiye ayrılır. Temel haritalar, orjinal yersel ve fotogrametrik ölçmelere ve tematik alımlara göre üretilmiş haritalardır. Türetme haritalar ise temel haritalardan ve daha büyük ölçekli türetme haritalardan yararlanılarak, kartografik genelleştirme ile üretilen haritalardır.

Ölçeklerine göre haritalar,

- Büyük ölçekli haritalar
- Orta ölçekli haritalar

- Küçük ölçekli haritalar

şeklinde üçe ayrılır.

Haritalar, işlenilen konunun içeriği bakımından,

- Topografik haritalar
- Tematik haritalar

olarak ikiye ayrılır.

Topografik haritalar, mekânsal olguların (objeler ve olaylar) konumlarını vurgulamak için kullanılır. Tematik haritalar ise birden fazla değişkenin mekânsal dokusunu vurgulamak için kullanılır. Kartografyada topografik haritalar ve tematik haritaların ayrımı haritaların sınıflandırılmasında kolaylık sağladığı içindir. Topografik haritalar aynı zamanda çeşitli öznitelik verilerinin gösterildiği tematik haritalardır. (Slocum ve ark., 2014). Topografik haritalar konusu yeryüzü olan haritalar olarak da nitelendirilir.

Tematik haritalar, topografik bir altlık üzerinde, çalışılan bölge ile mekânsal referanslı olan ve istenilen konularda bilgi aktarımını sağlayan kartografik ürünlerdir (URL1). Tematik haritaların üretilmesi sırasında, işlenmek istenen konuya göre değişmekle birlikte birçok istatistiksel veriye ihtiyaç duyulur. Tematik harita yapımı ile aktarılmak istenen konular iki veya daha fazla ise çok değişkenli istatistiksel analiz yöntemlerinden bahsedilebilir. Elde edilmiş birden fazla veri kümesinin ayrı haritalarda gösterilmek yerine tek bir haritada gösterilmesi, aktarılmak istenen konunun kullanıcı tarafından daha iyi anlaşılmasını sağlayacaktır. Aynı anda birden fazla konunun kullanıcıya aktarılması söz konusu olduğundan elde edilmiş verilerin birbirleriyle ilişkisinin olması hem yapılacak çalışma ve üretilecek ürün açısından hem de istatistiksel anlamda çok değişkenli istatistiksel analiz yöntemlerinin kullanılması aşamasında verimli sonuçların ortaya çıkması için önemlidir.

Çok değişkenli istatistiksel analiz yöntemlerinden biri olan kümeleme analizi ya da kümeleme, benzer elemanları barındıran kümelerin gruplanmasıdır. Aynı kümede bulunan elemanların birbirlerine benzer özellikte olması gerekmektedir. Kümeleme analizi, veri madenciliğinin temel problemlerinden birisidir. Kümeleme analizi aynı zamanda istatistiksel verilerin analizinde de çoğunlukla kullanılan bir tekniktir. Makine öğrenimi, örüntü tanıma, görüntü analizi, bilgi erişimi, biyoenformatik, veri sıkıştırma ve bilgisayar grafikleri alanlarında da çeşitli kullanımları mevcuttur (URL2).

Veri gruplarının birbirleriyle olan benzerlikleri ile sınıflandırılması ve son kullanıcıya uygun verilerin ortaya çıkarılması kümeleme analizinin amaçlarındandır. Son

yıllara baktığımızda istatistiksel sınıflandırma konularında oldukça popüler olan kümeleme analizi algoritmaları, küme sayıları ile ilgili ön bilginin sağlanması ile birlikte çok daha güvenilir sonuçlar verir (Erilli, 2012).

Kümeleme analizi, çok boyutlu veri kümelerinin birbirleri arasındaki ilişkiyi tespit ederek yakından ilişkili olan gruplara atanması ve aynı zamanda bir dizi makine öğrenme algoritmasıdır (URL3).

Kümeleme analizi ile birlikte veri gruplarının birbirleri arasındaki ilişki de göz önüne alınarak kümelenmesi işlemi sağlanmaktadır. İstatistiksel verilerin son kullanıcı tarafından daha iyi anlaşılmasını sağlayacak uygulamalar yapılırken kümeleme analizi aşaması büyük öneme sahiptir. Fazlaca veriye sahip veri gruplarından oluşan istatistiksel bilgiler birbirleri arasındaki benzerlik ilişkileri algoritma tarafından belirlenir ve bu veriler istenilen ya da belirlenmiş olan küme sayısı kadar kümelere ayrılmaktadır. Böylece bu bilgilerin çeşitli uygulamalarda kullanımı daha verimli hale gelmektedir. Burada veri madenciliğinden söz edilebilir. Belirli veri grubundan oluşan bir sistemden anlamlı bilgilerin çıkarımı sağlanmaktadır. Kümeleme analizi, birden fazla ilişkili veri grubu ile birlikte çalışılması durumunda son çıkacak ürünün anlaşılması, yorumlanması açısından büyük öneme sahiptir. Birçok istatistiksel veri ile yapılan uygulamalarda bu analiz işlemi kullanılmaktadır.

Kümeleme analizi yöntemlerinden olan K-Means algoritması küme sayısının kullanıcı tarafından belirlenmesi ile çalışmaktadır. En uygun küme sayısının belirlenmesi K-Means yönteminde karşılaşılan problemlerdendir. Yapılan çalışmada K-Means yönteminde en uygun küme sayısının belirlenmesi sorunu ele alınmış ve uygulama bu doğrultuda yapılmıştır.

İki uygulama yapılarak sonuçlar değerlendirilmiştir. Yapılan ilk uygulamada ülkemizdeki iller bazında 2019 yılında satışı gerçekleştirilen akaryakıt miktarı (benzin ve motorin (ton)), 2019 yılındaki kayıtlı motorlu kara taşıt sayısı ve her bir ile ait olan yüz ölçümü (km^2) elde edilen veri gruplarıdır. Burada taşıt sayısının yüz ölçümüne (km^2) oranına karşılık akaryakıt kullanım miktarının (m^3) yüz ölçümü (km^2) verisine oranlanması ile oluşan km^2'ye düşen araç sayısı ve buna karşılık kullanılan akaryakıt miktarı (m^3) veri grupları ile birlikte kümeleme analizi yapılmıştır. İkinci uygulamada ise 2018 yılına ait elde edilen veri grupları; il bazında nüfus sayıları, dağıtılan suyun abone sayısı, dağıtılan su miktarı (m^3), kişi başı arz edilen günlük su miktarı (m^3) verileri şeklindedir. Buna istinaden illerdeki su kullanımı verileri ile birlikte su tüketim alışkanlıklarına yönelik bir uygulama yapılmıştır. İstatistiksel veriler vasıtasıyla elde

edilen 2018 yılına ait ortalama kişi başı tüketilen su miktarı (m^3) ve yine 2018 yılına ait olan kişi başı yıllık arz edilen su miktarı (m^3/kişi-yıl) verileri ile kümeleme analizi yapılmıştır. Uygulamalar, kümeleme analizi yöntemlerinden K-Means yöntemi ile yapılmış ve en uygun küme sayısı bu doğrultuda belirlenmiştir. Hesaplanan DB değerleri ile ilk uygulamada 4 küme için, ikinci uygulamada 6 küme için en uygun küme sayıları belirlenmiştir. Aynı zamanda çıkan sonuçlar ile birlikte ilk uygulama için iller bazında araç yoğunluğu ve buna karşılık kullanılan akaryakıt oranı verileri, ikinci uygulama için ortalama kullanım için arz edilen su miktarına karşılık ortalama tüketilen su miktarı verileri ile Tematik Kartografya açısından çeşitli çıkarımlarda bulunulmuştur.

2. KAYNAK ARAŞTIRMASI

Akat (2007): Çalışmada, seçilmiş 52 ülke, askeri yapıları ve bu yapıyı temel olarak etkileyen değişkenler kullanılarak kümeleme analizi yardımıyla sınıflandırılmıştır. Kümeleme analizi Ward metodu ve K-Means algoritması ile yapılmış ve elde edilen veriler 2-12 arası kümeler oluşturularak değerlendirilmiştir. Yapılan değerlendirme sonucunda kümeler 8 grupta toplanmıştır. Uygulamada aykırı sonuçlara yol açan 9 ülke çıkarılıp 43 ülke ile değerlendirmeye devam edilmiştir. Birçok ülke farklı kümelerde yer alırken sonuca göre ABD'nin kümede tek başına bulunması dikkat çekmektedir. Bu küme "süper güç" kümesi olarak adlandırılmıştır. Kümelerin oluşmasında etki eden GSMH (Gayri Safi Milli Hasıla) verisi ABD'nin diğer kümelerden ayrışmasında etkili olmuştur.

Bildirici ve Afacan (2018): Yapılan çalışmada Türkiye için istatistik verilerin kümeleme analizi ile gruplanması ve veri gruplarına göre ortak özellikleri olan illerin belirlenmesi ele alınmıştır. Çalışmada kullanılan veriler iller bazında olup işsizlik ve suç üzerinedir. Elde edilmiş veriler kümeleme analizi yöntemlerinden hiyerarşik bir yöntem olan Ward yöntemiyle kümelenmiştir. 5, 6 ve 7'li olarak oluşturulmuş kümelere göre suç oranı ve işsizlik oranı karşılaştırılarak kümeler bazında işsizlik suç ilişkisi ele alınmıştır.

Bircik (2019): Yapılan çalışmada Erzurum ilinde yer alan toplu taşıma sisteminin verimliliği incelenmiştir. Çalışma kapsamında hatlar, hafta içi ve hafta sonu kapasite kullanım oranları ile hafta içi sabah ve akşam pik saatte taşınan yolcu sayıları kullanılarak hiyerarşik olmayan kümeleme analizi yöntemi kullanılmıştır. Elde edilen veriler 3 kümeye ayrılmış ve IBM SPSS Statistics yazılımı ile kümelenmiştir. Orta kümede yer alan hatların verimliliklerinin kabul edilebilir olduğu, düşük kümede kalan hatların kapasitelerinin altında, yüksek kümede yer alan hatların ise kapasitelerinin üzerinde yolcu taşıdıkları belirlenmiş. Çalışma sonucunda belirlenen kümelere istinaden hatların nasıl daha verimli olarak kullanılabileceği konusunda çözüm önerileri getirilmiştir.

Blashfield ve Aldenferder (1978): Makalede 1960'tan beri kümeleme analizi üzerine ilginin olduğundan bahsedilmektedir. Bu ilginin varlığının sebepleri olarak, kümeleme analizi teknikleri üzerine makale sayısındaki artış, konuyla ilgilenen bilim dallarının fazlalığı, kümeleme analizine yönelik yazılımların artması gösterilmiştir.

Blashfield ve ark. (1983): Yapılan çalışmada 23 farklı kümeleme analizi yöntemi karşılaştırılmıştır. Veriler, bir dizi sosyo-davranışsal değişkenle ilgili ve 750 alkol bağımlısı kişi özelliklerini kapsamaktadır. Yapılan uygulamaların sonucunda Ward

kümeleme analizi yönteminin diğer yöntemlere göre daha iyi sonuç verdiği anlaşılmaktadır.

Çağlar (2018): Yapılan çalışmada mekânsal verilerin analizinde veri madenciliği tekniklerinin kullanımı üzerinde durulmuştur. Tematik alan olarak Türkiye'deki trafik kazaları seçilmiş ve belirli yıllara ait olan trafik kaza istatistik veri setleri üzerinde K-Means, K-Medoids yöntemleri ve Birleştirici Hiyerarşik Kümeleme (AGNES) yöntemleri kullanılarak kümeleme analizi yapılmıştır. Kümeleme analizi sonuçları kullanılarak tematik haritalar üretilmiştir. AGNES kümeleme analizi sonucu kullanılarak üretilen haritaların birbirleriyle uyumlu olduğu görülmektedir. Küme sayısı kullanıcı tarafından belirlenmiş olan K-Means ve K-Medoids yöntemlerinde ise K-Medoids yönteminin kümeleri daha iyi ayırdığı gözlemlenmiştir. Ortaya çıkarılan çok değişkenli haritaların risk yönetimi açısından oldukça önemli olduğu görülmektedir.

Dinçer (2006): Yapılan çalışmada veri madenciliğinde bir kümeleme tekniği olan K-Means algoritması incelenmiştir. Bu algoritma kullanılarak bir yazılım geliştirilmiş ve bu yazılım gırtlak kanseri ameliyat verilerinin analizini yapmak üzerinedir. K-Means algoritması ile eldeki veriler kümelenmiş ve oluşan kümeler grafik üzerinde gösterilmiştir. Çalışmada kullanılan K-Means yöntemi tercih nedenleri küme sayısının parametrik olması, sonuçların grafik, yazı ve rakamla kolayca ifade edilebilmesi, uygulamanın kolay olması ve hızlı çalışması olarak sıralanmıştır. Geliştirilen yazılım, tıp doktorlarına geçmiş kayıtları analiz ederek, ileriye dönük tahminleri kolaylaştırmakta ve karar almada yardımcı olabilecek analiz aracıdır.

Han ve ark. (2012): Veri Madenciliği Kavramları ve Teknikleri isimli kitapta çeşitli uygulamalarda kullanılabilecek olan veri madenciliği tekniklerine ayrıntılı olarak yer verilmektedir. Özellikle büyük verilerin işlenmesiyle elde edilecek yararlı bilgilerin çıkarımı hakkında detaylı bilgiler vermektedir. Veri madenciliği kavramında kümeleme analizi algoritmalarının önemi büyüktür. Bu anlamda kümeleme analizi algoritmalarının nasıl çalıştığı, işlevi, kullanımı hakkında bilgiler verilmektedir.

Kalaycı (2020): SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri kitabında temel ve çok değişkenli istatistik tekniklerini uygulayabilme ve sonuçlarını yorumlayabilme, normal dağılım, normallik testi, kümeleme analizi, güvenilirlik analizi, çok boyutlu ölçekleme gibi konulara yer verilmiştir. Kitabın 7. Bölümünde sunulan kümeleme analizi, analizin amacı, analiz sonucunda karar verme süreçleri, hiyerarşik ve hiyerarşik olmayan yöntemlere yönelik uygulamalar ile anlatılmıştır.

Kırmızıbiber (2019): Yapılan çalışmada grid noktalarının kümeleme analizi yöntemlerinden Hiyerarşik ve K-Means yöntemleriyle anlamlı verilerin elde edilebilirliği araştırılmıştır. Kümeleme analizi sonucunda grid noktalarının yön ve yükseklik verileri karşılaştırılarak elde edilen kümelerin yapılan uygulama adına anlamlı olup olmadığı araştırılmıştır. Sonuç olarak yapılan çalışmaya göre hiyerarşik kümeleme yönteminin daha doğru sonuç verdiği saptanmış ve bu sonuca da kendiliğinden ulaşılmadığı, her defasında farklı küme sayısı denenerek ulaşıldığı görülmektedir.

Nacaroğlu (2010): Yapılan çalışmada deprem etkisiyle oluşan boru hasarlarının Coğrafi Bilgi Sistemleri (CBS) ve kümeleme analiziyle değerlendirilmesi yapılmıştır. Çalışma için 1994 Northridge depremi verileri kullanılmıştır. Uygulamada kümeleme analizi yöntemlerinden çıkarımlı küme algoritması ve bulanık c-ortalamalar algoritması kullanılmıştır. Çalışma için seçilen bölgede çeşitli çaplar (km) belirlenmiş ve veriler belirlenen bu bölgeler üzerinden oluşturulmuş. Bulanık c-ortalamalar kümeleme algoritması ile konuma göre (y, x) boru hasarı değerlendirilirken çıkarımlı kümeleme algoritması ile de boru hasarlarının değerlendirilmesi yapılmıştır. Boru hasarları verisi 2-25 arası kümelere bölünerek değerlendirilmiş. Kümeleme yöntemlerinin gömülü boru hatlarındaki deprem hasarlarının değerlendirilmesinde ve hasarların yoğun olduğu alanların belirlenmesinde kullanılması hususu incelenmiş ve yüksek hasar bölgeleri belirlenmeye çalışılmıştır.

Slocum ve ark. (2014): Tematik Kartografya ve Coğrafi Görselleştirme isimli kitapta verileri temsil eden işaretlerin uygun kullanımına yönelik geniş kapsamlı bilgiler yer almaktadır. Verilerin sınıflandırılmasında dikkat edilmesi gereken konular örneklerle birlikte açıklanmıştır. Etkili ve verimli harita tasarımının çeşitli yönlerinden bahsedilmektedir. Tematik haritaların kullanımı, üretimi, görsel tekniklerle bilgi iletimi hakkında detaylı bilgiler verilmektedir. Bunlara ek olarak haritacılık tarihi, ölçek bilgisi, renklendirme, harita elemanları vb. konulardan bahsedilmektedir.

Gülcemal (2019): Yapılan bu çalışmada OECD (İktisadi İş birliği ve Kalkınma Teşkilatı) ülkelerinin sigortacılık ve makroekonomik göstergelerinin sigorta pazar paylarını nasıl etkilediği incelenmiştir. Çalışmanın amacı, sigorta pazar payları açısından ülkelerin birbirlerine benzerlik ve farklılıklarının tespit edilmesidir. OECD ülkelerinin 2016 yılı sigortacılık verileri ve makroekonomik değişkenleri ele alınmış ve sigorta pazar payı üzerine çalışılmıştır. Elde edilmiş olan verilerin analizi, kümeleme analizi yöntemlerinden hiyerarşik olmayan yöntemler grubuna dahil olan K-Means yöntemi ile yapılmıştır. K-Means yöntemi ile hayat sektörü ve hayat dışı sektör için ülkeler

sınıflandırılmıştır. Uygulama için küme sayısının hesaplanması küçük örneklemlerde kullanılan yöntemler baz alınarak yapılmıştır. Bunun sonucunda 4 kümeye ayrılan ülkeler için doğru kümelere ayrıldıklarının belirlenmesi için diskriminant analizi yapılmıştır. Yapılan uygulamanın sonucunda belirlenen faktörler açısından ülkelerin benzer özellik gösterme durumlarına göre kümeleme işlemleri yapılmış ve çıkan sonuçlar ışığında çeşitli değerlendirmeler ortaya konulmuştur.

Güneş (2017): Yapılan çalışmada, havzaların bölgelere ayrılması üzerinde durulmuştur. Farklı kümeleme yöntemleri kullanılarak havza bazında çalışmalar yapılmıştır. Entegre havza yönetimi için gerekli görülen analizler araştırılarak çalışmaya dahil edilmiştir. Elde edilen kümeleme analiz sonuçlarının okuyucuya aktarılmasında etkili olması açısından farklı görselleştirme tekniklerinden yararlanılmıştır. Burada yapılan uygulamaların amacı, su kaynaklarının sürdürülebilir olarak yönetimi, havzaların geliştirilmesi, yatırım programlarına rehberlik sağlamak, havza yönetiminde iklim değişikliğinin etkileri olarak sıralanabilir. Çalışmaya yönelik elde edilen istatistiksel verileri için Hiyerarşik Ward, K-Means, Two Step Cluster Algoritması gibi kümeleme analizi yöntemleri kullanılmıştır. K-Means yönteminde kullanılan bir görsel metot olan gCLUTO algoritmasının, iklim değişikliği ve mevsimsel hareketleri iyi bir şekilde göstermesi, alınabilecek önlemler açısından önemli olduğu çıkarımında bulunulmuştur. Kümeleme yöntemleri kullanılarak homojen ve heterojen havza bölgeleri belirlenmiştir. Belirlenen bölgelerle birlikte yeni havza bölgelerinin oluşturulabileceğinden, havzaların bölgeselleştirilmesinde ana yöntem olarak kümeleme analizinin kullanılabileceğinden bahsedilmiştir.

Erilli (2012): Yapılan çalışmada; Kümeleme, Bulanık Kümeleme Analizi, Bulanık Sayılarda Kümeleme Analizi yöntemlerinden bahsedilmiş, aralarındaki farklara değinilmiştir. Çalışmanın uygulama bölümü, satranç oyunlarının maç skorlarına göre sınıflandırılması üzerinedir.

Yılmaz (2009): Çalışmada, Türkiye genelinde 19 su toplama havzasından alınan veriler kullanılmış ve elde edilen parametrelerin çok değişkenli istatistiksel analiz yöntemleri ile incelenmesi amaçlanmıştır. Veriler üzerinde Ana Bileşenler Analizi, Faktör Analizi ve Kümeleme Analizi uygulanmıştır. Parametreler üzerinde uygulanan ilk iki analizden sonra en son Kümeleme Analizi uygulamasında Ortalama Bağlantı Tekniği ve Pearson Uzaklık Ölçütü kullanılmış ve bunun sonucunda benzerlik matrisleri oluşturulmuştur. 3 adet küme elde edilmiş olup her bir kümenin ayrı bir faktöre bağlı istasyon verilerine göre biçimlendiği görülmüştür.

Karakuş (2020): Yapılan çalışmada, İstanbul'un iki yakasında yer alan ve nüfus bakımından iki büyük ve iki küçük ilçesine ait trafik kaza verileri elde edilmiştir. Uygulamada trafik kazalarına neden olan farklı parametrelerin kümelenmesi amaçlanmıştır. Kazaların cinsiyet, yaş, araç hızı, hava durumu, yol yüzeyi, hafta günleri, kaza türleri gibi parametrelere göre dağılımları tespit edilmiştir. Kümeleme Analizi yapılan çalışmada yöntem olarak K-Means tercih edilmiştir. Küme sayısının belirlenmesinde farklı sayıda kümeler için algoritma uygulanmış ve bunun sonucunda 12 kümeli analizin tutarlı sonuç verdiği yorumu yapılmıştır. Veri sayısı az olan kümeler çalışmaya dahil edilmemiştir. K-Means Yöntemi, kaza gruplarının oluşturulmasına, kümelerin genel karakterlerinin açıklanmasına yardımcı olmaktadır. Yapılan analizde SPSS yazılımı kullanılmıştır. Ortaya çıkan sonuç verilere göre kent içi trafik kazalarına etki eden sebepler tespit edilmiştir. Kent içi trafik kazalarında kaza türlerinin dağılımları belirlenmiştir.

Altınok (2019): Yapılan çalışmada, veri madenciliğine dahil olan hiyerarşik kümeleme algoritmalarından CLUCDUH ve ROCK algoritmaları seçilmiş ve veri grubu üzerinde yöntemler karşılaştırılmıştır. Uygulama, R yazılımı ile yapılmış ve aynı zamanda CLUCDUH algoritmasının R kodları geliştirilmiştir. Elde edilen kümeler için en uygun küme sayısının belirlenebilmesi için Dunn, Davies-Bouldin, Gamma gibi indeks hesaplama yöntemlerinden yararlanılmıştır. Yapılan hesaplama sonucunda ROCK algoritmasının daha iyi kümeler oluşturduğu belirlenmiştir.

Na ve ark. (2010): Bu makalede, Kümeleme Analizi yönteminin veri madenciliğinde kullanılan ana yöntemlerden biri olduğuna değinilmiştir. Kullanılan kümeleme algoritmalarının kümeleme sonuçları üzerinde doğrudan etkisi olduğundan bahsedilmiştir. Burada K-Means algoritmasının işleyişinden, eksiklerinden bahsedilmiştir. K-Means algoritması, kümeleme işlemi yaparken yinelemeli olarak çalışmaktadır. Burada her bir yineleme sonrasında bazı verilerin kaydedileceği ve sonraki yineleme için baz alınacak verilerden yararlanılacağı yani bir K-Means algoritması önerilmektedir. Geliştirilen yöntem ile her bir veri nesnesinin oluşan küme merkezlerine olan uzaklıklarının tekrar hesaplanmasını önleyerek çalışma süresinden tasarruf sağlanmaktadır.

Sinaga ve Yang (2020): Yayında, Kümeleme Analizi açısından K-Means Algoritmasının en çok bilinen ve kullanılan kümeleme yöntemlerinden olduğundan bahsedilmiş ve literatürde çeşitli K-Means eklentilerinin olduğu bilgisi verilmiştir. K-Means Algoritması ve çeşitli eklentileri analiz başlangıcında öngörülen "sınıf sayısı"

bilgisinden etkilenmektedir. Bu nedenle K-Means Algoritmasının tam olarak denetimsiz bir kümeleme yöntemi olmadığından bahsedilmiştir. Bu yayında, K-Means Algoritması için parametre seçimi (küme sayısı, yineleme sayısı vs.) olmadan yapılan analizlerde en uygun sayıda kümenin bulunabilmesi için denetimsiz bir öğrenme şeması oluşturulmuştur. Yani en uygun sayıda kümenin otomatik olarak bulunması amaçlanmıştır. Yayında sunulan yöntem U-K-Means (Unsupervised K-Means) olarak adlandırılmıştır. Diğer mevcut yöntemler ile karşılaştırılması yapılmış ve deneysel sonuçlar ışığında uygulanan algoritmanın olumlu yönlerine değinilmiştir.

Bhatia (2004): Bu yayında; kümelemenin, elde edilen verilerden etkili bir şekilde yararlanmak için kullanılabileceğinden bahsedilmiştir. K-Means Algoritmasının kümeleme teknikleri arasındaki popülerliğine değinilmiştir. K-Means Algoritmasının çalışma şeklinden bahsedilmiş ve kümeleme analizi işlemi başlamadan küme sayısının önceden belirlendiğine ve bu tekniğin belirlenen küme sayısına bağlı olduğuna değinilmiştir. Burada işlenen konu, K-Means Algoritması ile yapılan kümeleme analizinin başlangıç aşamasında belirlenmesi gereken k adet küme sayısı seçimine bağlı kalmadan kümeleme analizini gerçekleştirmek için bir algoritmanın sunulması, küme sayısı seçimine bakılmaksızın kümelerin iyileştirilmesi üzerinedir. Oluşan kümeler elde edilen tekniğe göre büyüyebilir.

Sayılan (2019): Yapılan çalışmada, kümeleme analizinin veri madenciliğinde yaygın olarak kullanıldığından, çok değişkenli bir istatistiksel yöntem olduğuna değinilmiştir. Çalışmadaki amaç, kümeleme analizi için kullanılan bazı algoritmaların farklı özelliklere sahip veri grupları üzerindeki etki ve performanslarının karşılaştırılmasıdır. Bu kapsamda AGNES DIANA, K-Means, K-Medoids, fuzzy c-ortalama, CLARA, DBSCAN, CURE, PAM, DENCLUE, STING, CLIQUE kümeleme analizi algoritmaları incelenmiştir. Algoritmaların uygulama aşamasında R istatistiksel programlama dili kullanılmıştır. Çalışmanın sonucunda farklı veri grupları ile uygulanan algoritmaların sonuç vermedeki başarı durumları tartışılmıştır.

3. MATERYAL VE YÖNTEM

3.1. Tematik Kartografya

Kartografyada haritalar işledikleri konular açısından ikiye ayrılırlar (Slocum ve ark., 2014). Bunlar;

- Tematik Haritalar
- Genel Referans Haritaları (topografik haritalar, fiziki haritalar)

Genel referans haritaları mekânsal özniteliklerin konumlarını vurgulamak için kullanılmaktadır. Topografik haritalar genel referans haritalara örnek olarak verilebilir. Topografik haritalarda kullanıcılar akarsuların, yolların, binaların ve benzeri birçok doğal ve yapay objelerin konumlarını belirleyebilirler. Tematik ya da istatistiksel haritalar ise nüfus yoğunluğu, ailelerin gelirleri ve günlük maksimum sıcaklıklar gibi bir ya da daha çok coğrafi özelliğin (ya da değişkenin) mekânsal dokusunu vurgulamak için kullanılırlar. Koroplet haritalar yaygın olarak kullanılan tematik harita türüne örnek olarak verilebilir. Örnek olarak idari birimler gibi sayım birimlerinin bir özniteliğin farklı birimlerini temsil etmek üzere renklerle doldurulduğu (farklı renk ya da değişen tonlarda renkler) haritalar verilebilir (Slocum ve ark., 2014).

Kartografyada genel referans ve tematik harita ayrımı genel olarak yapılıyor olsa da bu ayrım aslında üretilen haritaları sınıflandırmada kolaylık sağladığı için yapılır. Genel referans haritası da eş zamanlı birden fazla özniteliklerin gösterildiği bir tematik harita olarak görülebilir ve çok değişkenli tematik harita olarak tanımlanabilir. Ayrıca genel referans haritalarında temel vurgu mekânsal objelerin konumu olsa da bununla birlikte belirli bir özniteliğin dağılımı da gösterilebilir (Slocum ve ark., 2014).

3.2. Tematik Haritalar

Tematik haritalar belirli konunun işlendiği ve okuyucuya bir ya da birden fazla öznitelik hakkında bilgi vermek amacıyla kullanılan haritalardır. Dört temel tematik harita yapım yönteminden bahsedilebilir.

3.2.1. Koroplet harita

Koroplet haritalar genellikle iller, ilçeler gibi sayım birimlerinde toplanan verilerin gösterilmesi amacıyla kullanılır. Bu tür haritalar sayım birimlerinden elde edilen verilerin sınıflandırılarak gruplara ayrılması ve her gruba bir gri ton ya da renk atanması ile oluşturulur. Koroplet haritanın sayım birimleri sınırında ani veri değişimi olması durumunda kullanıma uygun olduğu ortaya çıkar. Harita kullanıcısı sayım birimlerinde tipik değerlere yoğunlaşmak isterse ani veri değişimleri olmasa da yine koroplet haritalar

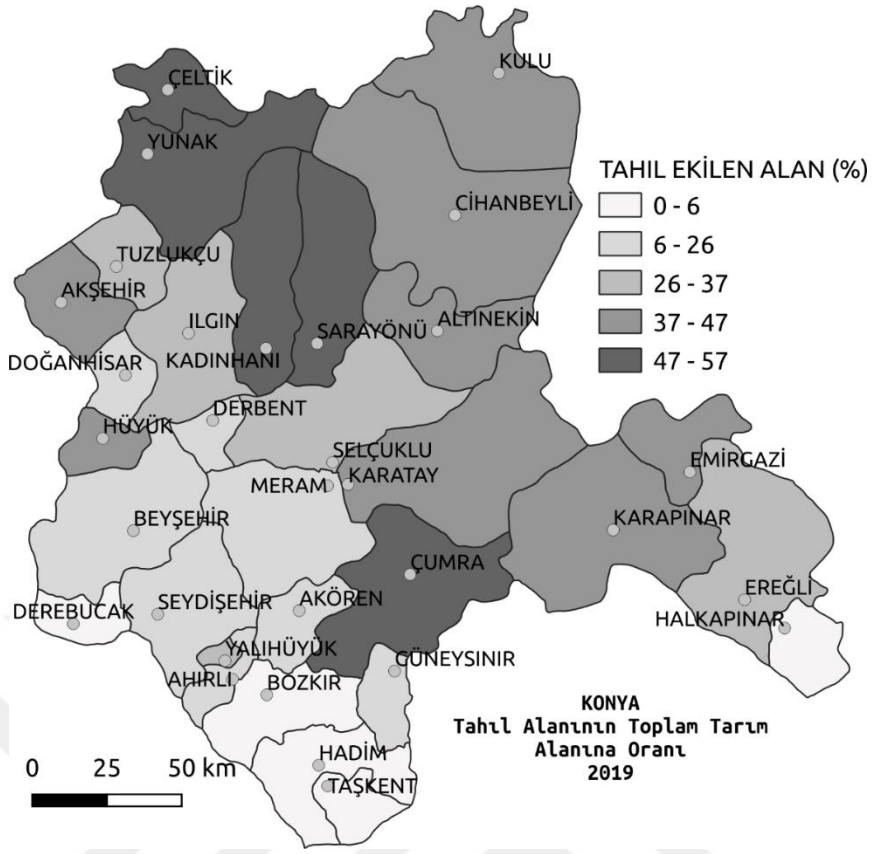
uygun bir tekniktir. Örneğin siyasetçiler ya da yöneticiler tarafından il, ilçe gibi birimler arasında karşılaştırma yapılmak istenirse yine bu tekniğin kullanılması uygundur (Slocum ve ark., 2014).

Koroplet haritalarda iki kısıtlamadan söz edilebilir (Slocum ve ark., 2014):

- Sayım birimleri içindeki olası değişimleri göstermezler.
- Sayım birimlerinin sınırları keyfidir. Büyük olasılıkla ele alınan olgunun değişimi ile uyumlu değildir.

Koroplet harita tekniğinde verilerin standartlaştırılması önemli bir konudur. Standartlaştırma, verinin sayım birimlerinin alanlarına göre dengeli olmasını sağlamaktadır. Örneğin il nüfusları ham veridir (mutlak büyüklük). Koroplet harita yapım tekniği için uygun değildir. Bunun yerine nüfus km^2 olarak il alanına bölünerek standartlaştırılan veri (nüfus yoğunluğu, km^2'ye düşen insan sayısı) koroplet harita ile gösterilebilir (Slocum ver ark., 2014).

Koroplet harita yapımında önemli kararlardan birisi de verilerin sınıflandırılıp sınıflandırılmamasıdır. Sınıflandırılmamış bir koroplet haritanın avantajı veriyi daha doğru göstermesidir. Örneğin gri tonların kullanıldığı bir haritada her gri ton veri ile ilişkilidir. Buna karşın sınıflandırılmış haritanın (5 sınıf varsa 5 gri ton kullanılır) yapımı ve okunması daha kolaydır (Slocum ve ark., 2014). Türkiye illeri bazında nüfus yoğunluğu haritası düşünelim. 81 ilin her birine ait farklı bir değer vardır. Sınıflandırma yapılmaz ise 81 değişik gri ton kullanılması gerekir. Veri 5 sınıfa ayrılırsa 5 gri ton kullanılacaktır. 5 gri tonun lejant ile karşılaştırılması harita kullanıcısı açısından çok daha kolay olacaktır. Bu nedenle sınıflandırılmamış harita yapmak tematik kartografya açısından yanlış olmamakla beraber sınıflandırılmış haritalar tercih edilir. Aşağıda Şekil 3.1'de Konya iline ait olan tahıl vb. ekilen alanların haritası verilmiştir. Burada elde edilen veriler standartlaştırılmış ve sınıflandırılarak harita üretimi gerçekleştirilmiştir.



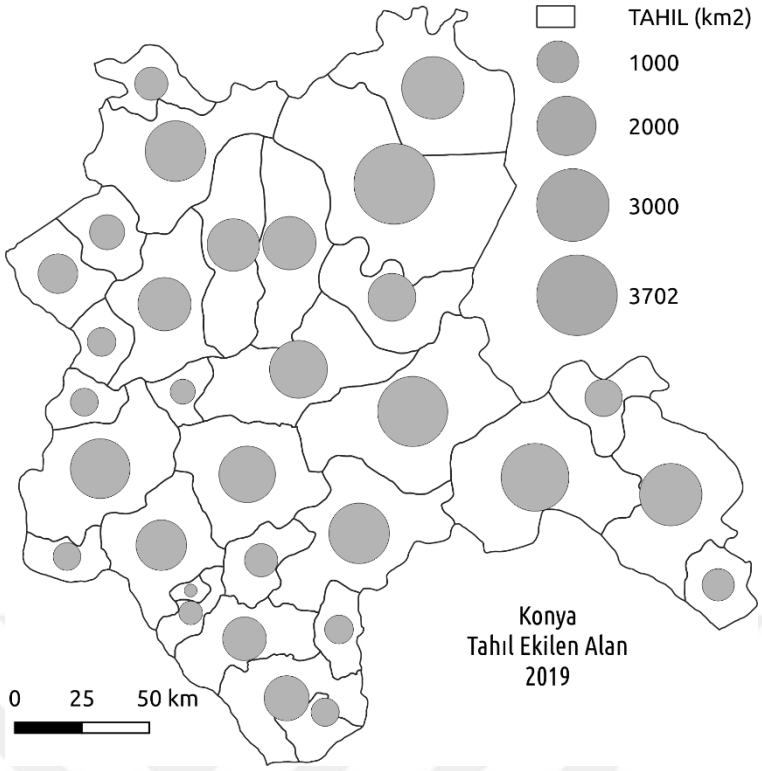
Şekil 3.1. Konya iline ait tahıl alanlarının toplam tarım alanına oranı koroplet haritası

3.2.2. İzoplet harita

İzaritmik (ya da kontur) haritası bilinen değerlere sahip olan nokta kümesinden enterpolasyonla elde edilen eşdeğerlik eğrileri (izolayn) ile oluşturulur. Örneğin meteoroloji istasyonlarında ölçülen sıcaklık değerlerinden yararlanılarak eş sıcaklık eğrileri, yüksekliği bilinen noktalardan eş yükseklik eğrileri oluşturulur. Bunlar eşdeğerlik eğrisi (izolayn) örnekleridir. İzoplet harita, izaritmik haritanın örneklem noktalarının sayım birimlerini temsil eden noktalar olduğu (alanların merkezleri ya da sentroitleri) biçimine denir (Slocum ve ark.,2014).

3.2.3. Orantılı işaret haritası

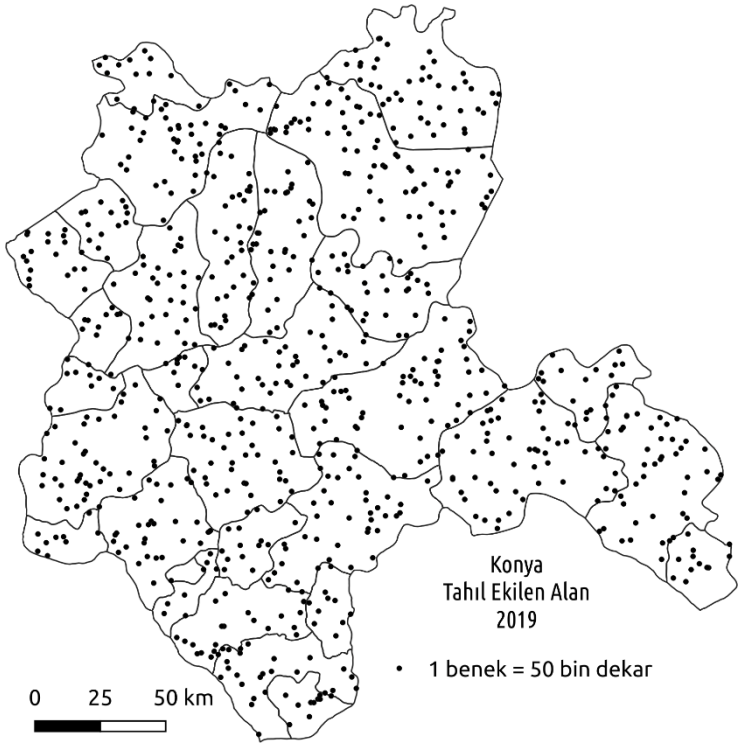
Orantılı işaret haritaları noktalara ait olan verilerle orantılı olarak işaret büyüklüklerinin değiştirilmesi ile oluşturulurlar. Harita üzerindeki noktalar, su kuyuları gibi gerçek konumlar olabildiği gibi, sayım birimlerinin orta noktaları (sentroit) gibi (alanı temsil eden noktalar) kavramsal konumlar olabilir (Slocum ve ark., 2014). Orantılı işaret hartası örneği Konya tahıl ekilen alan verisine uygulanırsa sayım birimi orta noktalarına işaretler yerleştirilir. Bu tarz haritalarda büyüklük görsel değişkeni kullanılır. Örnek harita Şekil 3.2'de verilmiştir.



Şekil 3.2. 2019 yılına ait Konya tahıl ekilen alan orantılı işaret haritası

3.2.4. Benek harita

Benek harita oluşturmak için bir beneğin aktarılmak istenen olgunun belli bir miktarına eşit kabul edilir. Beneklerin konumları belirlenirken olgunun ortaya çıkma olasılığına dikkat edilir. Olgu, gerçekte bir ya da birden çok alanı kapsar. Ancak harita



Şekil 3.3. 2019 yılı verilerini kapsayan Konya iline ait tahıl ekilen alan benek haritası

yapım tekniği nedeniyle olgu, nokta konumları ile temsil edilmektedir. Sayım birimlerinden elde edilen olgu alansal olduğu halde noktalarla (benekler ile) gösterilir. Benek haritaya Şekil 3.3'te örnek verilmiştir.

3.3. Çok Değişkenli Harita Yapımı

Kartograflar birden çok konuyu tematik haritalarda göstermeye ihtiyaç duyarlar. Örneğin bir ziraat mühendisi bir coğrafi bölge için o bölgede yetişen bitki türlerini, hava durumunu ve toprağın yapısal özelliklerini aynı anda göstermek isteyebilir. Birden çok konunun aynı anda gösterilmesine çok değişkenli haritalama denir. Eğer sadece iki özniteliğin gösterilmesi isteniyorsa iki değişkenli haritalamadan söz edilebilir (Slocum ve ark., 2014).

Çok değişkenli haritalamanın temel konusu farklı haritaların ayrı öznitelikler için gösterilip gösterilmemesi gerektiği (böyle bir durumda haritalar birbirleriyle karşılaştırılıyordur) ya da bütün özniteliklerin tek bir haritada görüntülenip görüntülenemeyeceğidir (böyle bir durumda haritalar birleştiriliyordur) (Slocum ve ark., 2014).

3.3.1. Haritaların karşılaştırılması

Her bir özniteliğin gösterilmesinde farklı haritaların kullanıldığı (haritaların birbirleriyle karşılaştırıldığı) iki değişkenli haritalama için kullanılan yaklaşımlar ele alınmaktadır. Haritaların birbirleriyle karşılaştırılması söz konusu olduğunda daha çok koroplet haritalar kullanılmaktadır (Slocum ve ark., 2014).

Harita karşılaştırma için daha iyi sonuçlar veren sınıflandırma yöntemleri ortalama-standart sapma, iç içe ortalamalar, yüzdelik dilimler ve eşit alanlardır (Slocum ve ark., 2014).

Haritaların karşılaştırılması için belirtilen sınıflandırma yöntemlerinin hangilerinin daha iyi sonuç verdiğine dair araştırmalar yapıp bir sonuca varan Lloyd ve Steinke (1976, 1977) haritaların görsel korelasyonunun (işaretleştirme için gri tonlu renklerin kullanıldığı varsayılarak) her bir haritada bulunan siyahlık değerlerinden etkilendiğini saptamıştır. Bu sonuca göre çeşitli öznitelik verisine sahip haritalar karşılaştırıldığında ve bu haritaların aynı istatistiksel korelasyona sahip olduğu düşünülürse hangi haritaların birbirine daha benzer olduğu konusunda bir sonuca varılabilir. Lloyd ve Steinke (1976, 1977), koroplet haritalar karşılaştırılırken eşit alanların kullanılmasını savunmuştur.

Olson (1972a) sınıflandırma yöntemlerinden hangilerinin iki öznitelik arasındaki korelasyonu koruduğunu saptamak için iki çalışma yapmıştır. Olson (1972b) ilk

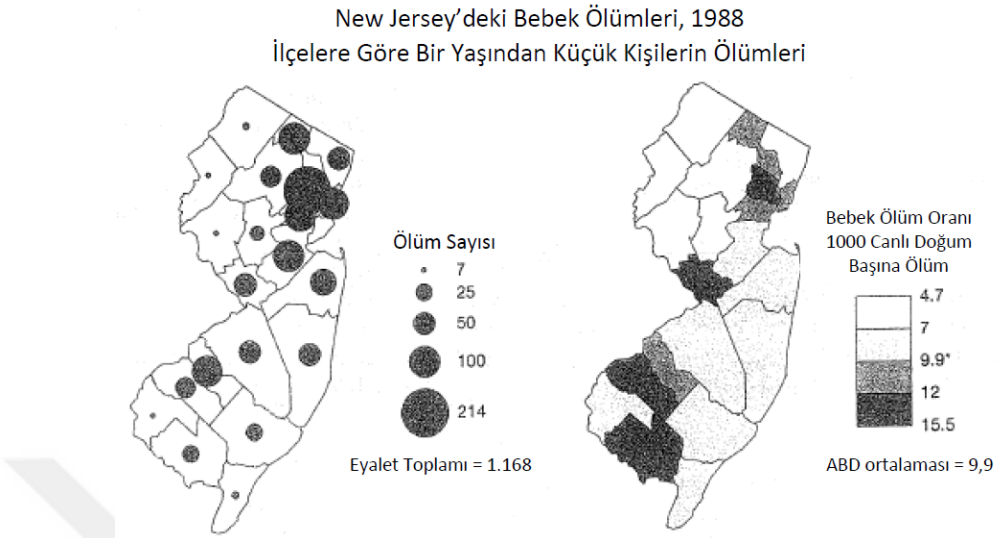
çalışmasında, kuramsal olarak normal dağılımlardan elde edilen 300 çift öznitelik üzerinde çalışmış ve yüzdelik dilimlerin öznitelikler arasındaki korelasyonu göstermede en iyi yöntem olduğunu belirlemiştir. İkinci çalışmasında ise Olson (1972a) (birincil olarak normal dağılmış) 300 çift gerçek öznitelik üzerinde çalışmış ve ortalama-standart sapma ve iç içe ortalamalar yöntemlerinin en etkili yöntemler olduğunu belirlemiştir.

Peterson (1979) ve Muller (1980) tarafında yapılan çeşitli çalışmalar, sınıflanmış ve sınıflanmamış koroplet haritaların insanlar tarafından görsel olarak karşılaştırılmasını ele almıştır. Yapılan çalışmalarda insanlar sınıflanmış ve sınıflanmamış ikili haritalar üzerinde benzer korelasyonlar olduğunu belirlemiştir. Yazarlar bu çalışmalarında koroplet haritalar yapılırken verilerin sınıflandırılması gerekliliğini belirtmiştir.

Brewer ve Pickle (2002)'ın yaptığı çalışmada ise harita karşılaştırma söz konusu olduğunda kullanılacak yöntemler arasından yüzdelik dilim yönteminin daha uygun olduğunu belirtmektedir. Yazarlar, çeşitli araştırmalar yaparak sınıflandırma yöntemlerinin harita karşılaştırma işlerinde karşılaştırılmasına istinaden yüzdelik dilim yönteminin etkili yöntem olduğu sonucuna varmıştır. Yazarlar yaptığı çalışmalar ile yüzdelik dilim yönteminin harita karşılaştırmayı kolaylaştırdığını ve harita okumaya yardımcı olduğu sonucuna varmıştır.

Haritaları karşılaştırırken üretim yöntemini sabit tutmak her ne kadar yaygın bir şekilde kullanılıyor olsa da aynı türden haritalar kullanmak yerine (aynı türden haritaları düşünürsek bunlara örnek olarak iki koroplet harita veya iki izoplet haritanın karşılaştırılması olabilir.) iki farklı türde tematik haritanın karşılaştırılması ile de kullanışlı bilgilerin elde edilmesi mümkündür. Bu şekilde bir yöntemi kullanmak, ham toplamlar ve standartlaştırılmış veriler karşılaştırılmak istendiğinde iyi bir yöntem olabilir. Şekil 3.4'de verilen New Jersey'deki bebek ölümlerinin ham sayısına ait orantısal bir işaret haritası, bebek ölüm oranının koroplet haritasıyla karşılaştırılmaktadır. Burada orantılı işaret haritası, bebek ölümü sorununun eyaletin kuzeydoğu bölgesinde olduğunu belirtmektedir. Ancak böyle bir harita tek başına yeterli olmamaktadır. Çünkü büyüyüp küçülen işaretler nüfusun bir fonksiyonudur. Daha fazla insan bulunan ilçelerde daha fazla bebek ölümü olma ihtimali fazladır. Buna karşın koroplet harita ham ölüm verisini, ölüm sayısını canlı doğum sayısına oranla göz önüne alarak standartlaştırmakta ve sorunun eyaletin üç farklı alanında olduğunu göstermektedir (en yüksek sınıftaki bölgeler). Ancak koroplet haritadaki yüksek bir oran, az sayıda ölüm varsa anlamlı olmayabilir. Böyle bir durumda iki harita birlikte gösterilebilir ve iletilmek istenen veriler ortaya çıkar. Şekil 3.4'de bulunan haritada kuzeyde bulunan yüksek oranlı alan en sorunlu

alan olarak göze çarpmaktadır. Çünkü ölüm sayısının fazla olduğu bölge ile çakışmaktadır.



Şekil 3.4. Orantısal işaret ve koroplet haritalamanın karşılaştırılması (Monmonier 1993)

3.3.2. İki özniteliğin aynı haritada birleştirilmesi

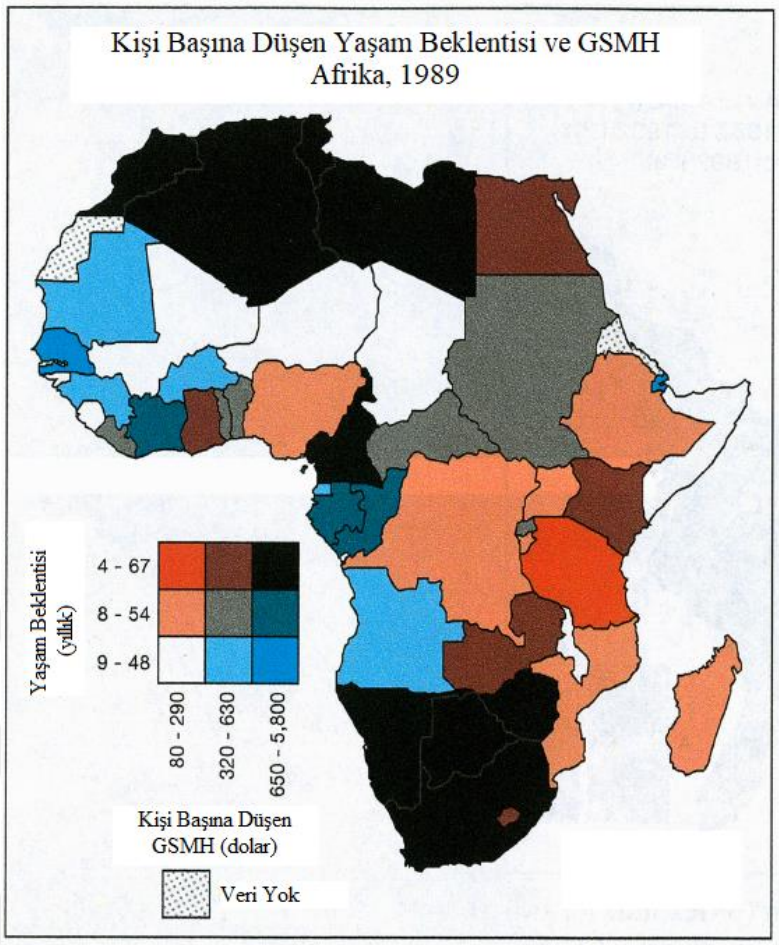
İki değişkenli haritalamaya yönelik olarak iki özniteliğin aynı haritada birleştirildiği yaklaşımlar bu kapsamda ele alınır. Bu yöntem yaygın olarak koroplet haritalarda kullanılmaktadır (Slocum ve ark., 2014).

ABD Nüfus İdaresi 1970'li yıllarda iki renkli koroplet haritanın birleştirilmesine yönelik bir yöntem olan iki değişkenli koroplet haritayı geliştirmiştir. ABD Nüfus İdaresi aynı zamanda iki değişkenli koroplet haritalara yönelik renk şemasının yapımında dikkate alınması gereken faktörlerin bir listesini oluşturmuştur (Eyton, 1984).

1. Renkler birbirlerinden ayırt edilebilir olmalıdır.
2. Renklerin birbiri arasındaki geçişi orantılı bir şekilde sağlanmalı ve yumuşak bir geçişe sahip olmalıdır.
3. Her bir dağılım için görsel olarak kategorileri birbirinden ayırt edilebilir veya dengeli olmalıdır ve bir bütün olarak iki dağılım, birbirinden ayrılabilir olmalıdır.
4. Lejanttaki renk düzeni belirli kurallara göre olmalıdır.
5. Haritadaki renk tonları açıktan koyuya doğru olmalıdır. Böylece aktarılmak istenen verideki sayısal değerler arasındaki değişim daha iyi bir şekilde gösterilebilir.
6. Uç değerler ana renklerle gösterilmelidir.

7. Pozitif ve negatif değerleri göstermek için ana köşegenin üstünde ve altındaki renk tonlarında bir düzen olmalıdır.
8. İlişkileri göstermek için pozitif köşegenler (sol alttan sağ üste uzanan) ve negatif köşegenler (sol üstten sağ alta uzanan) görsel olarak bir düzen içerisinde olmalıdır.
9. Renk şeması oluşturulurken lejantta bulunan çok sayıdaki renk dikkate alınmalı ve ayırt edilmelerindeki güçlük olabileceği hususuna dikkat edilmelidir.
10. Renk şeması, harita yapımında kullanılacak istatistiksel veriyi olabildiğince iyi yansıtacak şekilde açıklanmalıdır.
11. İki ayrı haritadaki renklerin birleşimi, dahil edilen belirli renklerin birleşimleri gibi görünmelidir.
12. Kullanılacak kategorilerin sayısı, okuyucunun anlayabileceği sınırı aşmamalıdır. 3 x 3 formundaki bir lejant hem teknik hem de görsel olarak 4 x 4 formundaki bir düzenden daha basittir ve okuyucuya daha çok bilgi aktarabilir.
13. Lejantta dikdörtgen formda bir düzen yerine diğer alternatifler dikkate alınmalıdır. Dikdörtgen form, harita yorumlama sorunlarına yol açar ve istatistiksel verinin mesajını etkiler.

ABD Nüfus İdaresi'nin yaklaşımına alternatif olarak Eyton (1984), ikişerli olarak bir araya gelerek oluşan renkleri (tamamlayıcı renkler) temel alarak bir iki değişkenli yöntem geliştirmiştir (Munsell renk çemberindeki karşılıklı olan renkler birbirinin tamamlayıcısıdır.). Eyton (1984), kullanılan renkleri tamamlayıcı renkler olan kırmızı ve cyan (gök mavisi) olarak seçmiştir. Çıkarmalı renkleri kullanarak (CMY), magenta (M) ve sarı (Y) rengi birleştirerek kırmızı rengi oluşturmuştur. Çıkarmalı renk karışımı yönteminde siyah renk (K) elde edilemediği için Eyton (1984), siyah istenen alanlar için siyah (K) kullanmıştır (Şekil 3.5).

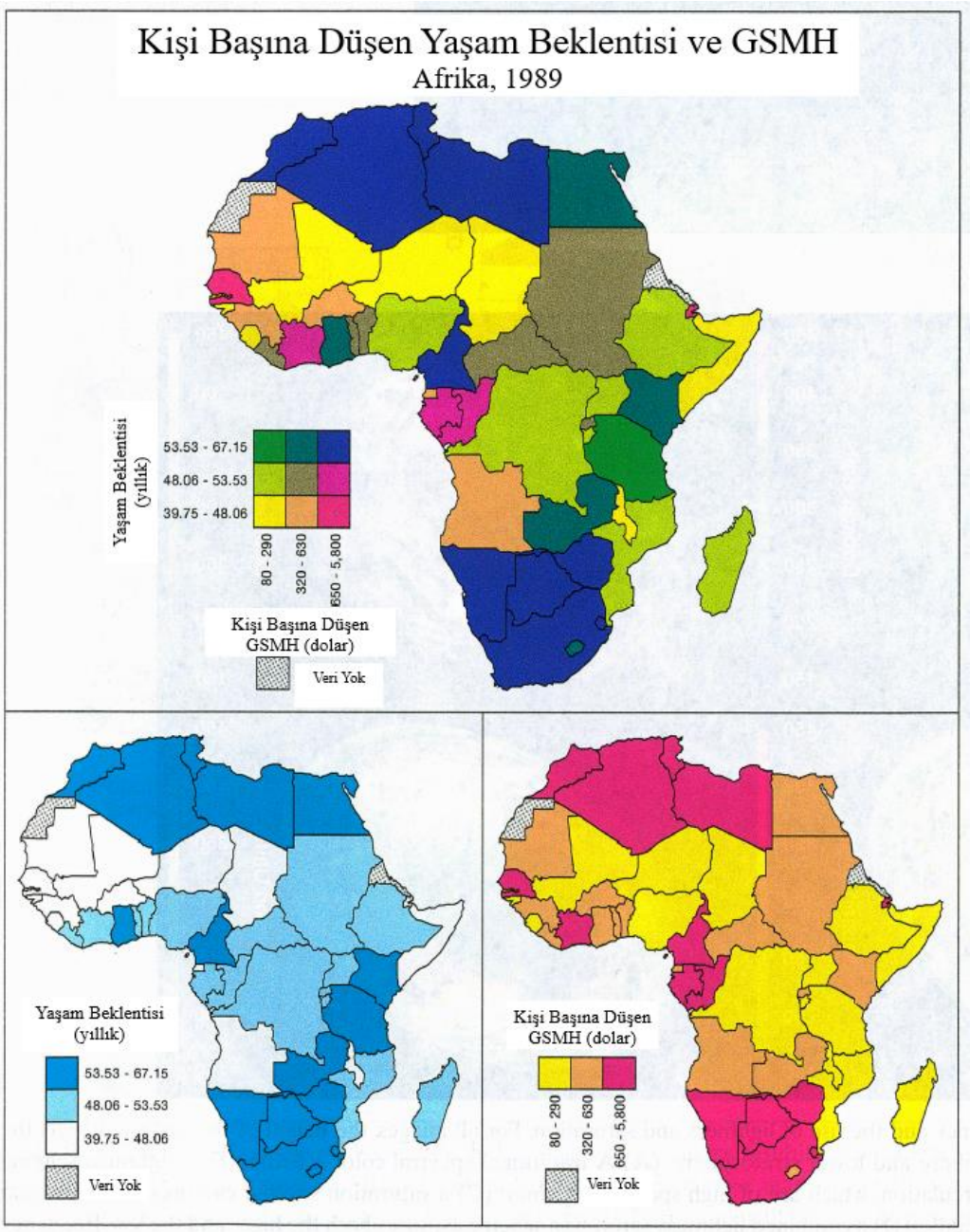


Şekil 3.5. Kırmızı ve Cyan (gök mavisi) tamamlayıcı renklerine sahip koroplet iki değişkenli bir harita (Eyton, 1984)

Eyton (1984), ABD Nüfus İdaresi'nin oluşturduğu iki değişkenli koroplet haritalarına yönelik renk şemalarının geliştirilmesinde dikkate alınması gereken faktörlerin çoğunun kendisinin oluşturduğu tamamlayıcı yöntemle açıklandığını ve kullanıcıların haritayı, ABD Nüfus İdaresi renklerine dayanan haritadan daha kolay anladığını öne sürmüştür. ABD Nüfus İdaresi'nin renk şemasında lejantın dikkatli bir şekilde incelenmesi gerekirken Eyton (1984)'ın oluşturduğu şema mantıksal olarak daha düzenli gösterilmektedir. Bu, haritadaki desenlerin daha kolay kavranmasını sağlamaktadır. Örneğin, Nüfus İdaresi'nin renk şemasını kullanan karşılık değerlere kıyasla kırmızımsı kahverengi değerlerin sağladığı kolaylıklar Eyton (1984)'ın şemasını kullanarak bulunabilir (Slocum ve ark., 2014).

İki değişkenli koroplet haritalar teknik olarak bir başarı gibi görünse de birçok araştırmacıdan ayrı ayrı dağılımlar hakkındaki bilginin iletilmesi ya da aralarındaki korelasyonun iletilmesindeki varsayılan başarısızlıklardan dolayı eleştiriler almıştır. Yapılan eleştirilere yanıt olarak Olson (1981), ABD Nüfus İdaresi'nin kullandıklarına benzer renk şemaları kullanarak deneysel bir çalışma yürütmüştür. Daha önceden yapılan

eleştirilere rağmen Olson (1981), iki değişkenli haritaların "homojen değer kombinasyonlu bölgeler hakkında bilgiler" sağladığını ve kullanıcıların elde edilen bu haritaların aktardığı bilgileri kolayca alabildiklerini, aksine bunlara karşı olumlu bir tavır takındıklarını bulmuştur. Ancak Olson (1981) iki değişkenli haritaları anlamada açıkça oluşturulmuş bir lejantın önemli olduğunun, hem iki değişkenli hem de ayrı ayrı oluşturulmuş haritaların gösterilmesi gerektiğinin ve bu haritalardan çıkarılabilecek bilginin türlerini açıklayan bir nota yer verilebileceğinin altını çizmiştir. Olson (1981)'ın bahsettiği gibi olan hem iki değişkenli olarak üretilen hem de ayrı ayrı verileri aktaran haritalar Şekil 3.6'da verilmiştir.



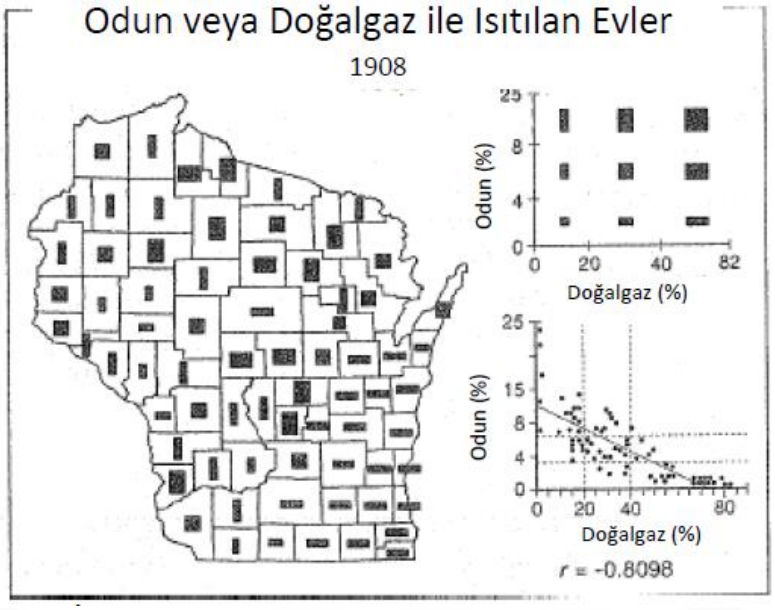
Şekil 3.6. İki değişkenli koroplet harita ve bu haritaya ait tek değişkenli haritalar (Olson, 1981)

Şekil 3.6'da verilen haritayı oluşturan renk şeması Olson (1981)'dan alınmıştır ve 1970'lerde ABD Nüfus İdaresi'nin kullandığı popüler renk şemalarına benzer özelliklere sahiptir. Her haritanın verilerinin sınıflandırılmasında kuantil sınıflandırma yöntemi kullanılmıştır.

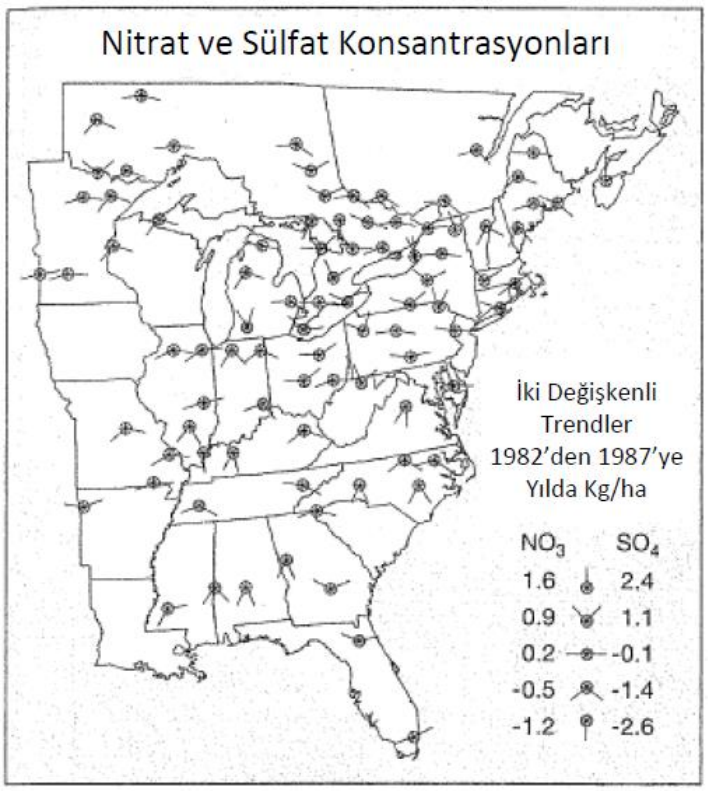
Niceliksel bir sınıflandırma olan kuantil sınıflandırmasında oluşturulan tüm sınıflar eşit sayıda öznitelik içerecek şekilde oluşturulur. Bu sınıflandırma yöntemi doğrusal olarak dağıtılan verilere oldukça uygundur. Kuantil sınıflandırma yönteminde her sınıfa aynı miktarda veri atanır. Bu sayede boş sınıf veya çok az ve çok fazla değer içeren sınıflar oluşmamaktadır (URL7).

İki değişkenli koroplet haritanın alternatiflerinden birisi de iki değişkenli nokta işaretli haritalardır. Bir dikdörtgenin eninin ve yüksekliğinin haritanın özniteliklerinin her birine göre orantılandığı, iki değişkenli nokta işaretinin bir şekli olan dikdörtgenel nokta işareti Şekil 3.7'de gösterilmektedir. Wilhelm (1983) ve Hartnett (1987) iki değişkenli nokta işareti yöntemini kesişen çizgili işarete (gösterilecek olan istatistiksel verinin değer olarak artması ya da azalmasına göre orantılı olarak kesişen çizgilerin sıklığının artması ya da azalması) bir alternatif olarak önermişlerdir. Hartnett (1987), dikdörtgen biçimli nokta işaretini incelemenin kesişen çizgilerden oluşan küçük kutuları incelemekten daha kolay olduğunu ve üretilen haritanın göze hitap ettiğini öne sürmüştür. Nokta işaretleri, alanlarla ilişkili olmak yerine nokta konumlarıyla daha ilişkili olduğundan iki değişkenli nokta işaretleri, nokta konumunun önem arz ettiği verilerin haritalanmasında daha uygundur (birbiriyle bağlantılı iki durumun konuma bağlı sayıları önemli olduğunda kullanılabilir).

İki değişkenli nokta işaretinin diğer bir şekli ise iki değişkenli ışın glifidir. Carr ve ark. (1992) bu formu doğu Birleşik Devletleri ve Kanada'daki nitrat (NO_3) ve sülfat (SO_4) konsantrasyonları arasındaki ilişkiyi incelemek için kullanmıştır. Şekil 3.8'de sol tarafa doğru yönelmiş ışınlar nitrat konsantrasyonunu temsil etmekteyken sağ tarafa doğru yönelmiş olan ışınlar sülfat konsantrasyonunu ifade etmektedir. Buradaki ışınlar ortadaki işaretin üst kısmına doğru çıktıkça yüksek değerleri ifade ederken ışınlar işaretin alt kısmına doğru indikçe düşük değerleri ifade eder. Küçük işaretlerin kısıtlı alanlara sıkıştırılabiliyor olması ışın glifi işaretlerinin avantajları arasındadır. Ancak bu işaretlerle temsil edilen veriler, dikdörtgen işaretle temsil edilen verilere göre daha zor yorumlanabilmektedir (Slocum ve ark., 2014).



Şekil 3.7. Özniteliklerin bir dikdörtgenel nokta işaretinin eni ve yüksekliğiyle gösterildiği iki değişkenli harita (Hartnett, 1987)



Şekil 3.8. Işın glifi işaretine yönelik iki değişkenli bir harita. Sağa ve sola işaret eden ışınlar sırasıyla sülfat ve nitrat konsantrasyonlarını temsil etmektedir (Carr ve ark., 1992)

İki değişkenli haritalamada yaygın yaklaşımlardan biri de orantısal ve koroplet işaretleri, ham toplam veriler için kullanılan orantısal işaretin boyutuyla ve standartlaştırılmış veriler için kullanılmakta olan koroplet harita ile birleştirmektir. Örnek olarak, burada bahsedilen uygulama Şekil 3.4'de iki harita halinde sunulmuş olan bebek ölüm verilerini tek bir haritada göstermek için kullanılabilir. Yani hem boyut hem de gölgeleme tek bir özniteliği temsil etmek için kullanılabilir. Burada en önemli noktalardan biri de lejanttır. Lejant, bir haritanın tek değişkeni mi, iki değişkeni mi ya da daha fazla değişkeni mi ifade ettiğini anlamak ve yorumlamak için oldukça önemlidir.

3.3.3. Çok değişkenli nokta haritalar

Tek değişkenli nokta haritalama, her bir özniteliğin haritada gösterilmesi için belirli bir işaret şekli veya rengi kullanılarak çok değişkenli bir harita oluşturulmak üzere genişletilebilir (Slocum ve ark., 2014).

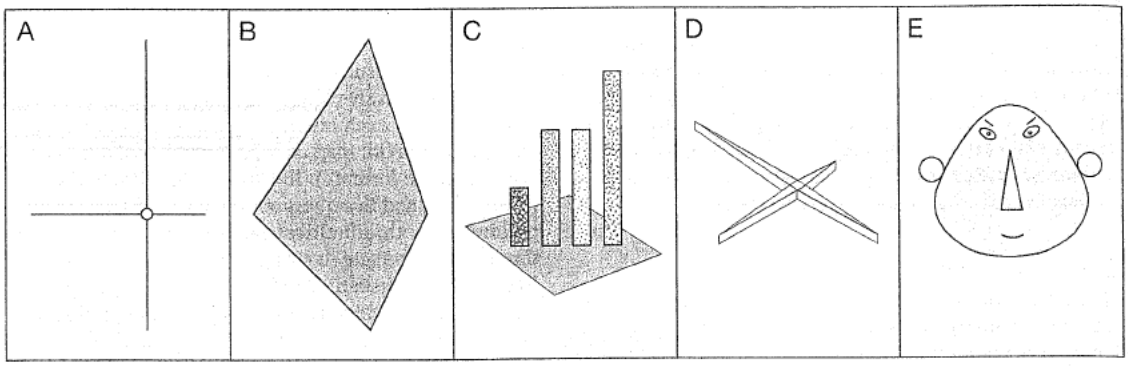
Pointilizm, 19.yy'da ressamlar tarafından, seçili renklerden oluşturulmuş çok küçük noktaları bakan kişiler tarafından birleştirileceği varsayılarak çeşitli renk karışımlarının oluşturulabilmesi için kullanılan bir tekniktir. Jenks (1953) bu ilkeyi farklı renkteki noktaların çeşitli mahsulleri temsil etmesi için kullanmıştır. Haritaya bakan kişilerin renkleri görsel bir şekilde birleştirerek karışımlar oluşturabildiğini ve ekim uygulamalarının geçişsel doğasına dair daha gerçekçi bir görüntü elde ettiğini öne sürmüştür. Ayrıca Jenks (1953), eğer noktalar yeteri kadar büyük olursa haritanın seçili alanlarındaki ayrı mahsullerin konumuyla ilgili ayrıntı verebileceği konusuna dikkat çekmiştir. Jenks (1953) noktaların renklerinin seçimi konusunda faydalı önerilerde bulunmuştur:

1. Temsil edilen mahsulün rengi gerçekteki renklerine yakın olmalıdır.
2. Tütün ya da sebze gibi yüksek değere sahip daha az yer kaplayan mahsuller, daha geniş ve daha yaygın olarak yetiştirilen mahsullerden daha yoğun tonda olmalıdır.
3. Geniş alanlardaki mahsullerin karakterini değiştirmeye meyilli olan yer fıstığı, soya fasulyesi gibi seçili küçük mahsuller orta derecede yüksek yoğunluklu renklerde olmalıdır.

3.3.4. Çok değişkenli nokta işaretli haritalar

Çok değişkenli verilerin bir işaret ile gösterilmesiyle çok değişkenli nokta işaretli haritalar üretilir (Slocum ve ark., 2014).

Konu olarak birbirleriyle ilgili veya ilgili olmayan iki belirgin öznitelik formu çok değişkenli nokta işaretleri kullanılarak haritalanmaktadır. İlgili öznitelikler aynı birimlerle ölçülmektedir ve daha büyük bir bütünün parçasıdır. Örnek olarak bir nüfustaki çeşitli etnik gruplarının yüzdeleri verilebilir. Bu tür öznitelikler daha çok istatistiksel olarak pasta dilimi şeklinde gösterilirler. Gösterilen grafikte bir daire her özniteliğin oranını temsil eden bölümlere ayrılmaktadır. İlgili olmayan öznitelikler ise birbirlerine benzer olmayan birimlerle ölçülmektedir. Dolayısıyla büyük bir bütünün parçası değildir. Örneğin, yüzde olarak kentsel gelir ve gelirin medyanı olarak düşünülebilir. Birbirleriyle ilgili olmayan özniteliklerin temsil edilmesi için kullanılan çok değişkenli işaretlere



Şekil 3.9. Çok değişkenli nokta işaretlerine örnekler. (A) bir çok değişkenli ışın glifi ya da yıldız: ışın uzunlukları, özniteliğin değerleriyle orantılıdır; (B) bir poligonal glif veya kar tanesi: A'da gösterilen ışınların uç noktalarına bir poligon bağlanır; (C) üç boyutlu çubuklar: çubukların uzunluğu, özniteliklerin büyüklüğüyle orantılıdır; (D) veri çivileri: çivinin sivri uçları, her bir özniteliğin büyüklüğüyle orantılıdır; (E) Chernoff yüzleri: ayrı ayrı yüz özellikleri ayrı özniteliklerle ilişkilendirilir (örneğin gözlerin boyutu) (Slocum ve ark.,2014)

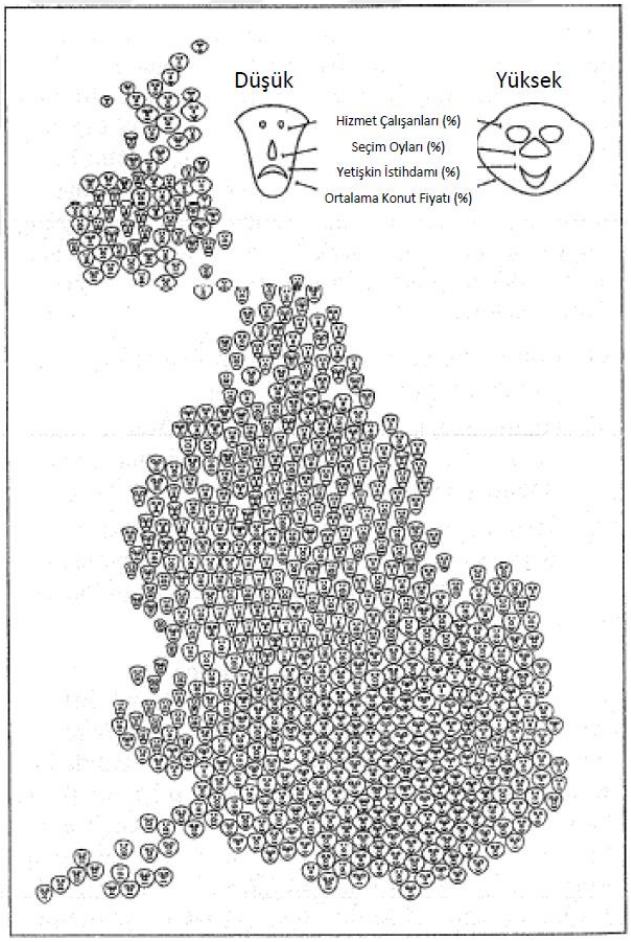
genellikle glif adı verilir (Slocum ve ark., 2014). Bunlardan birkaç tanesi Şekil 3.9'da gösterilmiştir.

Çok değişkenli ışın glifi ya da yıldızdaki (Şekil 3.9A) ışın uzunlukları aktarılmak istenen özniteliklerin değerleriyle orantılı olarak ışınların bir iç çemberden uzatılmasıyla oluşturulur. Işın glifleri, Anderson (1960) tarafından ilk kez oluşturulduğunda ışınlar çemberin sadece üst kısımlarından uzatılmaktaydı fakat sonraları ışınların tüm yönlere uzatılması daha yaygın hale gelmiştir. Çok değişkenli ışın gliflerini meydana getiren ışınlar eğer birbirlerine bağlıysa, poligonal glif veya kar tanesi meydana gelir (Şekil 3.9B).

Urbana-Champaign'deki Illinois Üniversitesi Ulusal Süper Bilgisayar Uygulamaları Merkezinden Cox (1990) çeşitli çok değişkenli yeni nokta işaretleri geliştirmiştir. Bunlardan biri de üç boyutlu çubuklardır (Şekil 3.9C). Çubukların yüksekliği, çeşitli özniteliklerin büyüklüğüyle doğru orantılıdır. Cox (1990)'un uygulamasında çubuklar şekilde (Şekil 3.9C) görüldüğü gibi farklı dokuların aksine farklı

renklere sahiptir. Diğer bir teknik olan veri çivisinde Ellson (1990) üçgensel olan sivri uçlar merkezi kare olan alandan uzatılmaktadır. Her bir özniteliğin büyüklüğüne orantılı şekildedir (Şekil 3.9D). Eğer çivilerin sivri uçları farklı renklerde gösterilirse çubuklarda olduğu gibi daha kolay ayırt edilebilmektedir. Veri çivilerinin üç boyutlu olmasının avantajlarından birisi de farklı açılardan bakıldığında kolayca gösterilebilmesidir.

Daha çok ilgi çeken çok değişkenli nokta işaretlerinden olan Chernoff yüzünde (Şekil 3.9E) belirgin olan yüz özellikleri çeşitli özniteliklerle ilişkilendirilmiştir. Örneğin Chernoff yüzündeki yanakların tombulluğu bir özniteliği temsil ederken gözlerin boyutu ise diğer bir özniteliği temsil etmektedir. Dorling (1993)'in kartograflar ile yaptığı çalışmasından alınan Şekil 3.10, Chernoff yüzlerini göstermektedir. Burada yüzün eni, ortalama konut fiyatını (tombul yanaklar daha pahalı konutları göstermektedir); ağız stili, yetişkin istihdamı yüzdesini (yüksek istihdam için gülücük); burun boyutu, ortalama oyu (daha büyük burun daha yüksek yüzdeyi göstermektedir); göz boyutu ve konumu, hizmet



Şekil 3.10. Çok değişkenli verileri göstermek için Chernoff yüzlerininin kullanıldığı bir kartogram (Dorling, 1993)

sektöründe istihdam edilenlerin yüzdesini (buruna yakın daha büyük gözler, hizmet

sektöründeki istihdam edilenlerin yüksek yüzdesini) ve yüzün toplam alanı, bir seçim bölgesindeki seçmenlerin sayısını temsil etmektedir.

3.4. Veri Madenciliği

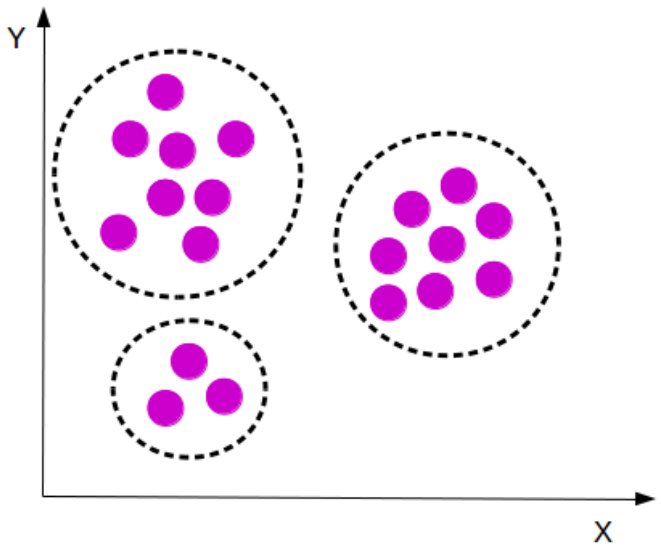
Veri madenciliği, verinin analiz edilmesi ve o veriden değer yaratmaya yönelik çalışmaları ifade eden bilgi keşfi, veriden doğrudan görülemeyen ve faydalı olacak bilgilerin elde edilmesi olarak tanımlanmaktadır (Frawley ve ark., 1992).

Veri madenciliği, büyük hacimli verilerin içinden yararlı bilginin çıkarımı, bilgiyi madenleme işidir. Büyük hacimli veri yığınları içinden gelecekle ilgili tahminde bulunabilmemizi sağlayan gerekli bağıntıların bilgisayar yazılımları kullanılarak aranması olarak da tanımlanabilir (URL4). Günümüzde veri madenciliği ile birlikte büyük hacimli verilerden işe yarar bilgi çıkarımı oldukça kolay hale gelmiştir. Alışılagelmişin dışında veri yığınları düşünüldüğünde veri madenciliğinin önemi de ortaya çıkmaktadır. Bu veri yığınları içerisinden istenilen bilginin elde edilmesinde harcanan emek, zaman, maliyet gibi unsurlar oldukça azalmıştır.

Elde edilen bilginin oldukça önemli olduğu günümüzde yeni değişimlerin temelini veri, bilgi ve enformasyon kavramları oluşturmaktadır. Sistem politikalarının ve stratejik kararların temelinde veri ve verilerden meydana gelen bilgiler vardır. Bu kararların amaca uygun olarak verilebilmesi için güvenilir, güncel, doğru ve zamanında ulaşılan bilgiye ihtiyaç vardır. Bu kavramalar aslında hayatımızın her alanında karşımıza çıkmaktadır. Sağlık, alışveriş, otomasyon verileri gibi birçok verinin veri tabanlarında saklanmasında da büyük zorluklar karşımıza çıkmaktadır. Verilerin depolanması ve saklanması başlı başına bir problem haline gelmiş ve konu üzerine araştırma yapanlar, oluşan bu problemlere veri tabanları ve dosya sistemlerindeki gelişmelerle birlikte çözüm üretmek için çalışmalar yapmışlardır. Araştırmacılar oldukça büyük bir hacme sahip veri tabanlarının herhangi bir araç kullanmadan analiz etmek ve sonucunda karar verme aşamasında kullanımın imkânsız olduğunu belirlemişler ve bu aşamada veri tabanından bilgi keşfi kavramı ortaya çıkmıştır. Veri tabanında bilgi keşfi; veriden yararlı bilgi çıkarma sürecidir (Özdemir ve ark., 2006; Skillicorn, 2009).

Veri madenciliğinde kullanılan önemli yöntemler arasında “veri sınıflandırma” yöntemleri gelmektedir. Veri sınıflandırma ile birlikte belirli hacme sahip verilerden istenilen çıkarım yapılabilir hale gelmekte ve böylesi istatistiksel veri yığınları ile uğraşan araştırmacılar için oldukça yol gösterici sonuçlar ortaya çıkmaktadır.

Veri madenciliği kavramı düşünüldüğünde birçok veri sınıflandırma yöntemi vardır. Bu yöntemlerden birisi de "kümeleme analizi" dir. Kümeleme analizi bir denetimsiz öğrenme metodudur. Denetimsiz öğrenme metodunda çalışma yapılacak istatistiksel verilerin niteliği hakkında bilgiye ihtiyaç yoktur. Uygulama sonucunda ayrılan sınıflara göre bir değerlendirme yapılır. Elde edilen veriler herhangi bir konuya ait olabilir ve yapılan çalışma açısından konu önemli değildir. Çıkan ürün sonucunda kullanıcı bir çıkarımda bulunur. Aşağıdaki şekil 3.11'de iki boyutlu verilerin belirli bir kümeleme analizi algoritmasıyla üç sınıflara ayrılmış şekli gösterilmiştir. Algoritma ile yapılan uygulama sonucunda kullanıcı yaptığı çalışma konusuna göre rahatlıkla bir çıkarımda bulunabilir. Bu durum çok değişkenli analiz yöntemlerinin bir sonucu olarak karşımıza çıkar.



Şekil 3.11. Kümeleme Analizi sonucunda sınıflara ayrılmış veri grupları

3.5. Kümeleme Analizi

Bir veri kümesindeki istenen özelliklerin birbirleriyle olan uyumlarına göre gruplandırılması işlemine kümeleme analizi denilmektedir. Oluşan gruplardan her birine "küme" denilirken kümeleme analizi işlemine ise kısaca "kümeleme" denilir. Yapılan kümeleme analizi işleminde her bir kümenin içindeki öğelerin birbirleriyle olan benzerliklerinin fazla olması gerekirken ortaya çıkan kümeler arasındaki benzerliğin ise az olması beklenir (Dinçer, 2006).

Veri toplama birimlerince toplanmış istatistiksel verilere kümeleme analizi uygulanarak elde edilmiş bu verilerin birbirine benzeyen özellikleri belirlenip elde edilen sonuçlar bir tematik harita ile kullanıcılara sunulabilir. Böylece kümeleme analizi ile elde

edilmiş kümelerin ilgili verilere ihtiyacı olan bireyler tarafından daha iyi anlaşılması sağlanır (Bildirici ve Afacan, 2018).

Bireylerin sınıflandırılmasını kapsayan yöntemler, literatürde Q analiz yöntemleri olarak bilinir ve kümeleme analizi de bu yöntemlere dahildir. Kümeleme analizinin özünde, verilerin bazı benzerlik ölçütlerine göre gruplara ayrılması ile birlikte işe yarar, faydalı bilgilerin sağlanmasını vardır. Kümeleme analizi, çok değişkenli istatistiksel analiz türüdür (Tatlıdil, 2002).

Kümeleme analizi teknik olarak ikiye ayrılabilir. Bunlar hiyerarşik ve hiyerarşik olmayan yöntemlerdir. Her bir gözlemin farklı kümeler olarak düşünüldüğü ve daha sonra kümelerin kademeli olarak birleştirildiği yöntemlere hiyerarşik yöntemler denir. Hiyerarşik olmayan yöntemler ise belirli sayıda küme ve ilişkili öğenin varsayıldığı ve ardından gözlemlerin kümeler arasında taşındığı ve bu sayede sınıflandırmanın geliştirilmeye çalışıldığı yöntemlerdir (Slocum ve ark., 2014). Hiyerarşik olmayan yöntemlerin kullanımında küme sayısının kullanıcı tarafından belirlenmesi beklenmektedir. Analiz, bu inisiyatif üzerinden devam eder.

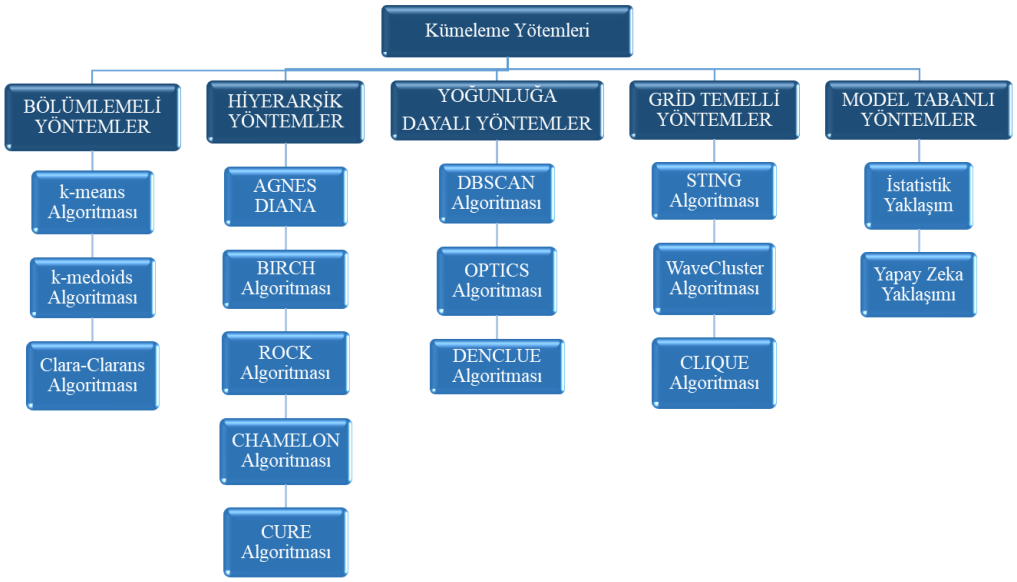
Aşağıda kümeleme analizi yöntemlerinin bazı genel özellikleri verilmiştir (Han ve ark., 2012):

- **Ölçeklenebilir.** Kümeleme algoritmaları, küçük veri kümesinde yani daha az veri içeren kümelerde daha iyi çalışırken daha fazla veri içeren büyük veri kümelerinde standart şekliyle çalışmayabilir. Böylesi durumlarda mutlaka ölçeklendirme algoritmaları kullanılmalıdır.
- **Farklı nesne tiplerine uyum sağlayabilmelidir.** Birçok kümeleme analizi algoritması sayısal veri tipleriyle çalışması için tasarlanmıştır. Ancak ikili (binary), kategorik ve sıralı veri tipleriyle de çalışabilmektedir.
- **Kümelerin belirli bir şekli olmayabilir.** Kümeleme algoritmaları, kümeleri belirlerken Öklid veya Manhattan mesafe ölçütlerinden yararlanır. Kümeleme analizi algoritmaları mesafe ölçütleri ile birlikte birbirine benzeyen boyut ve yoğunluktaki bütünsel kümeleri bulabilirler. Fakat ortaya çıkan kümeler yine de farklı şekillerde olabilirler.
- **Gürültülü verilerin kullanılabilmesi.** Birçok veri grubu bilinmeyen, hatalı ya da eksik veriler içerir. Kümeleme algoritmaları böyle veriler nedeniyle kalitesiz kümeler oluşmasına neden olabilirler.

- **Yüksek boyutlu veri gruplarının kümelenmesi.** Bir veri kümesi ya da veri grubu birçok boyut ya da nitelik içerebilir. Kümeleme algoritmalarının birçoğu iki ya da üç boyut içeren veri grupları için daha iyi sonuçlar vermektedir.
- **Veri kümelerinde birtakım kısıtlamalar olabilir.** Gerçek dünyadaki uygulamaların birtakım kısıtlamalar ile kümeleme işlemi yapması gerekebilir. Örneğin bir şehirde yeni kurulacak ATM'lerin konumlarının araştırılması için bir göreviniz olsun. ATM'lerin konumlarına karar vermek için belirli durumları göz önüne almak gerekir. Bunlar şehirden geçen nehirler, otoyol ağları, oluşturulacak kümeler için müşterilerin türü, sayısı gibi kısıtlamaların göz önünde bulundurulması gerekir. Buna göre hane halkı kümelenir. Bu tarz bir işlemde dikkat edilmesi gereken nokta istenilen özellikteki kümelerin belirtilen kısıtlamaları karşılamasıdır.

Kümeleme algoritmaları ölçeklenebilir yapılara sahip, farklı nitelik türleri bulunan, gürültülü verilerin analiz edilebildiği, çeşitli sınırlamaların uygulanabileceği yapılara sahiptirler. Ayrıca kullanılacak kümeleme yöntemleri istenen küme sayısına göre farklılık gösterebilir. Kümelerin farklı özelliklere sahip olmasına, istenilen benzerlik özelliklerine ve alt-uzay kümelerinin istenip istenmediğine göre kullanılacak algoritma belirlenir (Han ve ark., 2012).

Kümeleme analizi için kullanılan algoritmalar Şekil 3.12'de verilmiştir.



Şekil 3.12. Kümeleme Yöntemleri

3.5.1. Kümeleme analizinde veri türleri

Kümeleme analizinde veriler yapı olarak matris şeklindedir. Kümeleme analizi algoritmalarında kullanılan matrisler iki ana başlıkta incelenebilir (Han ve ark., 2012).

1. **Veri Matrisi (Data Matrix):** Bu veri yapısında *n* tane nesne, *p* tane değişkenden bahsedilebilir. Her bir satır nesneye karşılık gelmektedir. Sütunlar ise nesneye ait öznitelik bilgilerini kapsamaktadır. Örneğin nesnelerin bir ülkedeki şehirleri temsil ettiğini düşünürsek değişkenler ise şehrin nüfusu, yüz ölçümü ve orada yaşayan insanları eğitim durumunu temsil edebilir. Veri matrisi, aşağıdaki denklemde (3.1) gösterilmiştir.

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix} \quad (3.1)$$

2. **Farklılık Matrisi (Dissimilarity Matrix):** Bu veri yapısı ise nesnelerin birbirleriyle olan yakınlık bilgisinin tutulduğu matris türüdür ve *nxn* boyutundadır. Farklılık matrisinin şekli genel olarak aşağıdaki denklemde (3.2) gösterilmiştir.

$$\begin{bmatrix} 0 & & & & \\ d\,(2,1) & 0 & & & \\ d\,(3,1) & d\,(3,2) & 0 & & \\ \vdots & \vdots & \vdots & 0 & \\ d\,(n,1) & d\,(n,2) & \cdots & \dots & 0 \end{bmatrix} \quad (3.2)$$

Veri matrisi, satır (nesne) ve sütun (nitelik) olmak üzere iki ayrı indisten oluşur. Bu nedenle veri matrisinden "iki modlu" matris olarak bahsedilebilir. Farklılık matrisi ise "tek modlu" matris olarak bilinir. Nesneler arasındaki uzaklık fonksiyonu değişme özelliğine sahiptir. Bu nedenle farklılık matrisinin asal köşegeni altında ve üstünde kalan değerler simetriktir. Veri matrisinin uzaklık fonksiyonu, farklılık matrisindeki gibi değişme özelliğine sahip olmadığından uygulamada çoğunlukla farklılık matrisi kullanılır (Han ve ark., 2012).

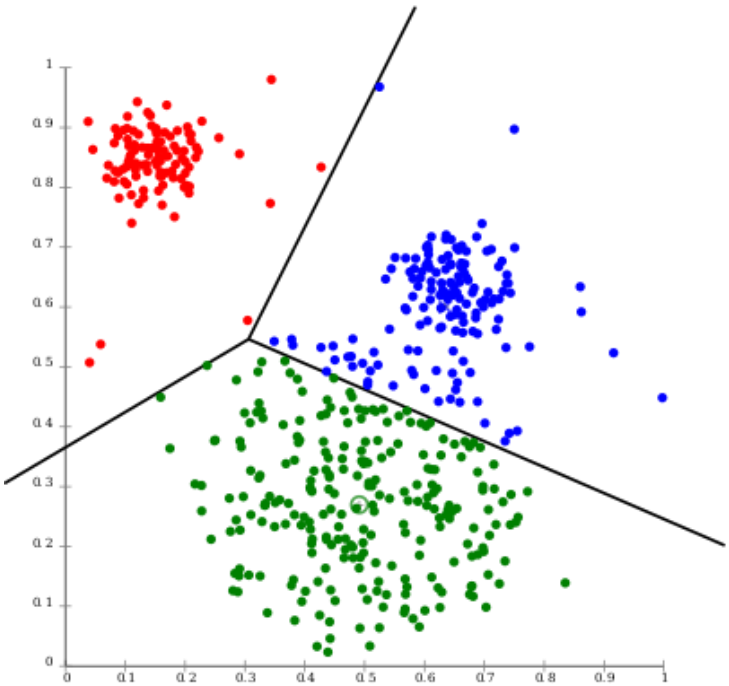
3.6. Kümeleme Yöntemleri

Kümeleme yöntemlerinin sınıflandırılması Şekil 3.13'de gösterilmiştir. Genel olarak kümeleme yöntemlerinden bahsedilirken iki gruba ayrılır. Bunlar; hiyerarşik yöntemler ve bu grubun dışında kalanlar ise hiyerarşik olmayan yöntemlerdir.

3.6.1. Bölümlemeli yöntemler

Kümeleme analizi yapılırken bölümlemeli yöntemler seçildiğinde *n* adet veri daha önceden belirlenmiş olan *k* adet küme sayısına (*k*<n*)* ayrılır. İşlem sonucunda oluşacak küme sayısı daha önceden belli olmalıdır. Kullanıcı, elde ettiği verinin durumuna göre oluşacak küme sayısını belirlemeli, seçtiği algoritmaya kümeler arasında oluşacak mesafeyi (en az ya da en çok olacak şekilde) ve işlem sonucunda beklenen iç benzerlik kriterlerini uygulamanın başlangıcında algoritmaya tanımlamalıdır (Silahtaroğlu, 2013).

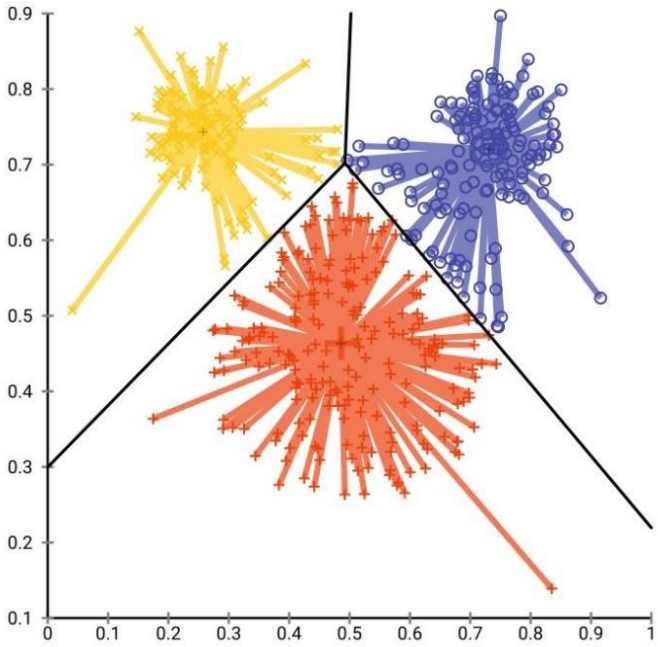
- **K-Means Algoritması:** İlk olarak 1967 yılında Mac Quenn tarafından bulunan algoritma kümelerin en uygun çözüme ulaşılıncaya kadar yenilendiği döngüsel bir algoritmadır. Genel mantığı, *n* adet veri nesnesinden oluşan veri kümesinin *k* adet kümeye bölünmesidir. *k* sayısı kullanıcının bilgisine göre belirlenir. Algoritmanın amacı, bölümleme işlemi sonucunda ortaya çıkan kümelerin iç benzerliklerinin üst düzey olması ve kümeler arasındaki benzerliğin ise en az olmasını sağlamaktır (Silahtaroğlu, 2013). İyi ve kararlı küme sonuçları vermesi, verilerin işleniş



Şekil 3.13. K-Means Algoritması ile Veri Gruplarının Kümelenmesi (URL15)

sıralaması ve ilk merkez seçiminin kümelemeye etkisinin olmaması, merkezi nesneler kümeyi temsil ettiği için gürültülü veriye karşı duyarlı olmaması avantajları arasındadır. Dezavantajı ise uygun küme sayısının belirlenmesi aşamasında birden fazla denemeye ihtiyaç duyulmasıdır (Han ve ark., 2012). K-Means algoritması ile veri gruplarının kümelenmiş hali için bir örnek Şekil 3.13'te verilmiştir.

- **K-Medoids Algoritması:** Temeli, elde edilen verinin yapısal özelliklerini temsil eden *k* adet temsilciyi bulmaya dayanmaktadır. En yaygın kullanılan k-medoids algoritması 1987 yılında Kauffman ve Rousseuw tarafından geliştirilmiştir. Temsilci nesne kümenin merkezine en yakın noktadadır ve medoid olarak adlandırılır. Asıl amaç, veri gruplarını *k* adet kümeye bölerken birbirine çok benzeyen nesnelerin birlikte bulunduğu, farklı kümelerdeki nesnelerin birbirine benzemediği kümeleri bulmaktır. Temsilci nesne, farklı nesnelere olan ortalama uzaklığın en az olduğu, kümenin merkezi nesnesidir. Bölünme metodu, her bir nesne ve o nesnenin referans noktası arasındaki benzerliklerin toplamının küçültülmesi işlemine dayanır. Amaç, *k* adet temsilci nesnenin bulunmasıdır. Bu yüzden algoritma "k-medoids metodu" olarak adlandırılır. *k* adet temsilci nesnenin tespit edilmesiyle her bir nesne en yakın olduğu temsilciye atanır. *k* tane küme oluşturulmuş olur (Işık ve Çamurcu, 2007). İyi ve kararlı kümelerin oluşması, verilerin işleniş sıralarının ve ilk merkez seçiminin kümeleme üzerinde etkisinin olmaması, merkezi elemanlar kümeyi temsil ettiği için gürültülü veriye



Şekil 3.14. K-Medoids Algoritması ile Veri Gruplarının Kümelenmesi (URL16)

karşı duyarlı olmaması avantajları arasındadır. Dezavantajı ise k-means yönteminde olduğu gibi uygun küme sayısının belirlenmesi için birden fazla denemeye ihtiyaç olmasıdır (Han ve ark., 2012). K-medoids algoritması ile veri gruplarının kümelenmiş hali için bir örnek Şekil 3.14'te verilmiştir.

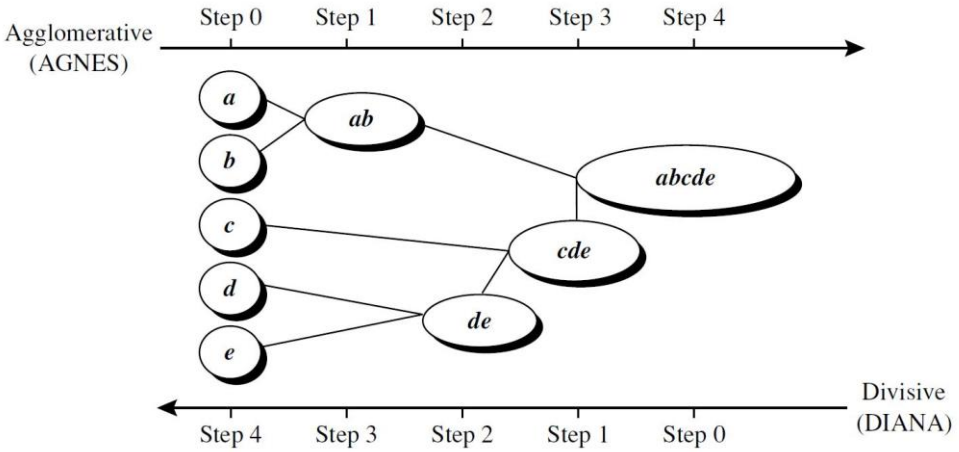
3.6.2. Hiyerarşik yöntemler

Hiyerarşik kümeleme yöntemi, veri nesnelerini belirli bir hiyerarşik tablo şeklinde ya da kümeleri ağaç yapısı şeklinde gruplayarak çalışır. Hiyerarşik kümeleme yöntemleri genel olarak iki tür metot kullanırlar. Kümelerin hiyerarşik olarak ayrıştığı düşünülürse bu ayrışma aşağıdan yukarı olduğunda "birleştirici", yukarıdan aşağıya olduğunda ise "ayırıcı" kümeleme yöntemlerinden bahsedilir (Han ve ark., 2012).

- **Birleştirici Hiyerarşik Kümeleme (AGNES):** Hiyerarşik kümelemede aşağıdan yukarı kümeleme metodunun kullanıldığı yöntemdir. Bu yöntemde her nesne kendi kümesini oluşturur ve kümeler yinelemeli olarak birleşir. Her bir nesne tek bir kümede ya da belirli bir kümede oluncaya kadar işlem yinelenir. İşlem adımları nesnelerin belirli koşullar altında ve benzerlik durumlarına göre birleştirilmesiyle devam eder. Birçok hiyerarşik kümeleme yöntemi birleştirici yöntemler içerisinde yer alır. Diğer yöntemler, küme içi benzerliklerin farklı koşullar altında belirlenmesi ile ayrışırlar (Han ve ark., 2012).
- **Ayırıcı Hiyerarşik Kümeleme (DIANA):** Hiyerarşik kümelemede yukarıdan aşağıya kümeleme metodunun kullanıldığı yöntemdir. Bu algoritmada işlem tüm nesnelerin tek bir kümeye yerleştirilmesi ile başlamaktadır. Daha sonra oluşan bu küme alt kümelere ayrılır ve alt kümeler de yinelemeli olarak daha küçük alt kümelere ayrılır. En alt düzeye bakıldığında kümeler yeterli tutarlılığa ulaştığında işlem durur. En alt düzeydeki kümeler ya sadece bir nesne içermektedir ya da alt düzeydeki küme içerisindeki nesneler birbirine oldukça benzer durumdadır (Han ve ark., 2012).

Kullanıcı, Birleştirici ve Ayırıcı Hiyerarşik Kümeleme yöntemlerinde istenilen sayıda kümeyi işlemin sonlandırılması koşulu olarak belirleyebilir (Han ve ark., 2012).

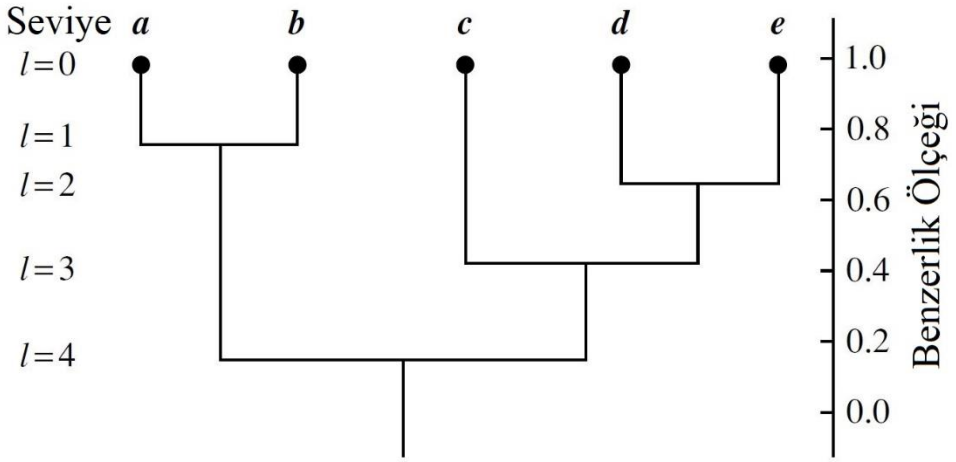
Birleştirici (AGNES) ve Ayırıcı (DIANA) Hiyerarşik Kümeleme yöntemlerinde veri nesnelerinin şeması Şekil 3.15'de gösterilmiştir.



Şekil 3.15. Birleştirici (AGNES) ve Ayırıcı (DIANA) Hiyerarşik Kümeleme Yöntemi İşlem Şeması (Han ve ark., 2012)

Şekil 3.15'da Birleştirici (AGNES) ve Ayırıcı (DIANA) Hiyerarşik Kümeleme algoritmalarının {a, b, c, d, e} veri setleri ile yapılan bir uygulamada işlem sırası görülmektedir. Birleştirici metodu incelediğimizde işlem, her bir nesnenin ayrı küme kabul edilmesiyle başlamaktadır. Daha sonra kümeler belirli kriterlere göre birleştirilir. Ayırıcı metodu incelediğimizde ise tüm nesnelerin başlangıçta tek bir küme olarak kabul edildiği görülmektedir. Daha sonra bu küme, içinde yakın komşuluk ilişkisi olan veriler ile maksimum Öklid uzaklığı gibi kriterlere göre kümelere ayrılır. Burada kümelere ayırma işlemi, her bir nesne tek bir küme oluşturuncaya kadar tekrarlanır (Han ve ark., 2012).

Hiyerarşik kümeleme yönteminde birbirini izleyen değişimleri göstermek amacıyla dendogram adı verilen yapı kullanılır. Dendogram, nesnelerin birlikte nasıl gruplandığını adım adım gösteren bir yapıdır. Aşağıdaki Şekil 3.16'da 5 nesne için bir dendogram gösterilmiştir. l=0 seviyesinde nesneler tekil küme halindedir. l=1 seviyesinde a ve b nesnelerinin birleşmesiyle ilk küme oluşmuştur. Sonraki seviyelerde ise aynı küme içerisinde kalmışlardır. Kümeler arası benzerlik ölçeğini göstermek için dikey eksen kullanılabilir. Örnek olarak, $\{a, b\}$ ve $\{c, d, e\}$ nesne grupları arasındaki benzerliğin yaklaşık 0.16 olduğu söylenebilir. Bu nesne grupları birlikte tek küme olmak için birleşmişlerdir (Han ve ark., 2012).



Şekil 3.16. {*a, b, c, d, e*} Veri Nesnelerinin Hiyerarşik Kümelenmesi İçin Dendogram Gösterimi (Han ve ark., 2012)

3.6.3. Kümeleme analizinde yöntem seçimi

Yapılacak çalışmaya göre hangi analiz yönteminin seçilmesi gerektiği konusunda kesin bir öneri yoktur (Orhunbilge, 2010).

Karagöz (2016), verinin durumuna göre hangi yöntemlerin kullanılabilir olduğu konusunu şöyle özetlemiştir:

- Veri hacmi az ise hiyerarşik kümeleme kullanılabilir.
- K-Means yöntemi, küme sayısı biliniyorsa, orta hacimli veriler için kullanılabilir.
- Veri hacmi büyükse ve veride kategorik veri, sürekli veriler aynı anda varsa iki aşamalı kümeleme (TwoStep Cluster) kullanılabilir.

İki aşamalı kümeleme, iki temel yaklaşım olan, hiyerarşik ve hiyerarşik olmayan kümeleme analizi yöntemlerinin dışında bir yaklaşımdır. Chiu ve arkadaşları (2001) tarafından tanıtılan iki aşamalı kümeleme, özellikle büyük hacimli veriler söz konusu olduğunda kullanılır.

3.6.4. K-Means yöntemi

Bu bölümde yukarıda bahsedilmiş olan yöntemin ayrıntılarına değinilmiştir. Yapılan uygulamada kümeleme analizi yöntemlerinden, hiyerarşik olmayan yöntemler grubuna giren ve bölümlemeli yöntemlerin içerisinde bulunan “K-Means Algoritması” yöntemi kullanılmıştır.

K-Means algoritması kümeleme analizi uygulamalarında en fazla kullanılan yöntemler arasındadır ve tekrarlı bölümleyici bir yapıya sahiptir. Uygulaması kolay olan bir yöntemdir. Büyük hacimli veri gruplarının hızlı ve etkin bir şekilde kümelenmesi

işlemini sağlamaktadır. “k” değeri uygulamaya başlamadan önce algoritmaya girilmesi istenen ve kümeleme analizi işlemi sonlandığında oluşması istenen küme değerini ifade eden sabit bir sayıdır. Bu sayede her bir verinin ait olduğu kümeye olan uzaklıkları toplamını küçültmektedir. Algoritma, aynı zamanda karesel hatayı en küçük yapacak olan k adet kümeyi tespit etmeye çalışır ve iteratif bir yaklaşım ile iyi çözüm veren kümeleme analizi algoritmaları arasında yer almaktadır. Bu yöntemde küme içi benzerliğin büyük, kümeler arası benzerliğin küçük olduğu sürece yapılan uygulamanın doğruluğundan söz edilebilir (URL5).

Verilerin her biri n-boyutlu reel vektör olmak üzere bir $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ veri kümesi ve bölünecek küme sayısı olan k küme sayısı verilsin. K-Means kümeleme yöntemi oluşacak karesel hatayı en aza indirgemek için N tane veriyi K adet $S = \{S_1, S_2, \ldots, S_K\}$ kümeye bölümlemeyi amaçlar. Başka bir deyişle;

$$\mu_i = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i \qquad (3.3)$$

burada μ_i, S_j’deki noktaların ortalaması olmak üzere

$$arg_S min \sum_{j=1}^{K} \sum_{x_i \in S_j} ||x_i - \mu_j||^2 \qquad (3.4)$$

bulmaktır (URL3).

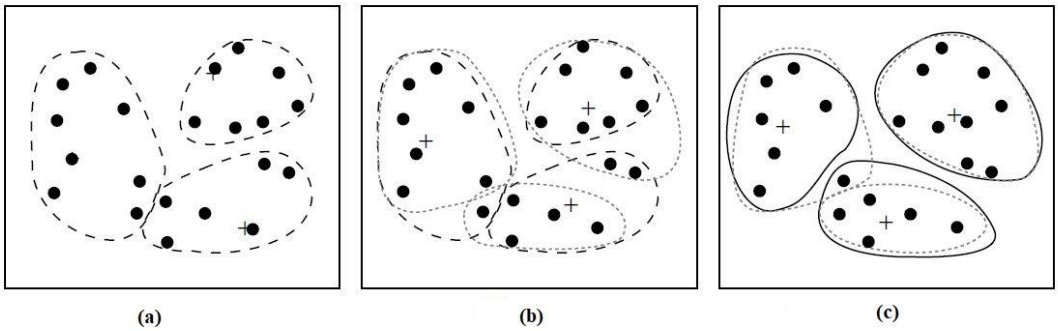
K-Means algoritmasının veri setine uygulandıktan sonra oluşan durumu Şekil 3.17’de gösterilmiştir. Bu şekilde görüldüğü gibi küme sayısı k=3 olarak seçilmiştir.

K-Means algoritmasının çalışma prensibine göre uygulama için önceden belirlenmiş küme sayısı olan k seçilir ve sisteme dahil edilir. Belirlenmiş olan K adet kümenin ortalama değerleri hesaplanarak küme merkezleri belirlenir ve nesnelerin merkeze olan uzaklıkları izlenir. Herhangi bir değişim olmayıncaya kadar algoritma kendini tekrar eder (URL5).

Algoritma temel olarak 4 basamaktan oluşur (URL5):

1. Her bir kümenin merkezlerinin belirlenmesi,
2. Merkez dışında kalan verilerin mesafelerine göre kümelenmesi,
3. Oluşan yeni kümelere göre yeni merkezlerin belirlenmesi, (ya da ilk belirlenen merkezlerin yeni oluşan merkez değerine kaydırılması)

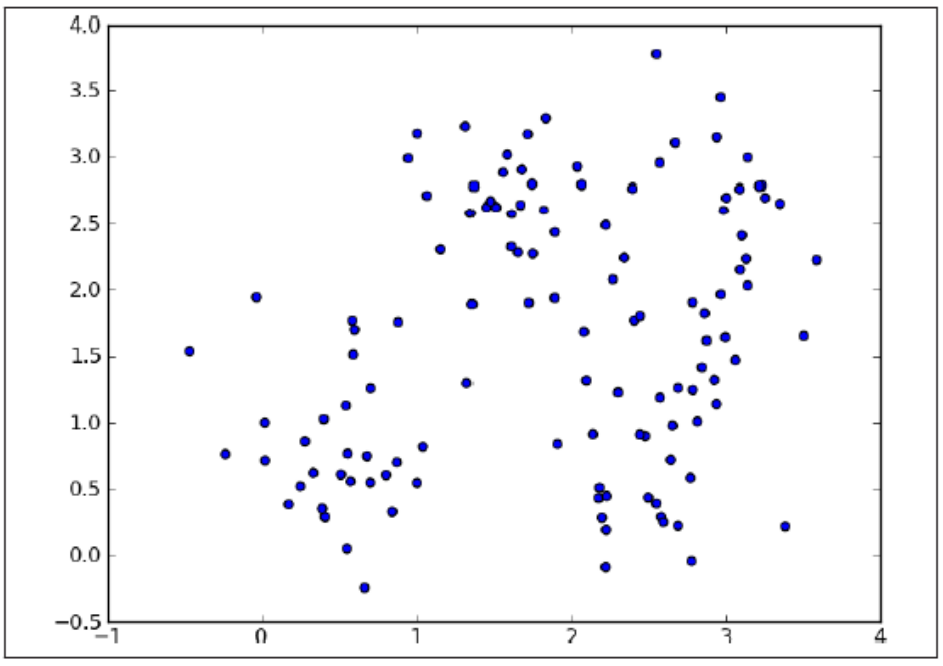
4. Kümeler kararlı hale gelene kadar 2. ve 3. adımların tekrarlanması.



Şekil 3.17. K-Means Algoritması ile Kümeleme aşamaları (Han ve ark., 2012)

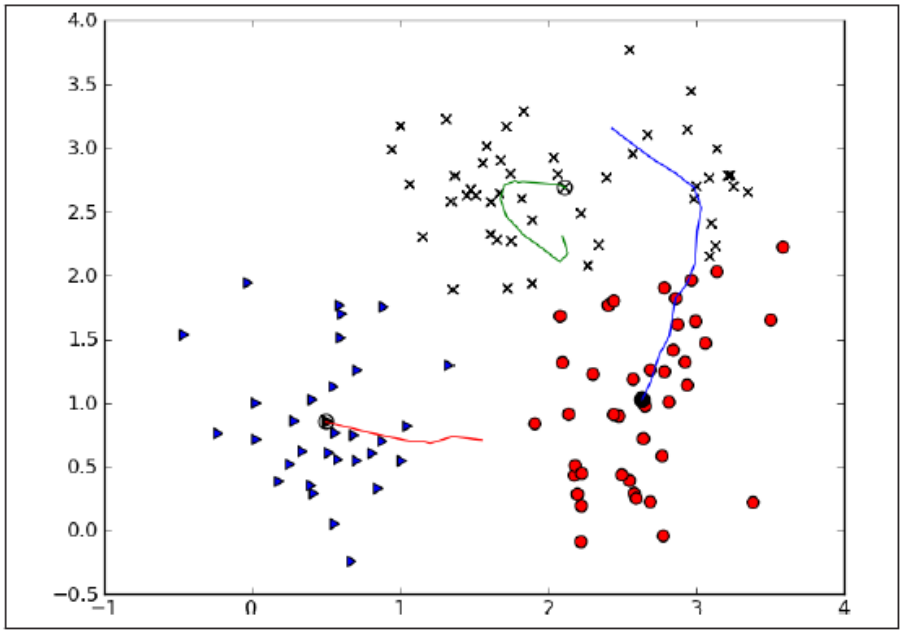
Şekil 3.17'de K-Means yöntemi kullanılarak bir grup nesnenin kümelenmesi işlemi gösterilmiştir. Her kümenin ortalama değeri "+" ile belirtilmiştir (Han ve ark., 2012).

Şekil 3.19'da K-Means algoritması kullanılarak yapılan ve 90 adet veri noktasından oluşan bir örnek gösterilmiştir. Şekil 3.18'de kümeleme analizi öncesi ilk veri noktaları gösterilmektedir. Örnekte küme sayısı "k=3" olarak belirlenmiştir (URL12).



Şekil 3.18. K-Means Algoritması ile yapılan kümeleme analizi öncesi durum (90 adet veri noktası) (URL12)

Şekil 3.19'da K-Means algoritması kullanılarak yapılan ve 16 yineleme sonrası veri noktalarının kümelerinin değişimi, oluşan küme merkezleri ve bu küme merkezlerinin değişim çizgisi gösterilmiştir. Algoritmanın yinelenmesi ile birlikte kümeler arası benzerliğin en az, küme içi benzerliğin ise üst seviyede olması hedeflenir. Böylece küme merkezleri kararlı hale gelecektir (URL12).



Şekil 3.19. K-Means Algoritması kullanılarak yapılan uygulamada küme merkezlerinin değişim çizgisi (URL12)

K-Means algoritması, farklı merkezler ile birlikte birden çok kez çalıştırılarak kararlı küme merkezlerinin bulunması sağlanır. K-Means yönteminde uygun küme sayısının belirlenmesi için farklı "k" değerlerinde algoritma uygulanır ve çıkan sonuçlara göre çeşitli geçerlilik indeks değerleri hesaplanarak kümelerin anlamlı olup olmadığına karar verilir (URL12).

3.7. Verilerin Normalize Edilmesi

İstatistikte, örneklem uzayındaki sayıların birbirleriyle ilişkilendirilmesi ve birbirleriyle hesaplamalar yapılabilmesi için kullanılan yöntemlere normalizasyon denilmektedir. Kümeleme analizi yapılırken daha sağlıklı sonuçlar alınabilmesi için kullanılacak verilerin normalize edilmesi, birbirleriyle ilişkili hale getirilmesi gerekmektedir. Bu çalışmada normalizasyon yöntemlerinden biri olan z-sayısı yöntemi kullanılacaktır.

İstatistiksel verilerin normalizasyonunda kullanılan z-sayısı, herhangi bir sayının istatistiksel dağılımdaki ortalamaya olan uzaklığının standart sapmaya bölünmesiyle elde

edilir. Buna standartlaştırılmış rastgele değişken denir. (3.5) formülüyle hesaplanmaktadır.

$$Z = \frac{X - \mu}{\sigma} \quad (3.5)$$

3.8. Küme Sayılarında Geçerlilik

Hiyerarşik olmayan kümeleme analizi yöntemlerinde, küme sayısı kullanıcı tarafından, analiz sonucuna göre belirlenir.

Kümeleme analizi sonucunda geçerli ve anlamlı sonuçların sağlanması için gerekli koşullardan birisi de küme sayısının doğru seçilmesidir (Çakmak, 1999).

Kümeleme analizi sonucu ortaya çıkan verilerin geçerliliği konusu, üzerinde oldukça durulması gereken bir konudur. Verilerin analiz edilmesi ve sonrasında kümeleme çözümünden sonra ortaya çıkan kümelerin güvenilirlik açısından herhangi bir garantisi yoktur. Bu nedenle belirlenen küme sayısında ortaya çıkan kümelerin tesadüfi bir çözüm mü yoksa kabul edilebilir bir çözüm mü olup olmadığının belirli testlerle belirlenmesi gerekmektedir (Çakmak, 1999).

Hiyerarşik olmayan yöntemler grubuna giren K-Means Algoritması, küme sayısının kullanıcı tarafından belirlendiği bir yöntemdir. Bu nedenle belirlenen küme sayısının geçerliliğinin test edilmesi, anlamlı olup olmadığının belirlenmesi gerekmektedir. Bunun içinse farklı küme sayılarında K-Means Algoritması ile işlem yapılması gerekir.

K-Means yöntemiyle yapılan kümeleme analizi sonucu ortaya çıkan farklı küme sayılarındaki analizler için oldukça kullanılan geçerlilik testi yöntemleri arasında Davies-Bouldin İndeksi de yer almaktadır.

3.8.1. Davies-Bouldin indeksi

Davies-Bouldin indeksi 1979 yılında David L. Davies ve Donald W. Bouldin tarafından belirlenmiştir ve kümeleme algoritmalarını değerlendirmek için kullanılmaktadır (Davies ve Bouldin, 1979). Davies-Bouldin indeksi, kümelerin içerisinde bulunan veri miktarını, özelliklerini değerlendirir ve küme içi benzerlik ile kümeler arası benzerlik oranına dayanan bir fonksiyon olarak tanımlanabilir. Burada indeks değerinin düşük olması ile kümeleme analizinin daha iyi sonuç verdiği ortaya çıkar. Belirli küme sayıları ile yapılan indeks hesaplamaları sonucu en düşük değere sahip analizin küme sayısı daha doğrudur. Diğer küme sayıları ile ortaya çıkan analizler düşünüldüğünde, indeks değerlerinin hesaplanmasıyla belirlenen küme sayısı ile yapılan kümeleme analizi sonucunda küme içi benzerlik fazla, kümeler arası benzerlik ise en

düşük seviyededir. Davies-Bouldin İndeksi'nin düşük çıktığı küme sayısı ile yapılan analizde gerçekten de kümelerin diğer kümelerden ayrıştığı, kümeler arası benzerliğin en alt seviyede olduğundan söz edilebilir (URL6).

Bir geçerlilik indeksi olan Davies-Bouldin indeksi DB katsayısıyla gösterilmektedir ve formülü (3.6) eşitliğinde verilmiştir.

$$\boldsymbol{DB} = \frac{1}{n}\sum_{i=1}^{n} max\left(\frac{S_n(Q_i)+S_n(Q_j)}{S_n(Q_i,Q_j)}\right) \quad (3.6)$$

Denklemde bulunan *n* küme sayısını, S_n kümenin her bir elemanının merkeze olan uzaklıklarının aritmetik ortalamasını ve $S_n(Q_i, Q_j)$ iki küme arasındaki uzaklığı göstermektedir. DB değerinin küçük olması, küme elemanlarının küme merkezlerine olan uzaklıklarının en az olmasını, kümelerin her birinin homojen yapıda olmasını ve kümeler arası uzaklığın en fazla olduğunu belirtir (Çağlar, 2018).

DB değerinin büyük ya da küçük olmasından bahsedebilmek için en az iki defa kümeleme işleminin yapılması ve her bir işlem için bu indeks değerinin hesaplanması gerekmektedir (Silahtaroğlu, 2013).

K-Means yöntemi kullanılan kümeleme analizi uygulamaları sonucunda oluşan kümeler için DB değerinin hesaplanmasıyla yapılan en uygun küme sayısı çıkarımı ve hesaplanan değere göre kümelerin gösterilmesi, RapidMiner yazılımında yapılan örnek çalışma ile sağlanmıştır. Burada bir veri grubu rastgele belirlenen sayılardan oluşturulmuştur. Diğer üç veri grubu ise hesaplanan DB değerleri ve kümelerin durumlarının gösterilmesi için oluşturulmuş veri gruplarıdır. Veri grupları, Çizelge 3.1, Çizelge 3.2, Çizelge 3.3 ve Çizelge 3.4'de verilmiştir. Oluşan kümeler ve bu kümelere ait DB değerleri ise Şekil 3.20, Şekil 3.21, Şekil 3.22 ve Şekil 3.23'de verilmiştir.

Çizelge 3.1. Rastgele oluşturulan 1. veri grubu

No	x	y
1	34	1
2	25	17
3	6	28
4	11	6
5	45	29
6	28	2
7	7	11
8	39	49
9	16	29
10	15	29
11	22	49
12	37	24
13	49	5
14	16	11
15	32	21
16	43	1
17	31	41
18	46	39
19	0	7
20	27	11
21	10	21
22	16	13
23	44	4
24	35	49
25	19	49
26	43	22
27	18	36
28	4	22
29	20	21
30	33	5

Çizelge 3.2. Oluşturulan 2. veri grubu

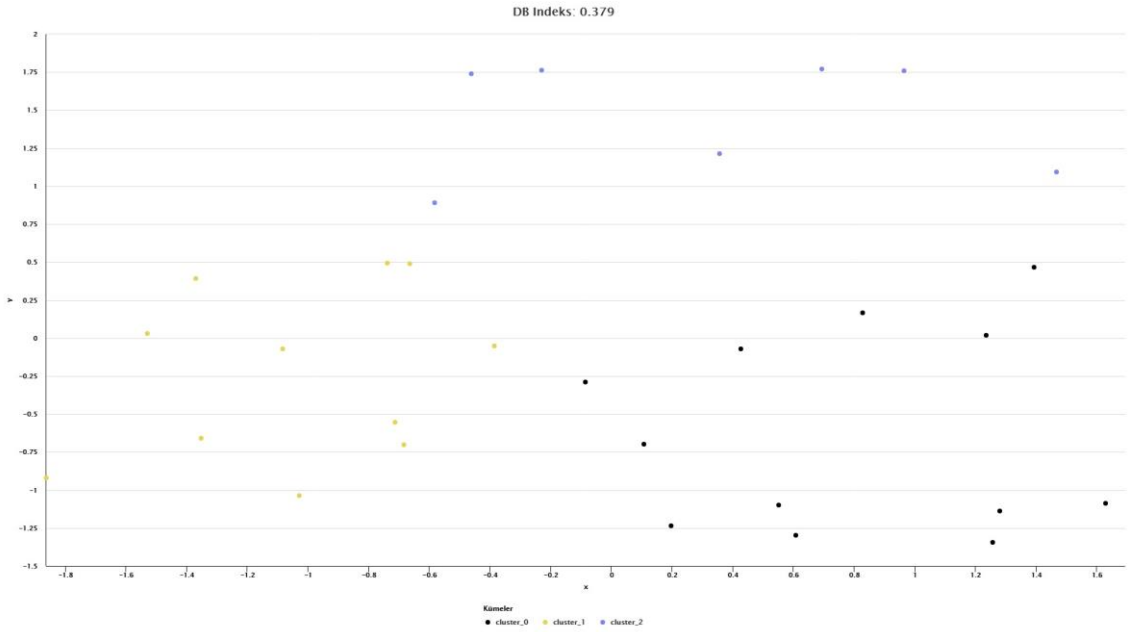
No	x	y
1	1	1
2	1	2
3	1	3
4	1	4
5	1	5
6	1	6
7	1	7
8	1	8
9	1	9
10	1	10
11	2	21
12	2	22
13	2	23
14	2	24
15	2	25
16	2	26
17	2	27
18	2	28
19	2	29
20	2	30
21	3	31
22	3	32
23	3	33
24	3	34
25	3	35
26	3	36
27	3	37
28	3	38
29	3	39
30	3	40

Çizelge 3.3. Oluşturulan 3. veri grubu

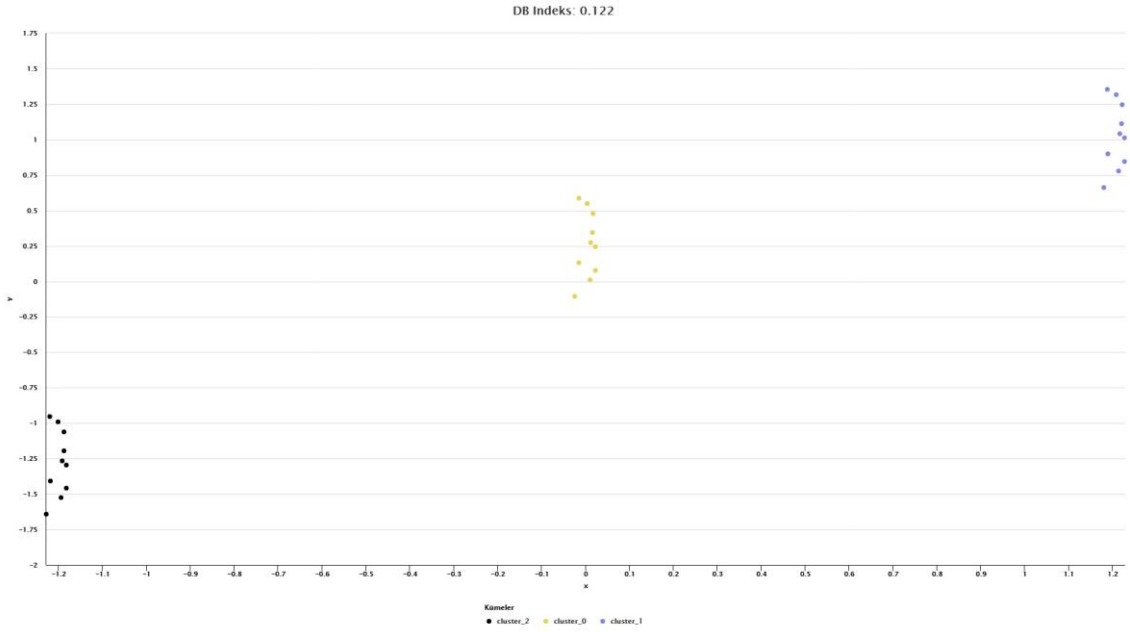
No	x	y
1	1	1
2	1	2
3	1	3
4	1	4
5	1	5
6	1	6
7	1	7
8	1	8
9	1	9
10	1	10
11	2	31
12	2	32
13	2	33
14	2	34
15	2	35
16	2	36
17	2	37
18	2	38
19	2	39
20	2	40
21	3	81
22	3	82
23	3	83
24	3	84
25	3	85
26	3	86
27	3	87
28	3	88
29	3	89
30	3	90

Çizelge 3.4. Oluşturulan 4. veri grubu

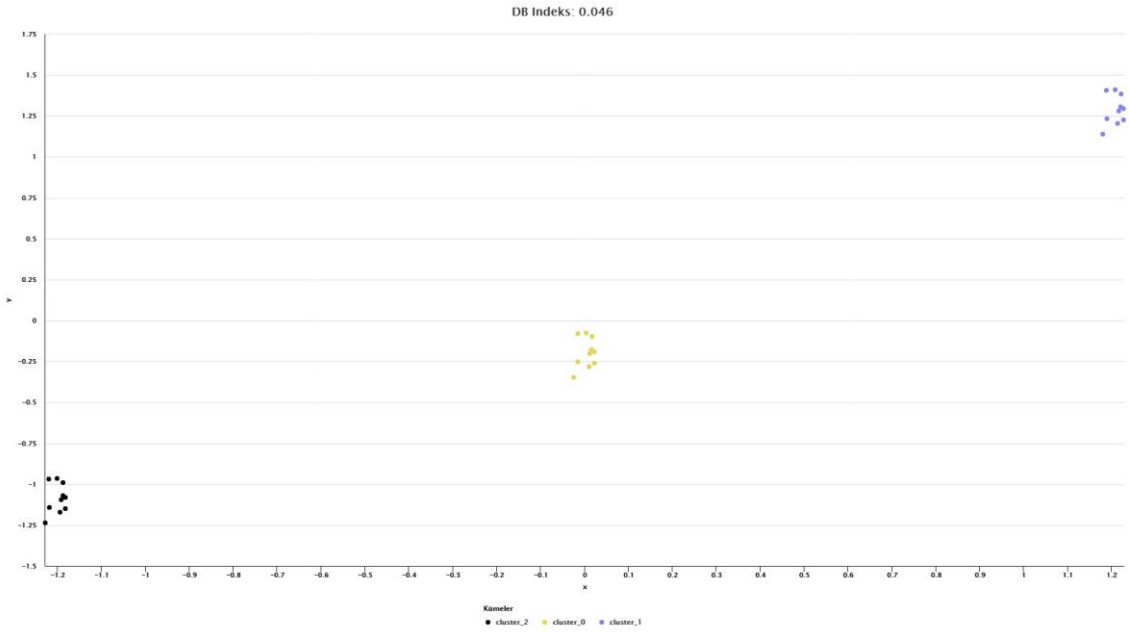
No	x	y
1	1	1
2	1	2
3	1	3
4	1	4
5	1	5
6	1	6
7	1	7
8	1	8
9	1	9
10	1	10
11	2	101
12	2	102
13	2	103
14	2	104
15	2	105
16	2	106
17	2	107
18	2	108
19	2	109
20	2	110
21	3	301
22	3	302
23	3	303
24	3	304
25	3	305
26	3	306
27	3	307
28	3	308
29	3	309
30	3	310



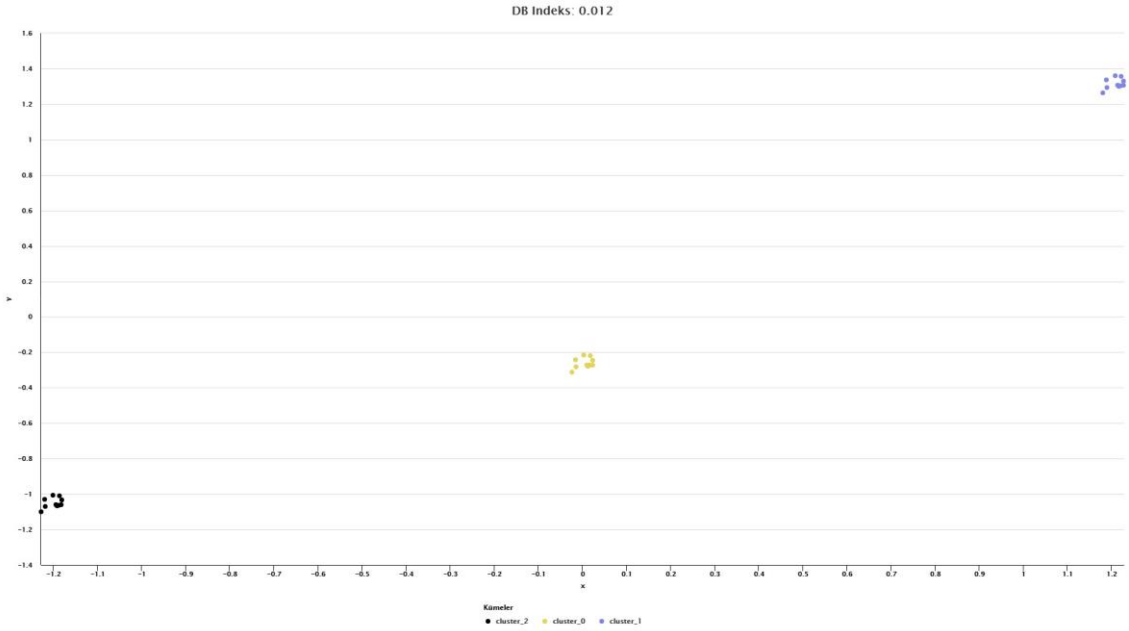
Şekil 3.20. Rastgele sayılarla oluşturulan 1. veri grubu ile yapılan kümeleme (DB Indeks: 0.379)



Şekil 3.21. Oluşturulan 2. veri grubu (DB Indeks: 0.122)



Şekil 3.22. Oluşturulan 3. veri grubu (DB Indeks: 0.046)



Şekil 3.23. Oluşturulan 4. veri grubu (DB Indeks: 0.012)

Oluşturulan veri grupları ile üç küme sayısına sahip olmak üzere planlanarak yapılan çalışmada hesaplanan DB değerleri ve oluşan kümeler gösterilmiştir. Üç küme sayısına sahip çalışma için belirlenen 2., 3. ve 4. veri grupları içerisindeki sayısal değerlerin sırasıyla gruplar halinde birbirlerinden uzaklaşması istenmiş ve değerler bu doğrultuda seçilmiştir. İkili veri grubu olarak rastgele oluşturulan sayısal değerli 1. veri grubunda üç küme için DB İndeksi 0.379 olarak belirlenmiştir. Üç küme için analiz sonucunda birbirinden en uzak olması istenen 4. veri grubundaki hesaplanan DB İndeksi 0.012'dir. 2. ve 3. veri grubunda hesaplanan DB İndeks değerleri ise sırasıyla 0.122 ve 0.046'dır. Hesaplanan indeks değerleri birbirleriyle karşılaştırıldığında, sıfıra yaklaşan değere sahip kümeleme işleminde kümeler arası benzerliğin en alt seviyede olduğu, diğer kümelerden ayrıştığı ve küme içi benzerliklerin ise üst seviyede olduğu görülmektedir. DB değeri en yüksek olan kümeleme işleminde ise diğerlerine göre kümelerin daha az ayrıştığı, küme içi ve kümeler arası benzerliklerin ise sıfıra yaklaşan diğer analiz işlemlerindeki gibi keskin olarak ayrışmadığı görülmektedir.

3.9. QGIS Desktop

QGIS yazılımı veri görüntüleme, düzenleme ve çözümleme yeteneklerini sağlayan birçok platformu destekleyen açık kaynaklı bir CBS yazılımıdır. İşlevsellik yönünden QGIS, haritaların raster veya vektör katmanlarının oluşmasına olanak verir. Vektör verileri nokta, çizgi ya da poligon özellikte saklanır. Bu tür bir yazılım için varsayılan bir durumdur. Çeşitli türde desteklenen raster görüntüler ve yazılım görüntülerinde jeoreferanslama gerçekleştirilir.

QGIS, yapılacak çalışmaya özel olarak oluşturulan eklentilerin, yazılım içerisine entegre edilmesine izin verir. Böylece yazılımdan ayrılmadan yapılan çalışmalar, veri yönetimi ve ortaya çıkan sonuçların yorumlanması açısından oldukça önemlidir.

QGIS eklentileri, yapılan uygulamalara ek işlevler sağlar. Kullanılmaya hazır birçok eklenti koleksiyonu vardır. QGIS eklenti yöneticisinden de doğrudan kurulabilir. QGIS yazılımı için özel olarak üretilen eklentiler, bağımsız kuruluşlar ve geliştiriciler tarafından oluşturulmuştur. Kullanılan eklentiler ile ilgili herhangi bir hata olması durumunda, eklentinin geliştiricisine hata hakkında e-mail gönderilebilir (URL8).

3.9.1. Attribute based clustering

Attribute Based Clustering, Eduard Kazakov tarafından oluşturulan ve ilk sürümü 26 Nisan 2016 yılında yayınlanan bir QGIS eklentisidir. Öznitelik verisine dayalı olarak kümeleme analizi yapılmasına olanak tanır.

Eklenti ile hiyerarşik kümeleme analizi veya K-Means Algoritmaları kullanılarak istenilen sayıda öznitelik verisine dayalı olan vektör katmanının verileri için kümeleme analizi yapılabilir. Hiyerarşik kümeleme algoritması için küme sayısının bilindiği durum ve küme sayısının bilinmediği durum olmak üzere iki yöntem vardır. K-Means yöntemi kullanılıyorsa küme sayısının bilinmesi gerekmektedir. Hiyerarşik kümeleme yönteminde, istenilen öznitelik verisi için ağırlık tanımlanabilir. Hiyerarşik kümeleme analizinde küme sayısının bilinmediği durumda en uygun küme sayısının tespit edilmesi için belirli algoritmalar tanımlanmıştır. Ayrıca herhangi bir analiz yöntemi için her bir nesnenin oluşan küme merkezlerine olan uzaklıkları da hesaplanır (URL9).

3.10. RapidMiner

Eski adı YALE (Yet Another Learning Environment) olan RapidMiner yazılımı, ilk kez 2001 yılında Ralf Klinkenberg, Ingo Mierswa ve Simon Fischer tarafondan Dortmund Teknik Üniversitesi bünyesinde bulunan yapay zeka biriminde geliştirilmiştir. Ancak 2006 yılı itibariyle Ingo Mierswa ve Ralf Klinkenberg tarafından kurulan Rapid-I isimli şirket tarafından geliştirilmeye başlanmıştır. 2007 yılında, yazılımın adı YALE'den RapidMiner'a çevrilmiştir. 2013 yılında ise Rapid-I şirketi RapidMiner ürününü markalaştırmıştır (URL10, URL11).

RapidMiner, makine öğrenmesi, veri madenciliği, metin madenciliği, tahmin edici analiz ve iş analizi gibi amaçlar için geliştirilmiş bir yazılımdır. Ticari bir yazılım olduğu için genel olarak iş ve ticaret uygulamalarında kullanılır. Aynı zamanda araştırma, eğitim, hızlı prototipleme ve uygulama geliştirme gibi amaçlarla da kullanılır. Veri madenciliği süreçlerinin bütün aşamaları yazılım tarafından desteklenmektedir. Bu nedenle veri hazırlama, sonuçları görselleştirme, doğrulama ve optimizasyon amaçları için de RapidMiner'ın kullanılması mümkündür. RapidMiner, açık çekirdek modeli ile geliştirilmiştir ve RapidMiner Temel Sürümü (RapidMiner Basic Edition) AGPL lisansı ile indirilebilir. Profesyonel lisansı ise belirli bir ücrete tabiidir (URL10, URL11).

RapidMiner yazılımı, eğitim için kullanılabilmesi için iki yıl süreyle eğitim lisansı ile de kullanılabilmektedir. Bu da eğitim amaçlı kullanımlarda büyük avantaj sağlamaktadır.

4. ARAŞTIRMA VE TARTIŞMA

Yapılan bu tez çalışmasında; istatistiksel verilerin analizinde kullanılan, veri madenciliği tekniklerinden olan kümeleme analizi yöntemlerinden K-Means Algoritması'nın Tematik Kartografya'da çok değişkenli harita yapımı için kullanılabilirliğinin araştırılması, küme sayısının kullanıcı tarafından girildiği K-Means Algoritması'nda en uygun küme sayısının tespit edilmesinde uygulanan işlemlerin çalışma prensibinin araştırılması amaçlanmıştır. Çalışmada iki uygulama yapılmış ve ilk uygulamada 2019 yılına ait istatistiksel veri olarak km^2'ye düşen taşıt sayısı ile buna karşılık kullanılan akaryakıt miktarı (m^3) veri grupları oluşturulmuştur. İkinci uygulamada ise 2018 yılına ait ortalama kişi başı tüketilen su miktarı (m^3) ve kişi başı yıllık arz edilen su miktarı ((m^3)/kişi-yıl) veri grupları oluşturulmuştur. Bu istatistiksel verilerin kümeleme analizi, K-Means Algoritmasıyla gerçekleştirilecek ve bu işlem için CBS yazılımlarından, açık kaynak kodlu bir yazılım olan QGIS yazılımından yararlanılacaktır. En uygun küme sayısının tespitinde kullanılan, bir geçerlilik indeksi hesaplama yöntemlerinden birisi olan Davies-Bouldin İndeksi değeri ise RapidMiner yazılımı ile hesaplanacaktır. Elde edilen sonuç veriler ile iller bazında Türkiye haritası altlığı kullanılacak olup bir tematik harita üretimi amaçlanmıştır ve bu işlem QGIS yazılımı ile gerçekleştirilecektir. Tematik harita üretiminde, küçük ölçekli harita yapımında tercih edilen bir projeksiyon sistemi olan Lambert Konform Konik Projeksiyonu kullanılacaktır.

4.1. Uygulama 1

4.1.1. Konu seçimi ve verilerin hazırlanması

Son yıllarda ülkemizde artan araç sayısı ve buna karşılık artan akaryakıt kullanımı NO_x gazlarında önemli ölçüde artışa sebep olmuştur. Bu da hava kirliliğinin artmasına sebep olup insan sağlığını tehdit etmektedir. Düzenlenmiş veriler iller düzeyinde olup 2019 yılına ait motorlu kara taşıt sayısının yüz ölçümüne (km^2) oranı ve buna karşılık kullanılan akaryakıt miktarını (km^2'de kullanılan akaryakıt miktarı) göstermektedir.

Kümeleme analizinde ilk aşama veri kümelerinin temin edilmesidir. Verilerin temininde kaynak olarak; akaryakıt kullanımı için T.C. Enerji Piyasası Düzenleme Kurumu'nun sitesi (www.epdk.gov.tr), motorlu kara taşıt sayısı için Türkiye İstatistik Kurumu'nun sitesi (www.tuik.gov.tr), yüz ölçümü (km^2) verisi için Harita Genel Müdürlüğü'nün sitesi (www.harita.gov.tr) kullanılmıştır. Kümeleme analizi buradan elde

edilen veriler ve verilerin düzenlenmesi, oranlanması ile elde edilen, uygulama için gerekli olan veri kümeleri ile yapılmıştır.

2019 yılına ait olan motorlu kara taşıt sayısı verisi; otomobil, minibüs, otobüs, kamyonet, kamyon, motosiklet, özel amaçlı araç, yol ve iş makinaları ve traktör sayılarının toplamından oluşmakta olup benzin ve motorin tüketen araçlar baz alınmıştır. İl bazında olan motorlu kara taşıt sayısı Ek1-a'da, yüz ölçümü (km^2) verileri Ek1-b'de gösterilmiştir. Ek1-c'de 2019 yılına ait, il bazında benzin ve motorin tüketiminin ton biriminde toplamı, Ek1-d'de motorlu kara taşıt sayısının yüz ölçümü (km^2) verisine oranı ve Ek1-e'de kullanılan akaryakıt miktarının (m^3) yüz ölçümüne (km^2) oranı bulunmaktadır. Akaryakıt miktarı verisi ton biriminde elde edilip benzin (740 kg/m^3) ve motorin (835 kg/m^3) yoğunlukları ayrı ayrı hesaba katılarak m^3 birimine çevrilmiştir.

Verilerin elde edilmesiyle yapılacak olan uygulamadaki amaç, il düzeyinde km^2'ye düşen araç sayısına (yüz ölçümü bakımından ildeki araç yoğunluğu) göre satışı gerçekleştirilen akaryakıt miktarıdır. Kümeleme analizi sonucuna göre çeşitli çıkarımlarda bulunulmuştur.

4.1.2. Verilerin normalize edilmesi

2019 yılına ait akaryakıt kullanımının (benzin-motorin kullanımı (m^3)) yüz ölçümüne oranı verisi için hesaplanan ortalama 47 ve standart sapma 108'dir.

Taşıt sayısının yüz ölçümüne (km^2) oranı verisi için hesaplanan ortalama 35 ve standart sapma 86'dır.

Bu iki istatistiksel verideki değerlerin her biri için ortalama ve standart sapma değerleri kullanılarak z-sayısı değerleri hesaplanmış ve artık kümeleme analizi için kullanılacak verilerin normalizasyon işlemi tamamlanmıştır. Z-sayısı yöntemi ile normalize edilen veriler Çizelge 4.1'de verilmiştir.

Çizelge 4.1. Z-Sayısı yöntemi ile verilerin normalize edilmesi

	X	**Y**		**X**	**Y**
P.Kodu	**z-sayısı**	**z-sayısı**	P.Kodu	**z-sayısı**	**z-sayısı**
1	0.1479	-0.0048	42	-0.1966	-0.2190
2	-0.2350	-0.2792	43	-0.1924	-0.1647
3	-0.2176	-0.2391	44	-0.2354	-0.2949
4	-0.3691	-0.3482	45	0.1109	0.0417
5	-0.1600	-0.2441	46	-0.2143	-0.2647
6	0.5173	0.5261	47	-0.3020	-0.2416
7	0.2302	0.0280	48	0.0619	-0.0436
8	-0.3412	-0.3279	49	-0.3572	-0.3767
9	0.2490	0.0709	50	-0.2390	-0.2741
10	-0.0172	-0.0656	51	-0.2304	-0.2499
11	-0.2110	-0.2154	52	-0.1331	-0.1479
12	-0.3774	-0.3709	53	-0.1598	-0.1725
13	-0.3711	-0.3778	54	0.2891	0.2694
14	-0.2420	-0.2022	55	0.0263	0.0639
15	-0.1843	-0.2182	56	-0.3606	-0.3060
16	0.5655	0.5669	57	-0.2816	-0.3415
17	-0.1293	-0.1881	58	-0.3360	-0.3695
18	-0.3213	-0.3262	59	0.1021	0.2560
19	-0.2424	-0.2761	60	-0.1954	-0.2525
20	-0.0101	0.0813	61	0.0923	0.1712
21	-0.3089	-0.2773	62	-0.3879	-0.4118
22	-0.0993	0.3542	63	-0.2496	-0.2636
23	-0.2461	-0.2861	64	-0.1152	-0.2488
24	-0.3435	-0.3605	65	-0.3594	-0.3869
25	-0.3470	-0.3661	66	-0.3119	-0.3362
26	-0.1638	-0.1616	67	0.1360	-0.0359
27	0.4809	0.5381	68	-0.2116	-0.2303
28	-0.2512	-0.2534	69	-0.3540	-0.3767
29	-0.3591	-0.3806	70	-0.2802	-0.3558
30	-0.3872	-0.3921	71	-0.2374	-0.0422
31	0.6320	0.5292	72	-0.2860	0.0114
32	-0.1713	-0.2890	73	-0.3542	-0.2661
33	0.0461	0.1758	74	-0.1458	-0.2346
34	8.4840	8.1844	75	-0.3571	-0.3688
35	0.9868	0.8390	76	-0.3134	-0.3423
36	-0.3508	-0.3763	77	0.5561	0.8406
37	-0.2866	-0.3441	78	-0.2184	-0.1915
38	-0.1435	-0.1881	79	-0.0145	-0.2210
39	-0.1628	-0.1204	80	0.1723	-0.1021
40	-0.2820	-0.3366	81	0.1167	0.0976
41	0.9591	2.3042			

X: MKT/Yüz Ölçümü km^2 (2019); **Y:** Akaryakıt (m^3)/Yüz Ölçümü (km^2)

4.1.3. Korelasyon katsayısının hesaplanması

İstatistikte, kovaryans iki değişkenin birlikte nasıl değiştiklerinin bir ölçüsüdür. Kovaryans, beklenen değerleri $E(X) = \mu$ ve $E(Y) = v$ olan X ve Y olarak tanımlanmış iki gerçel değerli rastgele değişken arasındaki ilişki (4.1) formülüyle tanımlanır (URL13):

$$cov(X, Y) = E((X - \mu)(Y - v)) \quad (4.1)$$

Burada E, beklenen değeri temsil etmektedir.

Kovaryansı sıfır olan iki rastgele değişkene korelasyonsuz değişken adı verilir.

Kovaryans $cov(X, Y)$ ölçümünün birimi X çarpı Y sonucunun ölçü birimidir. Buna karşılık, kovaryansın lineer regresyon analizi ile standartlaştırılmasıyla elde edilen korelasyon birimsizdir.

Korelasyon, olasılık kuramı ve istatistikte iki rastgele değişken arasındaki doğrusal ilişkinin yönünü ve gücünü belirtir. Genel istatistiksel kullanımda korelasyon, bağımsızlık durumundan ne kadar uzaklaşıldığını gösterir (URL14).

Kümeleme analizine başlamadan önce kümeleme analizi yapılacak istatistiksel verilerin birbirleriyle olan doğrusallığının, birbirleriyle olan ilişkilerinin belirlenmesi gerekmektedir. Birbirleriyle ilişkili olmayan verilerin kümelenmesi ile istenmeyen sonuçlar ortaya çıkabilir. Bunun önüne geçmek için korelasyon katsayısı hesaplanmalıdır.

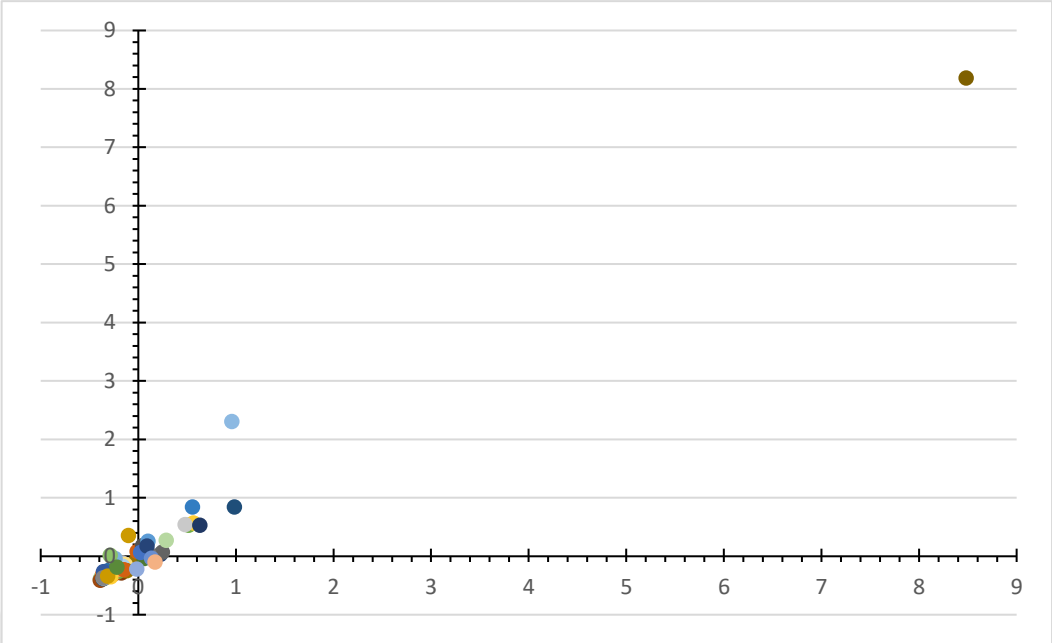
Korelasyon -1, 1 aralığında değişmektedir. Hesaplanan değer -1, 1 değerlerine yaklaştıkça veriler arasındaki ilişki güçlüdür. Sıfıra yaklaştıkça veriler arasındaki ilişki zayıflar. Korelasyon katsayısı sıfır ise korelasyon yoktur. Verilerin doğrusallığının kontrolünde bu duruma bakılır.

Elde edilen her iki istatistiksel veri grubunun korelasyon katsayısı (r) aşağıdaki (4.2) formülüyle hesaplanmıştır.

$$r\ (x, y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(x-\bar{x})^2}} \quad (4.2)$$

2019 yılına ait olan motorlu kara taşıt sayısının illerin yüz ölçümüne (km^2) oranı ve buna karşılık olarak kullanılan akaryakıt miktarı (m^3) verilerinin korelasyon katsayısı r = 0.98 olarak hesaplanmıştır. Bu da verilerin birbirleriyle korelasyonunun olduğunu ve kümeleme analizi için birbirleriyle ilişkileri bakımından bir sorun teşkil etmediğini göstermektedir.

Şekil 4.1'de normalize edilmiş verilerin grafik gösterimi yer almaktadır. Bu grafikten de verilerin birbirleriyle olan ilişkilerine göz atılabilir.



Şekil 4.1. Normalize edilmiş verilerin grafik gösterimi

4.1.4. Küme sayısının belirlenmesi

Kullanılacak olan en optimum küme sayısının (k) belirlenmesi için veriler 3, 4, 5, 6 ve 7'li kümelere ayrılmış ve oluşan kümelerin Davies Bouldin (DB) indeks değerleri hesaplanmıştır. Kümelerin Davies Bouldin indeks değerlerinin hesaplanması için RapidMiner Studio yazılımı kullanılmıştır. Hesaplanan değerler Çizelge 4.2'de verilmiştir.

Davies Bouldin değerinin küçük olması küme kalitesinin iyi olduğunu gösterir. Değerin büyük ya da küçük olduğunun söylenebilmesi için en az iki senaryo halinde kümeleme işleminin yapılması ve her bir senaryo için indeks değerinin hesaplanması gerekir (Silahtaroğlu, 2013).

Çizelge 4.2. Hesaplanan Davies Bouldin İndeksi Değerleri

Küme Sayısı	DB İndeksi
k = 3	0.183
k = 4	0.109
k = 5	0.154
k = 6	0.137
k = 7	0.182

Elde edilen Davies Bouldin indeks değerlerine göre küme sayısının “k = 4” olarak seçilmesi gerektiği ve uygulama için küme sayısının algoritmaya tanıtılmasında bu değerin kullanılması gerektiği görülmektedir. Elde edilen değerlere bakıldığında en uygun küme sayısı olan, dört adet küme için hesaplanan değerin diğer küme sayıları için hesaplanan değerlerden küçük ve diğerlerine göre sıfıra daha yakın bir değer olduğu görülmektedir. Bu da karşılaştırmalı olarak en uygun küme sayısının belirlenmesi için oldukça önemli bir noktadır.

4.1.5. Verilerin kümelenmesi

Yapılan hazırlık aşamalarından sonra herhangi bir veri uyuşmazlığı yoksa kümeleme analizi işlemine geçilebilir. Kümeleme analizinden önceki hazırlık aşamaları yapılacak analizin doğru sonuçlar vermesi ve sonuçların yorumlanabilir olması yönünden oldukça önemlidir.

K-Means yönteminin uygulaması için QGIS yazılımının bir eklentisi olan Attribute Based Clustering kullanılmıştır. Bu eklenti açık kaynak kodlu olup yazılıma entegre edilebilmektedir. Bu sayede öznitelik verilerinin K-Means algoritması ile analizi yapılabilmektedir. Ara yüzü oldukça sade olarak tasarlanmış olan eklenti entegre edilen verilerin normalizasyonu aşamasını da yapabilmektedir. Uygulama için Z-Sayısı yöntemi seçildiğinden öznitelik verisi olarak uygulama özelinde standartlaştırılmış veriler seçilmiştir. Aynı zamanda K-Means yöntemi için yineleme sayısı da uygulama öncesi algoritmaya tanımlanabilmektedir.

4.1.6. Kümelerin yorumlanması

Kümeleme analizi yapıldıktan sonra üzerinde çalışılan konu ile sonuç olarak çıkan kümelerin birbirleri ile ilişkileri değerlendirilmelidir. Üzerinde çalışılan uygulamanın amacı, K-Means yönteminde en uygun küme sayısı olmakla birlikte ortaya çıkan verilerin anlamlılığının yorumlanması gerekmektedir.

Üzerinde çalışılan konunun amacı, il düzeyinde km^2’ye düşen araç sayısına (taşıt yoğunluğu) karşılık satışı gerçekleştirilen akaryakıt miktarıdır (m^3). En uygun küme sayısına bakıldığında “k = 4” olarak ortaya çıkan değere göre yapılan uygulama düşünüldüğünde elde edilen her bir küme verilerinin ortalamaları alınarak küme bazında sağlıklı bir değerlendirme yapılması sağlanmıştır.

Elde edilen istatistiksel veriler ile oluşturulan motorlu kara taşıt sayısının yüz ölçümüne (km^2) oranı ile buna karşılık satışı gerçekleştirilen akaryakıt miktarı (m^3) verileri yardımıyla elde edilecek kümeler bazında bir kullanım değerlerinin ortaya çıkarılması amaçlanmış ve kümelere tekabül eden illerin akaryakıt kullanım verilerinin

ortaya çıkarılması sağlanmıştır. Uygulama dört adet küme ile sonuçlandığından iller bazında dört adet grup oluşmuştur.

Böylesi bir değerlendirme sonucunda küme gruplarındaki illerin akaryakıt kullanım oranları dikkate alındığında elde edilen kümelerin azdan çoğa doğru sıralanması 2, 4, 3, 1 şeklindedir. Burada baz alınan veri küme bazında benzin ve motorin kullanım oranına göredir. Taşıt sayısının yüz ölçümüne (km^2) oranı düşünüldüğünde, kümeler bazında taşıt yoğunluğu baz alınırsa azdan çoğa doğru yine 2, 4, 3, 1 şeklinde sıralama olmaktadır. Burada her bir küme için düşünülürse km^2'ye düşen araçların akaryakıt kullanım alışkanlığındaki değişimden kaynaklanmaktadır. Bu noktada akaryakıt kullanım miktarının taşıt yoğunluğuna oranı düşünülürse kümeler bazında araç yoğunluğuna göre akaryakıt kullanımı miktarından bağımsız veriler elde edilmiş olacaktır. Yani küme bazında ortalama km^2'ye düşen bir aracın kullandığı akaryakıt miktarı belirlenmiş olur. Bu da bize küme bazında akaryakıt kullanım alışkanlığının belirlenmesini sağlamaktadır. Bu veriler ışığında kümeler akaryakıt kullanım alışkanlığı açısından azdan çoğa doğru 1, 4, 2, 3 şeklinde sıralanabilir.

İllerin kümelere göre dağılımı Çizelge 4.3'de, kümeler bazında ortalama taşıt yoğunluğu ve buna karşılık kullanılan akaryakıt miktarı Çizelge 4.4'de verilmiştir.

Çizelge 4.3. K-Means Algoritmasına göre oluşturulan kümeler (k = 4)

Küme 1	İstanbul.
Küme 2	Adıyaman, Afyonkarahisar, Ağrı, Amasya, Artvin, Bilecik, Bingöl, Bitlis, Bolu, Burdur, Çanakkale, Çankırı, Çorum, Diyarbakır, Elazığ, Erzincan, Erzurum, Eskişehir, Giresun, Gümüşhane, Hakkari, Isparta, Kars, Kastamonu, Kayseri, Kırklareli, Kırşehir, Konya, Kütahya, Malatya, Kahramanmaraş, Mardin, Muş, Nevşehir, Niğde, Ordu, Rize, Siirt, Sinop, Sivas, Tokat, Tunceli, Şanlıurfa, Uşak, Van, Yozgat, Aksaray, Bayburt, Karaman, Kırıkkale, Batman, Şırnak, Bartın, Ardahan, Iğdır, Karabük, Kilis.
Küme 3	Ankara, Bursa, Gaziantep, Hatay, İzmir, Kocaeli, Yalova.
Küme 4	Adana, Antalya, Aydın, Balıkesir, Denizli, Edirne, Mersin, Manisa, Muğla, Sakarya, Samsun, Tekirdağ, Trabzon, Zonguldak, Osmaniye, Düzce.

Çizelge 4.4. 2019 yılına ait benzin ve motorin kullanım miktarı ve araç yoğunluğu.

Küme	Taşıt Yoğunluğu	Kullanılan Akaryakıt (m^3/km^2)
1	767	930
2	12	17
3	93	142
4	44	56

Araç yoğunluğu fazla olan kümelere gidildikçe akaryakıt kullanımının doğru orantılı olarak artacağı düşünülebilir. Ancak kümelerin akaryakıt kullanım alışkanlığı verisi de düşünülürse orantı bu şekilde olmayacaktır. Ek1'deki iki veri grubunun oranlanmasıyla kümeler bazında ortalama akaryakıt kullanımı alışkanlığı verisi elde edilecektir. Buna göre, 1 numaralı kümeye bakılırsa burada İstanbul hem akaryakıt kullanımı hem de araç yoğunluğu verisi dikkat çekici bir şekilde diğer illerden fazla olduğu için ayrı bir küme olarak karşımıza çıkmıştır. Ancak akaryakıt kullanım alışkanlığı verisi düşünülürse araç yoğunluğuna göre daha az akaryakıt kullanan küme olarak karşımıza çıkacaktır. Yüz ölçümüne (km^2) göre taşıt yoğunluğu sıralamasında üçüncü sırada olan üç numaralı kümeye bakıldığında ise akaryakıt kullanım alışkanlığı açısından ilk sırada yer aldığı görülmektedir. Akaryakıt kullanım miktarının araç yoğunluğu verisine oranlanması ile elde edilen kümeler bazında ortalama akaryakıt kullanım alışkanlığına ait veriler Çizelge 4.5'de verilmiştir. Buna göre akaryakıt kullanım alışkanlıkları açısından en az kullanım oranına sahip küme 1 iken en çok kullanım oranına sahip kümenin 3 olduğu görülmektedir.

Çizelge 4.5. 2019 yılına ait kümeler bazında ortalama akaryakıt kullanım alışkanlığı değeri.

Küme	Ort. Akaryakıt Kullanım Değeri (m^3)
1	1.21
2	1.42
3	1.53
4	1.27

4.1.7. Tematik harita üretimi

Bu uygulama ile elde edilmiş kümeler bir koroplet harita ile görselleştirilmiştir.

Üretilen tematik haritaların datumu ED50 ve Türkiye'de küçük ölçekli haritalar için kullanılan Lambert konform konik projeksiyonu, ölçeği ise 1/6000000 seçilmiştir.

Tematik harita üretiminde seçilen renkler, kartografya için renk tavsiyesi veren bir web tabanlı uygulama olan Color Brewer 2.0 uygulaması aracılığıyla seçilmiştir (URL16). Çok değişkenli harita yapımı için renk seçiminde nitelik ön plana çıkarılmıştır.

K-Means algoritması ile yapılan kümeleme analizi işlemi QGIS yazılımının bir eklentisi olan Attribute Based Clustering yardımıyla verilerin elde edilmesi sağlanmıştır. Uygulama sonucunda QGIS yazılımı ile harita üretimi yapılmıştır. Üretilen haritalar Ek3-a, Ek3-b, Ek3-c, Ek3-d ve Ek3-e'de verilmiştir.

4.2. Uygulama 2

4.2.1. Konu seçimi ve verilerin hazırlanması

Son yıllarda artan su kullanımı ve oluşan talebi karşılayacak içme ve kullanma suyu miktarının sınırlı miktarda olması bireylerin su tüketim alışkanlıklarını gözden geçirmeleri gerektiğini ortaya koymaktadır. Düzenlenmiş veriler iller düzeyinde olup 2018 yılına ait ortalama kişi başı tüketilen su miktarını (m^3) ve kişi başı yıllık arz edilen su miktarını (m^3/kişi-yıl, 2018) göstermektedir.

Verilerin temini için Türkiye İstatistik Kurumu'nun sitesi (www.tuik.gov.tr) kullanılmıştır. Kümeleme analizi buradan elde edilen toplam il nüfusları (2018), dağıtılan suyun abone sayısı (2018), dağıtılan su miktarı (m^3/yıl, 2018) ve kişi başı arz edilen günlük su miktarı (litre/kişi-gün, 2018) verilerinin düzenlenmesiyle oluşturulan uygulama için gerekli veriler ile yapılmıştır.

Toplam il nüfus verileri (2018) Ek2-a, il bazında dağıtılan suyun abone sayısı (2018) Ek2-b, dağıtılan su miktarı (m^3/yıl, 2018) Ek2-c, kişi başı arz edilen günlük su miktarı (litre/kişi-gün, 2018) Ek2-d'de verilmiştir. Verilerin düzenlenmesi ve birbirlerine oranlanması ile oluşturulan ve kümeleme analizi için kullanılacak ortalama kişi başı tüketilen su miktarı (m^3, 2018) Ek2-e, kişi başı yıllık arz edilen su miktarı (m^3/kişi-yıl, 2018) Ek2-f'de verilmiştir.

4.2.2. Verilerim normalize edilmesi

Ortalama kişi başı tüketilen su miktarı (m^3, 2018) verisi için hesaplanan ortalama 44 ve standart sapma 12'dir.

Kişi başı yıllık arz edilen su miktarı (m^3/Kişi, 2018) verisi için hesaplanan ortalama 86 ve standart sapma 24'dür.

Bu iki istatistiksel verideki değerlerin her biri için ortalama ve standart sapma değerleri kullanılarak z-sayısı değerleri hesaplanmış ve artık kümeleme analizi için kullanılacak verilerin normalizasyon işlemi tamamlanmıştır. Z-sayısı yöntemi ile normalize edilen veriler Çizelge 4.6'da verilmiştir.

Çizelge 4.6. Z-Sayısı yöntemi ile verilerin normalize edilmesi

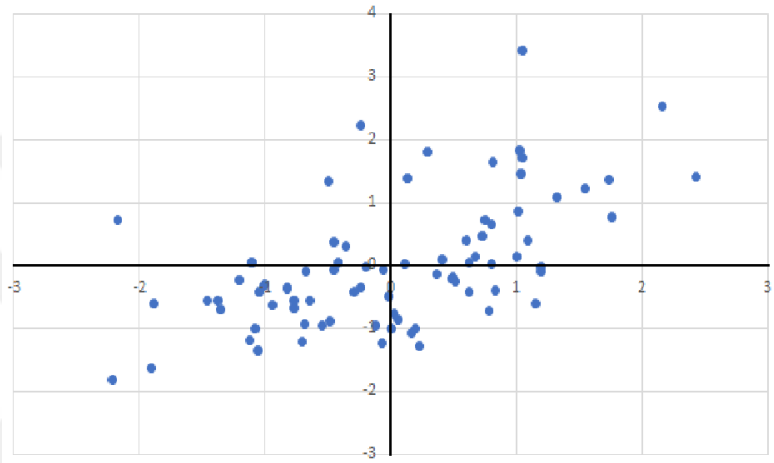
	X	Y		X	Y
P.Kodu	**z-sayısı**	**z-sayısı**	P.Kodu	**z-sayısı**	**z-sayısı**
1	0.48675	-0.21927	42	0.83994	-0.38697
2	-0.68061	-0.93580	43	-0.54472	-0.95105
3	-0.01350	-0.49369	44	1.74213	1.36625
4	-0.45317	0.36005	45	0.23161	-1.27120
5	0.59585	0.39055	46	1.02738	1.83886
6	0.62426	0.03990	47	-2.17643	0.72594
7	2.42505	1.41199	48	2.16294	2.54015
8	-1.20180	-0.23452	49	0.13607	1.39674
9	1.15187	-0.61565	50	0.51879	-0.24976
10	1.32145	1.09183	51	-0.23304	-0.35648
11	-0.70416	-1.21022	52	-1.35434	-0.70712
12	-0.05550	-0.08206	53	-1.37458	-0.55467
13	1.04773	1.71690	54	1.03232	1.45772
14	-1.11516	-1.19498	55	0.67899	0.14662
15	0.10977	0.00941	56	-0.66961	-0.09731
16	0.19468	-0.99679	57	-1.10183	0.05515
17	-0.76454	-0.67663	58	0.75225	0.71070
18	-0.35725	0.29907	59	0.03219	-0.76810
19	-0.29128	-0.43271	60	-0.42241	0.05515
20	1.00257	0.13137	61	0.29362	1.80837
21	-1.05475	-1.36268	62	0.80686	0.64972
22	-0.82642	-0.34123	63	0.00444	-0.99679
23	1.01381	0.84791	64	-0.47783	-0.89007
24	0.79842	0.02466	65	-0.81992	-0.37172
25	0.81730	1.65591	66	0.73159	0.46677
26	0.16525	-1.07301	67	-1.11102	0.05515
27	0.49585	-0.18878	68	-0.76438	-0.56991
28	-1.46350	-0.56991	69	-0.44825	-0.08206
29	-1.00198	-0.29550	70	-1.07275	-1.01203
30	-2.21481	-1.82004	71	-0.06677	-1.24071
31	0.06434	-0.87482	72	-0.49043	1.33576
32	1.08640	0.39055	73	-0.19228	-0.03633
33	0.36950	-0.14304	74	-1.88187	-0.60041
34	0.77705	-0.72237	75	-0.23938	2.21999
35	0.61928	-0.43271	76	-1.90599	-1.62185
36	1.04945	3.42438	77	1.55041	1.22904
37	-1.05056	-0.43271	78	1.75957	0.77168
38	1.19522	-0.09731	79	-1.01127	-0.32599
39	-0.64525	-0.56991	80	-0.12431	-0.96629
40	0.41490	0.10088	81	-0.94464	-0.63090
41	1.19444	-0.02108			

X: Ortalama kişi başı tüketilen su miktarı (m^3, 2018); **Y:** Kişi başı yıllık arz edilen su miktarı (m^3/Kişi-2018)

4.2.3. Korelasyon katsayısının hesaplanması

2018 yılına ait olan ortalama kişi başı tüketilen su miktarı (m^3) ve kişi başı yıllık arz edilen su miktarı (m^3/Kişi) verilerinin korelasyon katsayısı r = 0.57 olarak hesaplanmıştır. Bu da verilerin birbirleriyle orta düzeyde korelasyon olduğunu ve kümeleme analizi için birbirleriyle ilişkileri bakımından bir sorun teşkil etmediğini göstermektedir.

Şekil 4.2'de normalize edilmiş verilerin grafik gösterimi yer almaktadır. Bu grafikten de verilerin birbirleriyle olan ilişkilerine göz atılabilir.



Şekil 4.2. Normalize edilmiş verilerin grafik gösterimi

4.2.4. Küme sayısının belirlenmesi

Kullanılacak olan en optimum küme sayısının (k) belirlenmesi için veriler 3, 4, 5, 6 ve 7'li kümelere ayrılmış ve oluşan kümelerin Davies Bouldin (DB) indeksleri hesaplanmıştır. Hesaplanan değerler Çizelge 4.7'de verilmiştir.

Çizelge 4.7. Hesaplanan Davies Bouldin İndeksi Değerleri

Küme Sayısı	DB İndeksi
k = 3	0.409
k = 4	0.470
k = 5	0.505
k = 6	0.183
k = 7	0.417

Elde edilen değerlere göre küme sayısı "k=6" olarak belirlenmiştir.

4.2.5. Verilerin kümelenmesi

Yapılan hazırlık çalışmaları soncunda herhangi bir veri uyuşmazlığı olmadığı takdirde kümeleme analizi işlemine geçilir. Yine ilk uygulamada olduğu gibi kümeleme

analizi için K-Means yöntemi kullanılmıştır. Yöntemin uygulaması için QGIS yazılımının bir eklentisi olan Attribute Based Clustering kullanılmıştır.

Uygulamada 3, 4, 5, 6 ve 7 küme sayısına sahip kümeleme analizi yapılmıştır. QGIS yazılımında ortaya çıkan veriler, haritaya uygulanarak görüntülenebilmektedir.

4.2.6. Kümelerin yorumlanması

Üzerinde çalışılan uygulamanın amacı, Türkiye'deki su tüketim alışkanlıkları üzerine olduğundan ortaya çıkan altı küme üzerinden veri gruplarının ortalamaları alınmıştır. Böylece küme bazında daha sağlıklı bir değerlendirme yapılabilmektedir.

Çeşitli istatistiksel verilerin elde edilmesi ile oluşturulan 2018 yılına ait Ortalama Kişi Başı Tüketilen Su Miktarı (m^3) ve 2018 yılına ait Kişi Başı Yıllık Arz Edilen Su Miktarı (m^3/kişi-yıl) verileri yardımıyla bireylerin tüketimine göre artan su (m^3) ve arz edilen su miktarına göre tüketilen su (yüzde) değerleri hesaplanmıştır. Bu verilerin küme bazında değerlendirilebilmeleri için ortalama değerler elde edilmiştir. Ortalama tüketilen su miktarı (m^3) verisi ile illerin su kullanımı alışkanlıklarına göre gruplanması sağlanmıştır. Hesaplanan kullanım yüzdesi ile de küme bazında arz edilen suyun tüketim değerleri görülmüştür.

Değerlendirme sonucunda küme gruplarındaki illerin su tüketim alışkanlıkları azdan çoğa doğru olarak sırasıyla 2, 6, 5, 1, 3, 4 şeklinde olmuştur. Su artıran iller düşünüldüğünde ise en çoktan en aza doğru kümeler 4, 5, 6, 3, 2, 1 şeklinde sıralanır. İllerin kümelere göre dağılımı Çizelge 4.8'de verilmiştir.

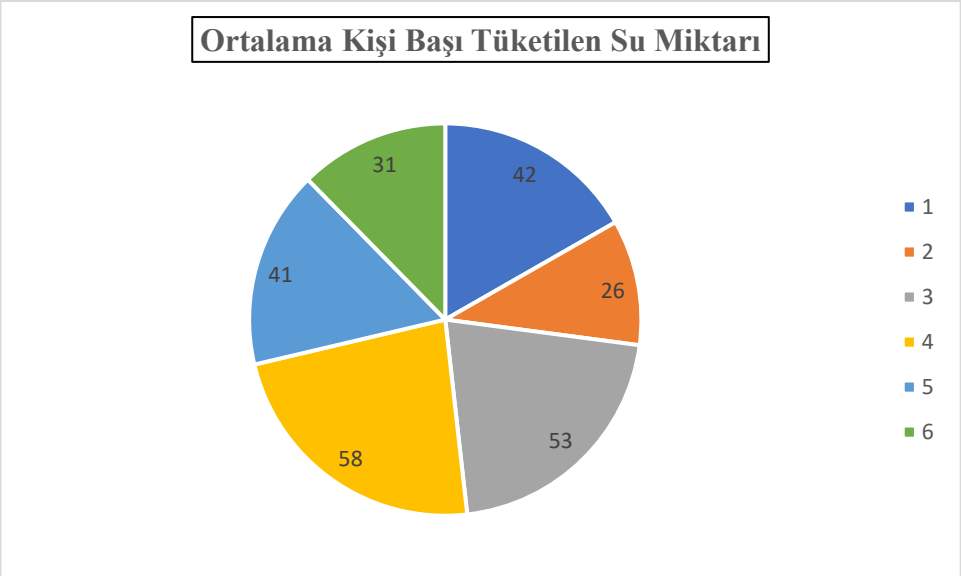
Çizelge 4.8. K-Means Algoritmasına göre oluşturulan kümeler (k = 5)

Küme 1	Adıyaman, Afyonkarahisar, Bilecik, Bursa, Çorum, Eskişehir, Hatay, Kütahya, Manisa, Niğde, Tekirdağ, Şanlıurfa, Uşak, Kırıkkale, Osmaniye.
Küme 2	Bolu, Diyarbakır, Hakkari, Karaman, Bartın, Iğdır.
Küme 3	Adana, Amasya, Ankara, Aydın, Denizli, Erzincan, Gaziantep, Isparta, Mersin, İstanbul, İzmir, Kayseri, Kırşehir, Kocaeli, Konya, Nevşehir, Samsun, Sivas, Tunceli, Yozgat.
Küme 4	Antalya, Balıkesir, Bitlis, Elazığ, Erzurum, Kars, Malatya, Kahramanmaraş, Muğla, Muş, Sakarya, Trabzon, Ardahan, Yalova, Karabük.
Küme 5	Ağrı, Bingöl, Burdur, Çankırı, Tokat, Bayburt, Batman, Şırnak.
Küme 6	Artvin, Çanakkale, Edirne, Giresun, Gümüşhane, Kastamonu, Kırklareli, Mardin, Ordu, Rize, Siirt, Sinop, Van, Zonguldak, Aksaray, Kilis, Düzce.

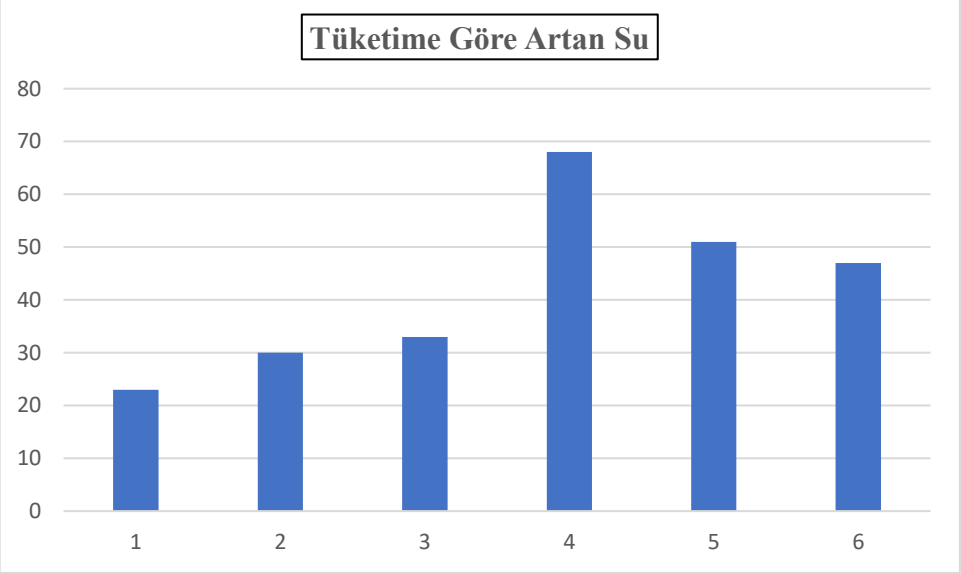
Belediyelerin kullanılması için arz ettiği su miktarı ve bireylerin tükettiği miktar verileri ayrı ayrı dikkate alınıp küme bazında tüketim yüzdesi oluşturulmuştur. Tüketim yüzdesi baz alındığında kümeler çoktan aza doğru sıralanırsa; 1. küme %65, 3. küme

%62, 2. küme %46, 4. küme %46, 5. küme%45, 6. küme %40 olarak sıralanır. Burada 1.Kümede yer alan illerin arz edilen suya göre tüketimlerinin en fazla olduğu görülmektedir. Arz edilen suya göre en az tüketen ise 6.Kümedeki illerdir. Kişi başı tüketim alışkanlığı düşünüldüğünde ise en fazla su tüketen küme 4. küme, en az su tüketen küme ise 2. kümedir. Arz edilen suya göre tasarruf sağlayan yani en fazla su artıran grup 4. küme, en az su artıran grup ise 1. kümede yer almaktadır.

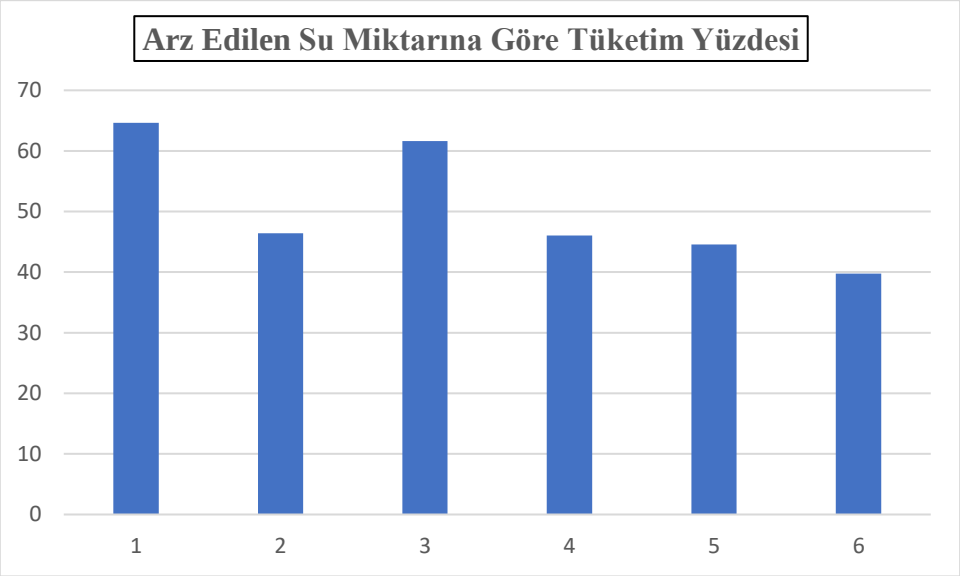
Küme bazında ortalama kişi başı tüketilen su miktarı Şekil 4.3'de, kullanıma göre artan su miktarı (m^3) Şekil 4.4'de ve arz edilen su miktarına (m^3) göre tüketilen su yüzdesi Şekil 4.5'de yer almaktadır.



Şekil 4.3 Küme bazında tüketilen ortalama su miktarı (m^3) pasta grafiği



Şekil 4.4. Küme bazında arz edilen su miktarına göre artan su miktarı (m^3)



Şekil 4.5. Küme bazında arz edilen su miktarına göre tüketim yüzdesi

Uygulama 1'de yapıldığı gibi kümeleme analizi ile birlikte ortaya çıkan sonuçların tematik haritalar ile gösterilmesi sağlanmıştır. Aynı şekilde veriler ile birlikte ortaya çıkan sonuçlar iki değişkenli tematik haritaların üretilmesine olanak sağlamıştır.

4.2.7. Tematik harita üretimi

Bu uygulama ile elde edilmiş kümeler bir koroplet harita ile görselleştirilmiştir. Uygulama 1 ile aynı projeksiyon ve ölçekte çalışılmıştır.

Üretilen haritalar Ek4-a, Ek4-b, Ek4-c, Ek4-d ve Ek4-e'de verilmiştir.

5. SONUÇLAR VE ÖNERİLER

Elde edilen istatistiksel verilerin kümeleme analizi yöntemiyle kümelere ayrılmasıyla birlikte son durumda verilerin artık daha ayırt edilebilir hale geldiği, yorumlanmasında kolaylık sağlandığı, son kullanıcıya ulaşacak olan tematik haritadaki kullanımında bilgiyi aktarma yönünden başarılı sonuçlar sağlanmıştır.

Bu çalışmanın amacı, kümeleme analizi yöntemlerinden olan K-Means Algoritması ile uygulamalar yapılması, algoritma çalışma prensiplerinin araştırılması ve K-Means kümeleme analizi yönteminin başlangıç aşamasında önemli kısımlarından birisi olan "en uygun küme sayısı" değerinin belirlenmesinde dikkat edilmesi gereken işlemlerin uygulamada gösterilmesidir. K-Means Algoritması, çalışma sistemi gereği, küme sayısının kullanıcı tarafından belirlenmesi gereken bir yöntemdir. Ancak küme sayısı belirlenirken kümeleme analizi aşamaları göz önünde bulundurulmalı, analiz için gerekli işlemler sırası ile yapılmalıdır. K-Means yönteminde küme sayısının belirlenmesi için çeşitli indeks değerleri mevcut olup uygulama özelinde DB değeri kullanılarak sonuca ulaşılmıştır. Değerlendirme sonucunda uygulama için doğru küme sayısı elde edilmiş ve çalışma bu doğrultuda gerçekleşmiştir.

Tematik harita üretiminde ortaya çıkan verilerin okunabilirliği göz önüne alınarak küme sayıları 3 ile 7 arasında seçilmiştir. Değerlendirmenin bu küme sayılarında yapılmasındaki amaç ortaya çıkan haritanın tematik harita üretimi açısından anlamlı olmasıdır. Küme sayısının daha fazla seçilmesi, ortaya çıkan sonucun aktarılmasını oldukça zorlaştıracaktır. Küme sayısının daha az seçilmesi ise verilerin değerlendirilmesini olumsuz etkileyecektir.

Seçilen her "k" değeri için tematik haritalar üretilerek illerin kümeler arası değişimi gözlemlenmiştir. DB değeri ile değerlendirme sonucunda en uygun küme sayısı belirlenerek uygulama sonuçları bu doğrultuda değerlendirilmiştir. Amaç, kümelerin en uygun düzeyde birbirinden ayrışmasını sağlamaktır. İlk uygulamada en iyi sonuç dört küme sayısında bulunurken ikinci uygulamada ise altı küme sayısında en iyi sonuca ulaşılmıştır. Bu sayede sonuçlara göre elde edilen verilerden daha anlamlı çıkarımlarda bulunulmuştur.

Daha az hacimli istatistiksel veriler söz konusu olduğunda bile en uygun küme sayısının belirlenmesi için geçerlilik indeks değerine başvurulabilir. Farklı sonuçlar ortaya çıkabilir ve bu da sonuç verilerin değerlendirilmesinde aktif rol oynayacaktır. Büyük hacimli verilerin değerlendirilmesi küme sayısı belirlemede yorumdan bağımsız

olduğundan mutlaka geçerlilik indeks değerleri hesaplanmalı ve çalışma yine bu doğrultuda sonuçlandırılmalıdır.

Kümeleme analizi yöntemleri çok değişkenli harita üretimi açısından önemli bir konuma sahiptir. Analiz yöntemleri ile birlikte birden çok konuya sahip istatistiksel verilerin birlikte değerlendirilmesi sağlanmakta ve çeşitli çıkarımlarda bulunulabilmektedir. Bu da istatistiksel araştırma yapan ve kapsamlı çıkarımlara gereksinim duyan şahıs ve kurumlar için kümeleme analizi yöntemlerinin ayrı bir öneme sahip olmasını sağlamaktadır.

Yapılan birinci çalışma için seçilen ve günümüzde hava kirliliği açısından belirli bir tehdit seviyesine ulaşan akaryakıt kullanımı ve araç yoğunluğu konularının önemine değinilmiştir. Belirlenen kümeler ile birlikte küme bazında çeşitli sonuçlara ulaşılmış, sonuç veriler ışığında çıkarımlarda bulunulmuş ve bu sonuçları değerlendirecek kurum ve kuruluşların bilgisine sunulmuştur. İkinci çalışmada ise gösterilmeye çalışılan konu Türkiye'nin iller bazında kişi odaklı su kullanım alışkanlıklarıdır. Elde edilen son verilere ve tematik haritaya bakıldığında kümeler bazında hangi illerin tasarruf ettiği hangi illerin daha fazla su sarfiyatı yaptığına dair ortalama olarak kişi bazında bir bilgi elde edilebiliyor. Verilere göre daha az su kullanmaya teşvik edilmesi gereken iller görülmektedir.

Sonuç olarak anlamlı küme sayısı uygulamadan uygulamaya değişmekte olup DB indeksinin önemi bu çalışmada görülmüştür. Gelecekteki çalışmalarda K-Means yönteminin DB indeksi ile dikkate alınarak uygulanması önerilir.

KAYNAKLAR

Akat, Y., 2007, Ülkelerin askeri benzerliklerine göre kümeleme analizi yardımıyla sınıflandırılması, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Altınok Y., 2019, Veri madenciliğinde hiyerarşik kümeleme algoritmalarının uygulamalı karşılaştırılması, Yüksek Lisans Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Anderson, E., 1960, A semigraphical method for the analysis of complex problems, *Technometrics* 2(3), 387-391.

Bhatia, S. K., 2004, Adaptive k-means clustering, In FLAIRS conferance, 695-699.

Bildirici, İ. Ö., 2018, Kartografya, Nesibe Necla ULUĞTEKİN, *Atlas Akademi Yayınevi,* Konya.

Bildirici, İ. Ö., ve Afacan, N., 2018, Kümeleme analizi sonuçlarının tematik haritalar ile görselleştirilmesi, *Geomatik Dergisi*, 3(1), 92-99.

Bircik, Ö. F., 2019, Erzurum kent içi ana toplu taşıma sistemlerinin kümeleme analizi yöntemiyle incelenmesi, Yüksek Lisans Tezi, Atatürk Üniversitesi Fen Bilimleri Enstitüsü, Erzurum.

Blashfield, K. R., Aldenferder, M. S., 1978, The literatüre on cluster analysis, multivariate behavioral research, Vol. 13, Pages 271-295.

Blashfield, K. R., Harvey A. Skinner, A. H., Morey, C. L., 1983, A comparison of cluster analysis techniques withing a sequential validation framework, multivariate behavioral research, Vol. 18, Pages 309-329.

Brewer, C. A., and Pickle, L., 2002, Evaluation of methods for classifying epidemiological data on choropleth maps in series, *Annals of the Association of American Geographers*, 92(4), 662-681.

Carr, D. B., Olsen, A. R., and White, D., 1992, Hexagon mosaic maps for display of univariate and bivariate geographical data, *Cartography and Geographic Information System*, 19(4), 228-236, 271.

Chiu, T., Fang, D., Chen, J., Wang, Y., Jeris, C., 2001, A robuts and scalable clustering algorithm for mixed type attributes in large database enviroment, *Proceeding of the Seventh ACM SIGKDD International Conferance on Knowledge Discovery and Data Mining,* San Francisco, California, USA, 263-268.

Cox, D. J., 1990, The art of scientific visualization, *Academic Computing*, 4(6), 20-22, 32-34, 36, 38, 40.

Çağlar, B., 2018, Mekansal verilerin kümeleme analizi ile değerlendirilmesi, Yüksek Lisans Tezi, Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü, Konya.

Çakmak, Z., 1999, Kümeleme analizinde geçerlilik problemi ve kümeleme sonuçlarının değerlendirilmesi, Dumlupınar Üniversitesi, Sosyal Bilimler Dergisi (3), Kütahya, 187-203.

Davies, David L., Bouldin, Donald W., 1979, A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224-227.

Dinçer, Ş. E., 2006, Veri madenciliğinde k-means algoritması ve tıp alanında uygulaması, Yüksek Lisans Tezi, *Kocaeli Üniversitesi Fen Bilimleri Enstütüsü,* Kocaeli, 24.

Dorling, D., 1993, Map design for census mapping, *The Cartographic Journal*, 30(2), 167-183.

Ellson, R., 1990, Visualization at work, *Academic Computing,* 4(6), 26-28, 54-56.

Erilli, N, A., 2012, Bulanık sayıların bulanık kümeleme analizinde kullanımı ve satranç oyuncularının sınıflandırılması, Doktora Tezi, Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü, Samsun, ii.

Eyton, J. R., 1984, Complementary-color two-variable maps, *Annals of the Association of American Geographers,* 74(3), 477-49.

Frawley, W. J., Piatetsky-Shapiro, G. and Matheus, C. J., 1992, Knowledge discovery in database: an overview. *AI Magazine,* v13.3, 55-70.

Gülcemal, M. E., 2019, OECD ülkelerinin sigorta Pazar paylarının çok değişkenli istatistiksel yöntemlerle incelenmesi, Yüksek Lisans Tezi, Kütahya Dumlupınar Üniversitesi Sosyal Bilimler Enstitüsü, Kütahya.

Güneş, M. Ş., 2017, Alternatif kümeleme yöntemlerinin karşılaştırmalı analizi ve bir uygulama, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Han, J., Pei, J., and Kamber, M., 2012, Data mining: concept and techniques, *Elsevier,* 225 Wyman Street, Waltham, MA 02451 USA.

Hartnett, S., 1987, Employing rectangular point symbols in two-variable maps, *Association of American Geographers Annual Meeting,* Portland, OR, 39.

Imhof, E., 2007, Cartographic relief presentation, *ESRI Press,* Redlands, California.

Işık, M., Çamurcu, A. Y., 2007, K-means, k-medoids ve bulanık c-means algoritmalarının uygulamalı olarak performanslarının tespiti, *İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü*, İstanbul.

Jenks, G. F., 1953, Pointilism as acartographic technique, *The Professional Geographer,* 5(5), 2-6.

Kalaycı, Ş., 2010, SPSS uygulamalı çok değişkenli istatistik teknikleri, *Asil Yayın Dağıtım,* Ankara.

Karagöz, Y., 2016, SPSS 23 ve AMOS 23 uygulamalı istatistiksel analizler, *Nobel Akademik Yayıncılık,* Ankara, 891.

Karakuş, Y. İ., 2020, Kentiçi trafik kazalarının kümeleme analizi ve lojistik regresyon modeli ile incelenmesi, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Kırmızıbiber, A., 2019, Hiyerarşik ve k-ortalamalar yöntemleriyle grid noktalarının kümelenmesi, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Lloyd, R. E., and Steinke, T. R., 1976, The decisionmaking process for judging the similarity of chorophleth maps, *The American Cartographers,* 3(2), 177-184.

Lloyd, R., and Steinke, T., 1977, Visual and statistical comparison of choropleth maps, *Annals of the Association of America Geographers,* 67(3), 429-436.

Monmonier, M., 1993, Mapping it out: expository cartography fort he humanities and social sciences, *University of Chicago Press,* Chicago, IL, 166.

Muller, J. C., 1980, Visual comparison of continuously shaded maps, *Cartographica 17,* 1, 4-51.

Na, S., Xumin, L., Yang, G., 2010, Research on k-means clustering algorithm: An improved k-means clustering algorithm, In 2010 Third International Symposium on intelligent information technology and security informatics, IEEE, China, 63-67.

Nacaroğlu, E., 2010, Deprem etkisiyle oluşan boru hasarlarının coğrafi bilgi sistemleri (cbs) ve kümeleme analizi ile değerlendirilmesi, Yüksek Lisans Tezi, Pamukkale Üniversitesi Fen Bilimleri Enstitüsü, Denizli.

Olson, J. M., 1972b, The effect of class interval system on chropleth map correlation, *Cartographica: The International Journal for Geographic Information and Geovisualization,* 9(1), 44-49.

Olson, J. M., 1981, Spectrally encoded two-variable maps, *Annals of the Association of America Geographers,* 71(2), 259-276.

Olson, J., 1972a, Class interval systems on maps of observed correlated distributions, *Cartographica: The International Journal for Geographic Information and Geovisualization,* 9(2), 122-131.

Orhunbilge, N., 2010, Çok değişkenli istatistik yöntemler, *İstanbul Üniversitesi İşletme Fakültesi Yayınları,* İstanbul.

Özdemir, A., Aslay, F. Y., Çam, H., 2010, Veritabanında bilgi keşfi süreci: Gümüşhane devlet hastanesi uygulaması, Sosyal Ekonomik Araştırmalar Dergisi, 348.

Peterson, M. P., 1979, An evaluation of unclassed crossed-line choropleth mapping, *The American Cartographer,* 6(1), 21-37.

Sayılan, A. B., 2019, Veri madenciliğinde bazı kümeleme algoritmalarının karşılaştırılması, Yüksek Lisans Tezi, Muğla Sıtkı Koçman Üniversitesi Fen Bilimleri Enstitüsü, Muğla.

Silahtaroğlu, G., 2013, Veri madenciliği (Kavram ve algoritmaları), Papatya Yayıncılık, İstanbul.

Sinaga, K. P., Yang, M. S., 2020, Unsupervised k-means clustering algorithm, IEEE Access, Vol.8, 80716-80727.

Skillicorn, D., 2009, Knowledge discovery for counterterorism and law enforcement, Taylor and FrancisGroup, USA.

Slocum, T. A., McMaster, R., Kessler, F., and Howard, H., 2014, Thematic cartography and geovisualization, *Pearson Education Limited,* UK.

Tatlıdil, H., 2002, Uygulamalı çok değişkenli istatistiksel analiz, *Akademi Matbaası,* Ankara.

Uçar D., Uluğtekin, N., 2006, Kartografyaya giriş, *Yayınlanmamış ders notu,* İstanbul.

Wilhelm, S. D., 1983, Two symbols for unclassed two-variable mapping: the bivariate box and the ratio bar, Doctoral dissertation, *University of North Carolina,* Chapell Hill, NC.

Yılmaz, V., 2009, Türkiye akarsuları su kalitesi parametrelerinin çok değişkenli istatistiksel analiz yöntemleri ile incelenmesi, Yüksek Lisans Tezi, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Konya.

URL1:https://www.hkmo.org.tr/hakkimizda/meslegimiz/harita_nedir.php [Ziyaret tarihi: 5 Nisan 2021]

URL2: https://tr.wikipedia.org/wiki/K%C3%BCmeleme_analizi [Ziyaret tarihi 5 Nisan 2021]

URL3: https://en.wikipedia.org/wiki/Clustering [Ziyaret tarihi: 5 Nisan 2021]

URL4: https://tr.wikipedia.org/wiki/Veri_madencili%C4%9Fi [Ziyaret tarihi: 10 Nisan 2021]

URL5: https://tr.wikipedia.org/wiki/K-means_kümeleme [Ziyaret tarihi: 1 Nisan 2021]

URL6: https://en.wikipedia.org/wiki/Davies–Bouldin_index [Ziyaret tarihi: 6 Temmuz 2020]

URL7: https://pro.arcgis.com/en/pro-app/help/mapping/layer-properties/data classification-methods.htm#ESRI_SECTION1_1BDD383C17164B948BF546CEADDA70E9 [Ziyaret tarihi: 4 Mart 2020]

URL8: https://plugins.qgis.org [Ziyaret tarihi: 15 Şubat 2021]

URL9: https://plugins.qgis.org/plugins/attributeBasedClustering [Ziyaret tarihi: 15 Şubat 2021]

URL10: https://tr.wikipedia.org/wiki/RapidMiner [Ziyaret tarihi: 10 Aralık 2020]

URL11: https://rapidminer.com [Ziyaret tarihi: 10 Aralık 2020]

URL12:https://bigdata-madesimple.com/possibly-the-simplest-way-to-explain-k-means-algorithm [Ziyaret tarihi: 01.09.2021]

URL13: https://tr.wikipedia.org/wiki/Kovaryans [Ziyaret tarihi: 10 Haziran 2020]

URL14: https://tr.wikipedia.org/wiki/Korelasyon [Ziyaret tarihi: 4 Mart 2020]

URL15: https://en.wikipedia.org/wiki/K-means_clustering [Ziyaret tarihi: 1 Kasım 2021]

URL16: https://en.wikipedia.org/wiki/K-medoids [Ziyaret tarihi: 1 Kasım 2021]

EKLER

Ek1-a İl bazında motorlu kara taşıt sayısı (2019)

P.Kodu	İL	MKT (Toplam)	P.Kodu	İL	MKT (Toplam)
1	ADANA	657202	42	KONYA	724397
2	ADIYAMAN	105856	43	KÜTAHYA	210652
3	AFYONKARAHİSAR	223265	44	MALATYA	176389
4	AĞRI	31692	45	MANİSA	590671
5	AMASYA	117601	46	KAHRAMANMARAŞ	235383
6	ANKARA	2033935	47	MARDİN	75935
7	ANTALYA	1101056	48	MUĞLA	506829
8	ARTVİN	38933	49	MUŞ	33608
9	AYDIN	456056	50	NEVŞEHİR	121818
10	BALIKESİR	484475	51	NİĞDE	107255
11	BİLECİK	68934	52	ORDU	136111
12	BİNGÖL	17111	53	RİZE	80210
13	BİTLİS	22235	54	SAKARYA	287793
14	BOLU	114889	55	SAMSUN	359612
15	BURDUR	134884	56	SİİRT	20493
16	BURSA	902981	57	SİNOP	59475
17	ÇANAKKALE	231148	58	SİVAS	160963
18	ÇANKIRI	52617	59	TEKİRDAĞ	269373
19	ÇORUM	171377	60	TOKAT	179184
20	DENİZLİ	410598	61	TRABZON	197471
21	DİYARBAKIR	122143	62	TUNCELİ	9359
22	EDİRNE	160603	63	ŞANLIURFA	253375
23	ELAZIĞ	126387	64	UŞAK	137574
24	ERZİNCAN	59815	65	VAN	77311
25	ERZURUM	119108	66	YOZGAT	106682
26	ESKİŞEHİR	287142	67	ZONGULDAK	155221
27	GAZİANTEP	518415	68	AKSARAY	125951
28	GİRESUN	91540	69	BAYBURT	15589
29	GÜMÜŞHANE	24772	70	KARAMAN	91378
30	HAKKARİ	9188	71	KIRIKKALE	68142
31	HATAY	492988	72	BATMAN	44882
32	ISPARTA	178205	73	ŞIRNAK	29310
33	MERSİN	619418	74	BARTIN	51553
34	İSTANBUL	4187776	75	ARDAHAN	19192
35	İZMİR	1425302	76	IĞDIR	28069
36	KARS	45160	77	YALOVA	65991
37	KASTAMONU	130296	78	KARABÜK	65686
38	KAYSERİ	378771	79	KİLİS	47244
39	KIRKLARELİ	133401	80	OSMANİYE	164595
40	KIRŞEHİR	68323	81	DÜZCE	111587
41	KOCAELİ	399064			

MKT: Motorlu Kara Taşıt Sayısı

Ek1-b İl bazında yüz ölçümü (2019)

P.Kodu	İl	Yüz Ölçümü (km^2)	P.Kodu	İl	Yüz Ölçümü (km^2)
1	ADANA	13844	42	KONYA	40838
2	ADIYAMAN	7337	43	KÜTAHYA	11634
3	AFYONKARAHİSAR	14016	44	MALATYA	12259
4	AĞRI	11099	45	MANİSA	13339
5	AMASYA	5628	46	KAHRAMANMARAŞ	14520
6	ANKARA	25632	47	MARDİN	8780
7	ANTALYA	20177	48	MUĞLA	12654
8	ARTVİN	7393	49	MUŞ	8650
9	AYDIN	8116	50	NEVŞEHİR	8650
10	BALIKESİR	14583	51	NİĞDE	7234
11	BİLECİK	4179	52	ORDU	5861
12	BİNGÖL	8003	53	RİZE	3835
13	BİTLİS	8294	54	SAKARYA	4824
14	BOLU	8313	55	SAMSUN	9725
15	BURDUR	7175	56	SİİRT	5718
16	BURSA	10813	57	SİNOP	5718
17	ÇANAKKALE	9817	58	SİVAS	28164
18	ÇANKIRI	7542	59	TEKİRDAĞ	6190
19	ÇORUM	12428	60	TOKAT	10042
20	DENİZLİ	12134	61	TRABZON	4628
21	DİYARBAKIR	15168	62	TUNCELİ	7582
22	EDİRNE	6145	63	ŞANLIURFA	19242
23	ELAZIĞ	9383	64	UŞAK	5556
24	ERZİNCAN	11815	65	VAN	20921
25	ERZURUM	25006	66	YOZGAT	13690
26	ESKİŞEHİR	13960	67	ZONGULDAK	3342
27	GAZİANTEP	6803	68	AKSARAY	7659
28	GİRESUN	7025	69	BAYBURT	3746
29	GÜMÜŞHANE	6668	70	KARAMAN	8678
30	HAKKARİ	7095	71	KIRIKKALE	4791
31	HATAY	5524	72	BATMAN	4477
32	ISPARTA	8946	73	ŞIRNAK	7078
33	MERSİN	16010	74	BARTIN	2330
34	İSTANBUL	5461	75	ARDAHAN	4934
35	İZMİR	11891	76	IĞDIR	3664
36	KARS	10193	77	YALOVA	798
37	KASTAMONU	13064	78	KARABÜK	4142
38	KAYSERİ	16970	79	KİLİS	1412
39	KIRKLARELİ	6459	80	OSMANİYE	3320
40	KIRŞEHİR	6589	81	DÜZCE	2492
41	KOCAELİ	3397			

Ek1-c İl bazında benzin ve motorin (ton) tüketimi (2019)

P.Kodu	İL	Benzin+Motorin (Ton)	P.Kodu	İL	Benzin+Motorin (Ton)
1	ADANA	528242	42	KONYA	783896
2	ADIYAMAN	101284	43	KÜTAHYA	280255
3	AFYONKARAHİSAR	243440	44	MALATYA	151523
4	AĞRI	85108	45	MANİSA	565845
5	AMASYA	95247	46	KAHRAMANMARAŞ	218718
6	ANKARA	2189755	47	MARDİN	151120
7	ANTALYA	826487	48	MUĞLA	436264
8	ARTVİN	70013	49	MUŞ	44206
9	AYDIN	363524	50	NEVŞEHİR	123445
10	BALIKESİR	477016	51	NİĞDE	118946
11	BİLECİK	81380	52	ORDU	148983
12	BİNGÖL	45043	53	RİZE	89281
13	BİTLİS	41547	54	SAKARYA	301695
14	BOLU	171465	55	SAMSUN	432203
15	BURDUR	138331	56	SİİRT	65608
16	BURSA	960485	57	SİNOP	46850
17	ÇANAKKALE	213884	58	SİVAS	161615
18	ÇANKIRI	72539	59	TEKİRDAĞ	379495
19	ÇORUM	175177	60	TOKAT	162553
20	DENİZLİ	559138	61	TRABZON	249854
21	DİYARBAKIR	211496	62	TUNCELİ	14906
22	EDİRNE	433209	63	ŞANLIURFA	292502
23	ELAZIĞ	123467	64	UŞAK	91319
24	ERZİNCAN	77242	65	VAN	87239
25	ERZURUM	150799	66	YOZGAT	119568
26	ESKİŞEHİR	338697	67	ZONGULDAK	117923
27	GAZİANTEP	590091	68	AKSARAY	139224
28	GİRESUN	113014	69	BAYBURT	19085
29	GÜMÜŞHANE	31726	70	KARAMAN	60469
30	HAKKARİ	26453	71	KIRIKKALE	168151
31	HATAY	474319	72	BATMAN	178842
32	ISPARTA	114970	73	ŞIRNAK	106648
33	MERSİN	872793	74	BARTIN	41146
34	İSTANBUL	4168543	75	ARDAHAN	28660
35	İZMİR	1344383	76	IĞDIR	29904
36	KARS	52319	77	YALOVA	90227
37	KASTAMONU	104333	78	KARABÜK	89628
38	KAYSERİ	371066	79	KİLİS	26612
39	KIRKLARELİ	180143	80	OSMANİYE	97944
40	KIRŞEHİR	57045	81	DÜZCE	117593
41	KOCAELİ	829240			

Ek1-d Motorlu kara taşıt sayısının yüz ölçümü (km^2) verisine oranı (2019)

P.Kodu	İL	MKT/Yüz Ölçümü (km^2)	P.Kodu	İL	MKT/Yüz Ölçümü (km^2)
1	ADANA	47.5	42	KONYA	17.7
2	ADIYAMAN	14.4	43	KÜTAHYA	18.1
3	AFYONKARAHİSAR	15.9	44	MALATYA	14.4
4	AĞRI	2.9	45	MANİSA	44.3
5	AMASYA	20.9	46	KAHRAMANMARAŞ	16.2
6	ANKARA	79.4	47	MARDİN	8.6
7	ANTALYA	54.6	48	MUĞLA	40.1
8	ARTVİN	5.3	49	MUŞ	3.9
9	AYDIN	56.2	50	NEVŞEHİR	14.1
10	BALIKESİR	33.2	51	NİĞDE	14.8
11	BİLECİK	16.5	52	ORDU	23.2
12	BİNGÖL	2.1	53	RİZE	20.9
13	BİTLİS	2.7	54	SAKARYA	59.7
14	BOLU	13.8	55	SAMSUN	37.0
15	BURDUR	18.8	56	SİİRT	3.6
16	BURSA	83.5	57	SİNOP	10.4
17	ÇANAKKALE	23.5	58	SİVAS	5.7
18	ÇANKIRI	7.0	59	TEKİRDAĞ	43.5
19	ÇORUM	13.8	60	TOKAT	17.8
20	DENİZLİ	33.8	61	TRABZON	42.7
21	DİYARBAKIR	8.1	62	TUNCELİ	1.2
22	EDİRNE	26.1	63	ŞANLIURFA	13.2
23	ELAZIĞ	13.5	64	UŞAK	24.8
24	ERZİNCAN	5.1	65	VAN	3.7
25	ERZURUM	4.8	66	YOZGAT	7.8
26	ESKİŞEHİR	20.6	67	ZONGULDAK	46.4
27	GAZİANTEP	76.2	68	AKSARAY	16.4
28	GİRESUN	13.0	69	BAYBURT	4.2
29	GÜMÜŞHANE	3.7	70	KARAMAN	10.5
30	HAKKARİ	1.3	71	KIRIKKALE	14.2
31	HATAY	89.2	72	BATMAN	10.0
32	ISPARTA	19.9	73	ŞIRNAK	4.1
33	MERSİN	38.7	74	BARTIN	22.1
34	İSTANBUL	766.9	75	ARDAHAN	3.9
35	İZMİR	119.9	76	IĞDIR	7.7
36	KARS	4.4	77	YALOVA	82.7
37	KASTAMONU	10.0	78	KARABÜK	15.9
38	KAYSERİ	22.3	79	KİLİS	33.5
39	KIRKLARELİ	20.7	80	OSMANİYE	49.6
40	KIRŞEHİR	10.4	81	DÜZCE	44.8
41	KOCAELİ	117.5			

Ek1-e Akaryakıt miktarının (benzin-motorin-m^3) yüz ölçümüne (km^2) (2019)

P.Kodu	İL	Akaryakıt (m^3)/ Yüz Ölçümü (km^2)	P.Kodu	İL	Akaryakıt (m^3) /Yüz Ölçümü (km^2)
1	ADANA	46	42	KONYA	23
2	ADIYAMAN	17	43	KÜTAHYA	29
3	AFYONKARAHİSAR	21	44	MALATYA	15
4	AĞRI	9	45	MANİSA	51
5	AMASYA	20	46	KAHRAMANMARAŞ	18
6	ANKARA	104	47	MARDİN	21
7	ANTALYA	50	48	MUĞLA	42
8	ARTVİN	11	49	MUŞ	6
9	AYDIN	54	50	NEVŞEHİR	17
10	BALIKESİR	40	51	NİĞDE	20
11	BİLECİK	24	52	ORDU	31
12	BİNGÖL	7	53	RİZE	28
13	BİTLİS	6	54	SAKARYA	76
14	BOLU	25	55	SAMSUN	54
15	BURDUR	23	56	SİİRT	14
16	BURSA	108	57	SİNOP	10
17	ÇANAKKALE	27	58	SİVAS	7
18	ÇANKIRI	12	59	TEKİRDAĞ	74
19	ÇORUM	17	60	TOKAT	20
20	DENİZLİ	56	61	TRABZON	65
21	DİYARBAKIR	17	62	TUNCELİ	2
22	EDİRNE	85	63	ŞANLIURFA	18
23	ELAZIĞ	16	64	UŞAK	20
24	ERZİNCAN	8	65	VAN	5
25	ERZURUM	7	66	YOZGAT	11
26	ESKİŞEHİR	29	67	ZONGULDAK	43
27	GAZİANTEP	105	68	AKSARAY	22
28	GİRESUN	19	69	BAYBURT	6
29	GÜMÜŞHANE	6	70	KARAMAN	8
30	HAKKARİ	5	71	KIRIKKALE	42
31	HATAY	104	72	BATMAN	48
32	ISPARTA	16	73	ŞIRNAK	18
33	MERSİN	66	74	BARTIN	22
34	İSTANBUL	930	75	ARDAHAN	7
35	İZMİR	137	76	IĞDIR	10
36	KARS	6	77	YALOVA	137
37	KASTAMONU	10	78	KARABÜK	26
38	KAYSERİ	27	79	KİLİS	23
39	KIRKLARELİ	34	80	OSMANİYE	36
40	KIRŞEHİR	10	81	DÜZCE	57
41	KOCAELİ	295			

Ek2-a 2018 yılına ait toplam il nüfus verileri

P. Kodu	İL	Nüfus	P. Kodu	İL	Nüfus
1	Adana	2220125	42	Konya	2205609
2	Adıyaman	624513	43	Kütahya	577941
3	Afyonkarahisar	725568	44	Malatya	797036
4	Ağrı	539657	45	Manisa	1429643
5	Amasya	337508	46	Kahramanmaraş	1144851
6	Ankara	5503985	47	Mardin	829195
7	Antalya	2426356	48	Muğla	967487
8	Artvin	174010	49	Muş	407992
9	Aydın	1097746	50	Nevşehir	298339
10	Balıkesir	1226575	51	Niğde	364707
11	Bilecik	223448	52	Ordu	771932
12	Bingöl	281205	53	Rize	348608
13	Bitlis	349396	54	Sakarya	1010700
14	Bolu	311810	55	Samsun	1335716
15	Burdur	269926	56	Siirt	331670
16	Bursa	2994521	57	Sinop	219733
17	Çanakkale	540662	58	Sivas	646608
18	Çankırı	216362	59	Tekirdağ	1029927
19	Çorum	536483	60	Tokat	612646
20	Denizli	1027782	61	Trabzon	807903
21	Diyarbakır	1732396	62	Tunceli	88198
22	Edirne	411528	63	Şanlıurfa	2035809
23	Elazığ	595638	64	Uşak	367514
24	Erzincan	236034	65	Van	1123784
25	Erzurum	767848	66	Yozgat	424981
26	Eskişehir	871187	67	Zonguldak	599698
27	Gaziantep	2028563	68	Aksaray	412172
28	Giresun	453912	69	Bayburt	82274
29	Gümüşhane	162748	70	Karaman	251913
30	Hakkari	286470	71	Kırıkkale	286602
31	Hatay	1609856	72	Batman	599103
32	Isparta	441412	73	Şırnak	524190
33	Mersin	1814468	74	Bartın	198999
34	İstanbul	15067724	75	Ardahan	98907
35	İzmir	4320519	76	Iğdır	197456
36	Kars	288878	77	Yalova	262234
37	Kastamonu	383373	78	Karabük	248014
38	Kayseri	1389680	79	Kilis	142541
39	Kırklareli	360860	80	Osmaniye	534415
40	Kırşehir	241868	81	Düzce	387844
41	Kocaeli	1906391			

Ek2-b İl bazında dağıtılan suyun abone sayısı (2018)

P.Kodu	İL	Dağıtılan Suyun Abone Sayısı	P.Kodu	İL	Dağıtılan Suyun Abone Sayısı
1	Adana	802952	42	Konya	972268
2	Adıyaman	130428	43	Kütahya	212754
3	Afyonkarahisar	231953	44	Malatya	314426
4	Ağrı	75098	45	Manisa	594200
5	Amasya	116127	46	Kahramanmaraş	413119
6	Ankara	2300689	47	Mardin	137567
7	Antalya	1262526	48	Muğla	461872
8	Artvin	51322	49	Muş	56025
9	Aydın	627617	50	Nevşehir	115264
10	Balıkesir	702541	51	Niğde	105184
11	Bilecik	83485	52	Ordu	268147
12	Bingöl	72369	53	Rize	95888
13	Bitlis	66140	54	Sakarya	469367
14	Bolu	103998	55	Samsun	567323
15	Burdur	94433	56	Siirt	55835
16	Bursa	1327455	57	Sinop	77516
17	Çanakkale	216134	58	Sivas	188539
18	Çankırı	58153	59	Tekirdağ	540356
19	Çorum	178544	60	Tokat	187774
20	Denizli	521307	61	Trabzon	415573
21	Diyarbakır	339322	62	Tunceli	25097
22	Edirne	147716	63	Şanlıurfa	455580
23	Elazığ	209259	64	Uşak	130766
24	Erzincan	75025	65	Van	198576
25	Erzurum	204101	66	Yozgat	129613
26	Eskişehir	426721	67	Zonguldak	193559
27	Gaziantep	484611	68	Aksaray	125515
28	Giresun	146582	69	Bayburt	18303
29	Gümüşhane	33200	70	Karaman	92582
30	Hakkari	29002	71	Kırıkkale	106952
31	Hatay	511669	72	Batman	124029
32	Isparta	167518	73	Şırnak	43558
33	Mersin	816193	74	Bartın	55961
34	İstanbul	6428080	75	Ardahan	15827
35	İzmir	1909180	76	Iğdır	30858
36	Kars	49435	77	Yalova	138582
37	Kastamonu	122081	78	Karabük	90428
38	Kayseri	571719	79	Kilis	24788
39	Kırklareli	123331	80	Osmaniye	151244
40	Kırşehir	83535	81	Düzce	97910
41	Kocaeli	795067			

Ek2-c Dağıtılan su miktarı (m^3/yıl, 2018)

P.Kodu	İL	Dağıtılan Su Miktarı(m^3/Yıl)	P.Kodu	İL	Dağıtılan Su Miktarı(m^3/Yıl)
1	Adana	111404491	42	Konya	119996750
2	Adıyaman	22614744	43	Kütahya	21867998
3	Afyonkarahisar	32065632	44	Malatya	51966858
4	Ağrı	21010518	45	Manisa	67374239
5	Amasya	17376509	46	Kahramanmaraş	64853645
6	Ankara	285242245	47	Mardin	15186001
7	Antalya	178024855	48	Muğla	67951573
8	Artvin	5216062	49	Muş	18760898
9	Aydın	63820256	50	Nevşehir	15084838
10	Balıkesir	73798909	51	Niğde	15159795
11	Bilecik	8028487	52	Ordu	21730340
12	Bingöl	12286210	53	Rize	9729084
13	Bitlis	19877681	54	Sakarya	57313885
14	Bolu	9669972	55	Samsun	70097761
15	Burdur	12327189	56	Siirt	12054031
16	Bursa	139798278	57	Sinop	6849485
17	Çanakkale	19035356	58	Sivas	34500461
18	Çankırı	8671976	59	Tekirdağ	46079418
19	Çorum	21926134	60	Tokat	24077707
20	Denizli	57916805	61	Trabzon	38673118
21	Diyarbakır	54977781	62	Tunceli	4763530
22	Edirne	14184184	63	Şanlıurfa	90407252
23	Elazığ	33645044	64	Uşak	14200050
24	Erzincan	12724247	65	Van	38820975
25	Erzurum	41567049	66	Yozgat	22570266
26	Eskişehir	40364310	67	Zonguldak	18627732
27	Gaziantep	102012881	68	Aksaray	14512337
28	Giresun	12185004	69	Bayburt	3208032
29	Gümüşhane	5267591	70	Karaman	7940231
30	Hakkari	5114914	71	Kırıkkale	12483361
31	Hatay	72645085	72	Batman	23057872
32	Isparta	25316839	73	Şırnak	22044686
33	Mersin	88503292	74	Bartın	4345853
34	İstanbul	808426542	75	Ardahan	4103767
35	İzmir	223652165	76	Iğdır	4255180
36	Kars	16440652	77	Yalova	16496119
37	Kastamonu	12185614	78	Karabük	16222297
38	Kayseri	81513406	79	Kilis	4597724
39	Kırklareli	13220074	80	Osmaniye	22909265
40	Kırşehir	11928861	81	Düzce	12819261
41	Kocaeli	111804020			

Ek2-d Kişi başı arz edilen günlük su miktarı (litre/kişi-gün, 2018)

P. Kodu	İL	K.B.A.E.G. Su Miktarı (Litre/Kişi-Gün)	P. Kodu	İL	K.B.A.E.G. Su Miktarı (Litre/Kişi-Gün)
1	Adana	222	42	Konya	211
2	Adıyaman	175	43	Kütahya	174
3	Afyonkarahisar	204	44	Malatya	326
4	Ağrı	260	45	Manisa	153
5	Amasya	262	46	Kahramanmaraş	357
6	Ankara	239	47	Mardin	284
7	Antalya	329	48	Muğla	403
8	Artvin	221	49	Muş	328
9	Aydın	196	50	Nevşehir	220
10	Balıkesir	308	51	Niğde	213
11	Bilecik	157	52	Ordu	190
12	Bingöl	231	53	Rize	200
13	Bitlis	349	54	Sakarya	332
14	Bolu	158	55	Samsun	246
15	Burdur	237	56	Siirt	230
16	Bursa	171	57	Sinop	240
17	Çanakkale	192	58	Sivas	283
18	Çankırı	256	59	Tekirdağ	186
19	Çorum	208	60	Tokat	240
20	Denizli	245	61	Trabzon	355
21	Diyarbakır	147	62	Tunceli	279
22	Edirne	214	63	Şanlıurfa	171
23	Elazığ	292	64	Uşak	178
24	Erzincan	238	65	Van	212
25	Erzurum	345	66	Yozgat	267
26	Eskişehir	166	67	Zonguldak	240
27	Gaziantep	224	68	Aksaray	199
28	Giresun	199	69	Bayburt	231
29	Gümüşhane	217	70	Karaman	170
30	Hakkari	117	71	Kırıkkale	155
31	Hatay	179	72	Batman	324
32	Isparta	262	73	Şırnak	234
33	Mersin	227	74	Bartın	197
34	İstanbul	189	75	Ardahan	382
35	İzmir	208	76	Iğdır	130
36	Kars	461	77	Yalova	317
37	Kastamonu	208	78	Karabük	287
38	Kayseri	230	79	Kilis	215
39	Kırklareli	199	80	Osmaniye	173
40	Kırşehir	243	81	Düzce	195
41	Kocaeli	235			

K.B.A.E.G. Su Miktarı: Kişi Başı Arz Edilen Günlük Su Miktarı

Ek2-e Ortalama kişi başı tüketilen su miktarı (m^3, 2018)

P. Kodu	İL	O.K.B.T. Su Miktarı (m^3, 2018)	P. Kodu	İL	O.K.B.T. Su Miktarı (m^3, 2018)
1	Adana	50	42	Konya	54
2	Adıyaman	36	43	Kütahya	38
3	Afyonkarahisar	44	44	Malatya	65
4	Ağrı	39	45	Manisa	47
5	Amasya	51	46	Kahramanmaraş	57
6	Ankara	52	47	Mardin	18
7	Antalya	73	48	Muğla	70
8	Artvin	30	49	Muş	46
9	Aydın	58	50	Nevşehir	51
10	Balıkesir	60	51	Niğde	42
11	Bilecik	36	52	Ordu	28
12	Bingöl	44	53	Rize	28
13	Bitlis	57	54	Sakarya	57
14	Bolu	31	55	Samsun	52
15	Burdur	46	56	Siirt	36
16	Bursa	47	57	Sinop	31
17	Çanakkale	35	58	Sivas	53
18	Çankırı	40	59	Tekirdağ	45
19	Çorum	41	60	Tokat	39
20	Denizli	56	61	Trabzon	48
21	Diyarbakır	32	62	Tunceli	54
22	Edirne	34	63	Şanlıurfa	44
23	Elazığ	56	64	Uşak	39
24	Erzincan	54	65	Van	35
25	Erzurum	54	66	Yozgat	53
26	Eskişehir	46	67	Zonguldak	31
27	Gaziantep	50	68	Aksaray	35
28	Giresun	27	69	Bayburt	39
29	Gümüşhane	32	70	Karaman	32
30	Hakkari	18	71	Kırıkkale	44
31	Hatay	45	72	Batman	38
32	Isparta	57	73	Şırnak	42
33	Mersin	49	74	Bartın	22
34	İstanbul	54	75	Ardahan	41
35	İzmir	52	76	Iğdır	22
36	Kars	57	77	Yalova	63
37	Kastamonu	32	78	Karabük	65
38	Kayseri	59	79	Kilis	32
39	Kırklareli	37	80	Osmaniye	43
40	Kırşehir	49	81	Düzce	33
41	Kocaeli	59			

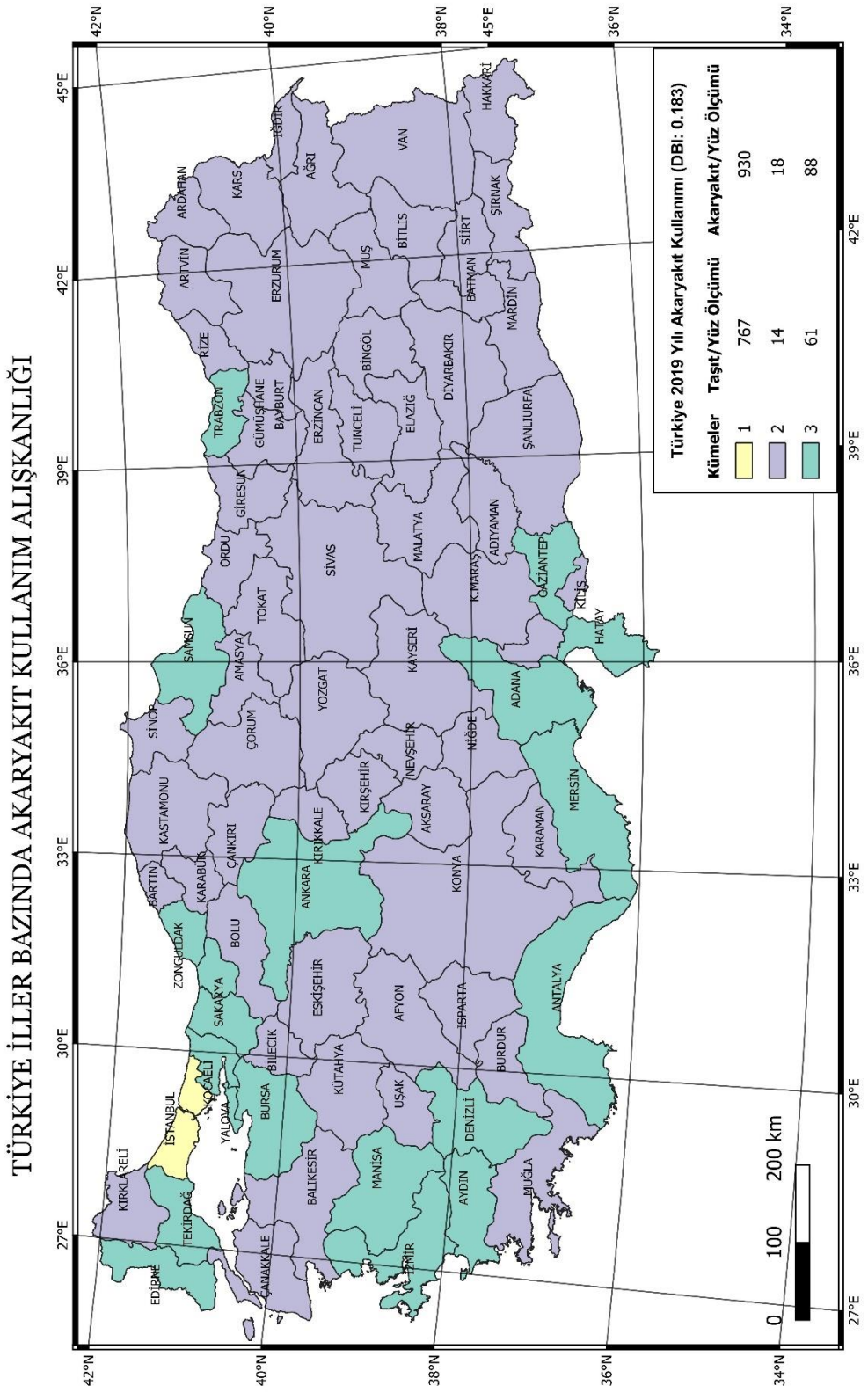
O.K.B.T. Su Miktarı: Ortalama Kişi Başı Tüketilen Su Miktarı

Ek2-f Kişi başı yıllık arz edilen su miktarı (m^3/Kişi-2018)

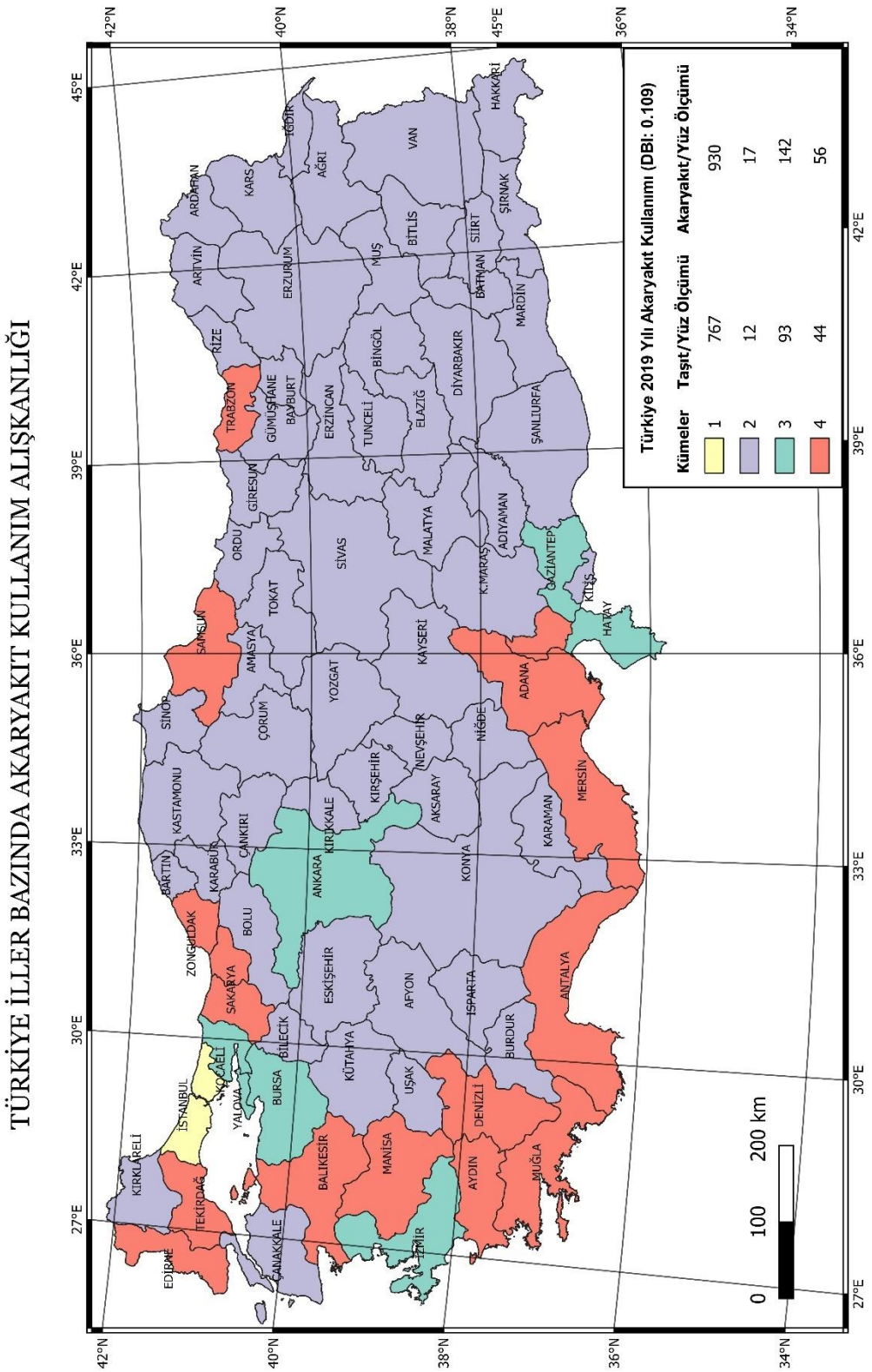
P. Kodu	İL	K.B.Y.A.E. Su Miktarı (m^3/Kişi-2018)	P. Kodu	İL	K.B.Y.A.E. Su Miktarı (m^3/Kişi-2018)
1	Adana	81	42	Konya	77
2	Adıyaman	64	43	Kütahya	64
3	Afyonkarahisar	74	44	Malatya	119
4	Ağrı	95	45	Manisa	56
5	Amasya	96	46	Kahramanmaraş	130
6	Ankara	87	47	Mardin	104
7	Antalya	120	48	Muğla	147
8	Artvin	81	49	Muş	120
9	Aydın	72	50	Nevşehir	80
10	Balıkesir	112	51	Niğde	78
11	Bilecik	57	52	Ordu	69
12	Bingöl	84	53	Rize	73
13	Bitlis	127	54	Sakarya	121
14	Bolu	58	55	Samsun	90
15	Burdur	87	56	Siirt	84
16	Bursa	62	57	Sinop	88
17	Çanakkale	70	58	Sivas	103
18	Çankırı	93	59	Tekirdağ	68
19	Çorum	76	60	Tokat	88
20	Denizli	89	61	Trabzon	130
21	Diyarbakır	54	62	Tunceli	102
22	Edirne	78	63	Şanlıurfa	62
23	Elazığ	107	64	Uşak	65
24	Erzincan	87	65	Van	77
25	Erzurum	126	66	Yozgat	97
26	Eskişehir	61	67	Zonguldak	88
27	Gaziantep	82	68	Aksaray	73
28	Giresun	73	69	Bayburt	84
29	Gümüşhane	79	70	Karaman	62
30	Hakkari	43	71	Kırıkkale	57
31	Hatay	65	72	Batman	118
32	Isparta	96	73	Şırnak	85
33	Mersin	83	74	Bartın	72
34	İstanbul	69	75	Ardahan	139
35	İzmir	76	76	Iğdır	47
36	Kars	168	77	Yalova	116
37	Kastamonu	76	78	Karabük	105
38	Kayseri	84	79	Kilis	78
39	Kırklareli	73	80	Osmaniye	63
40	Kırşehir	89	81	Düzce	71
41	Kocaeli	86			

K.B.Y.A.E. Su Miktarı: Kişi Başı Yıllık Arz Edilen Su Miktarı

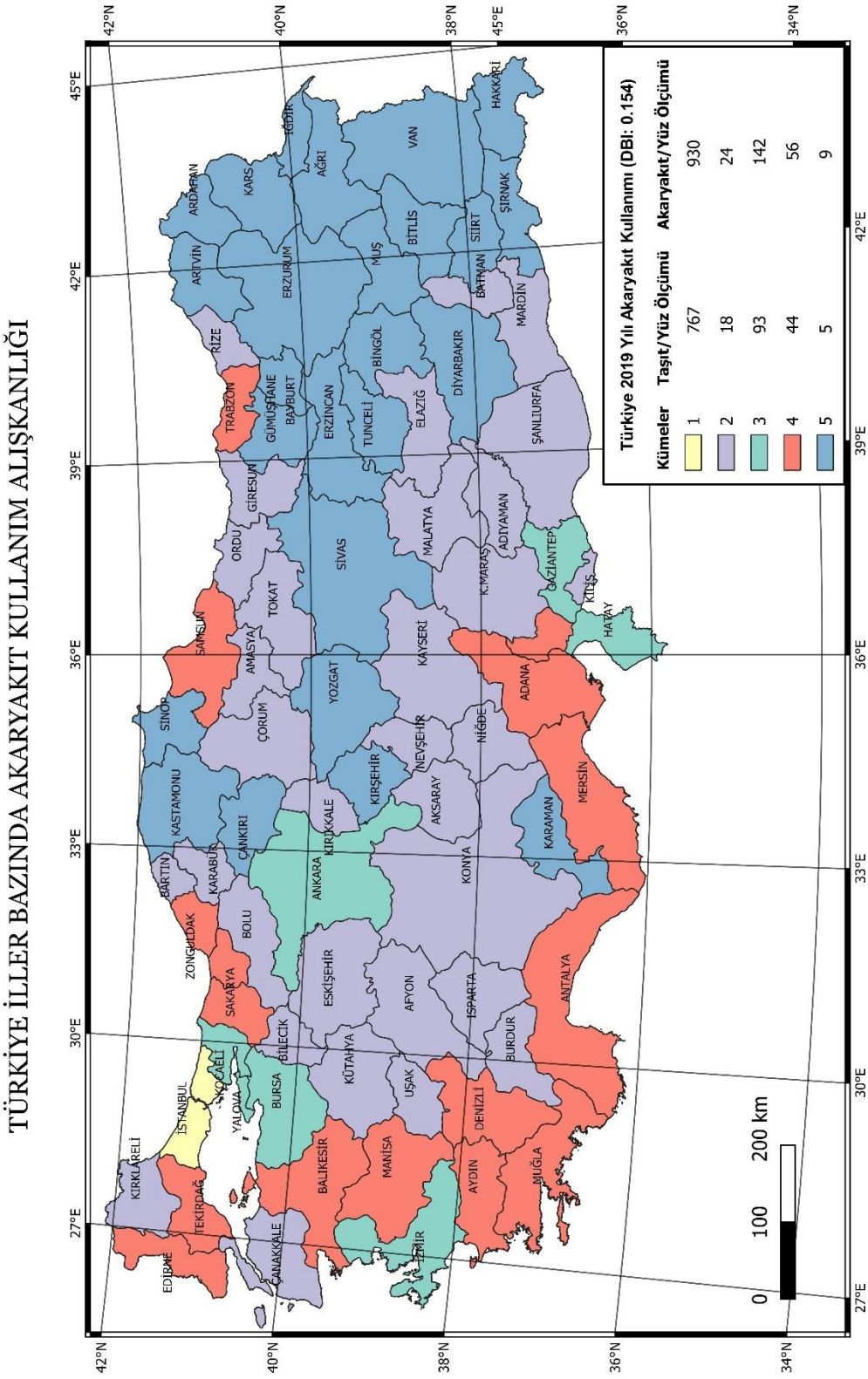
Ek3-a K-Means yöntemi kullanılarak yapılan kümeleme analizi sonucu elde edilen 3 kümeli tematik harita



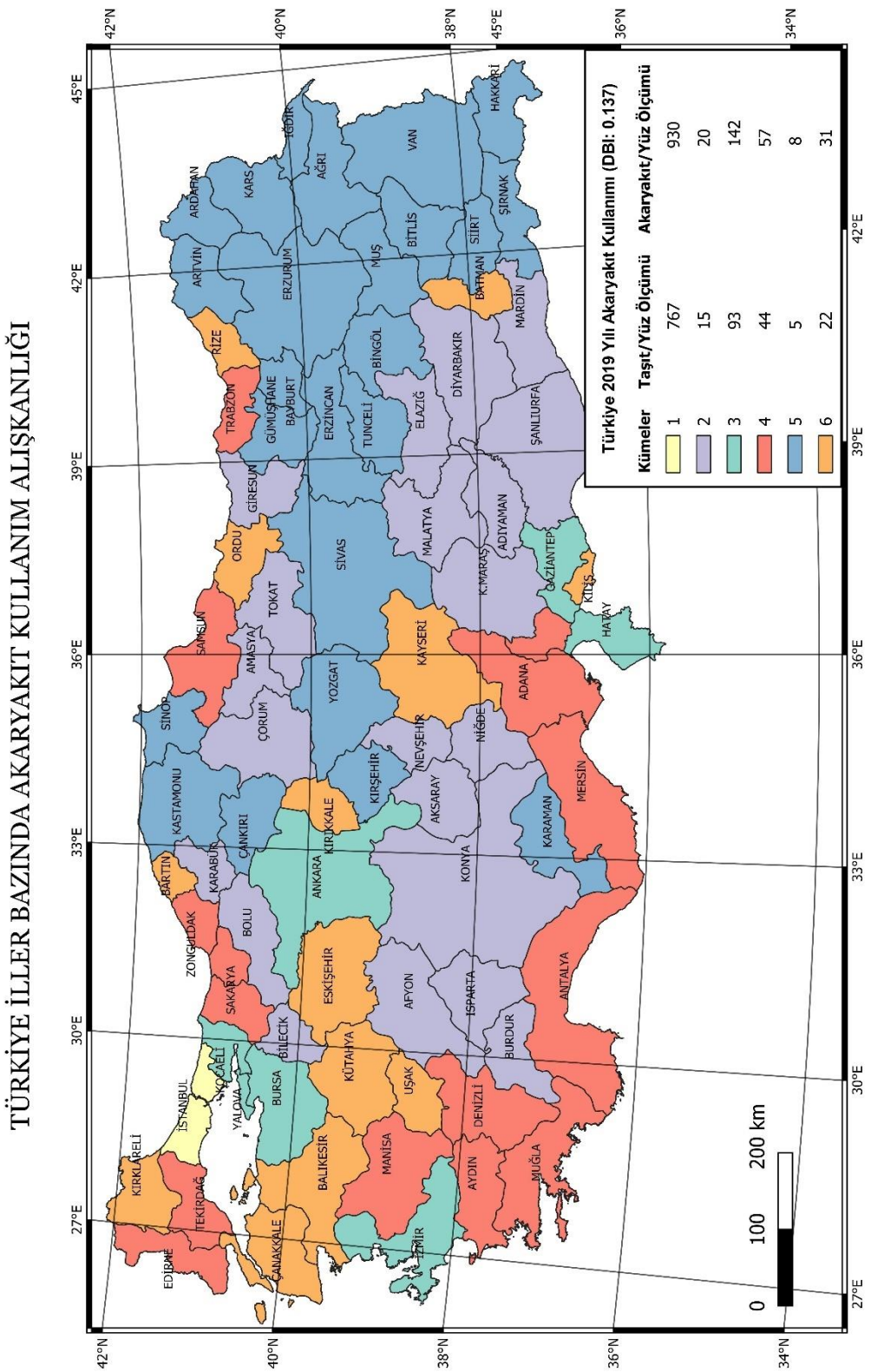
Ek3-b K-Means yöntemi kullanılarak yapılan kümeleme analizi sonucu elde edilen 4 kümeli tematik harita



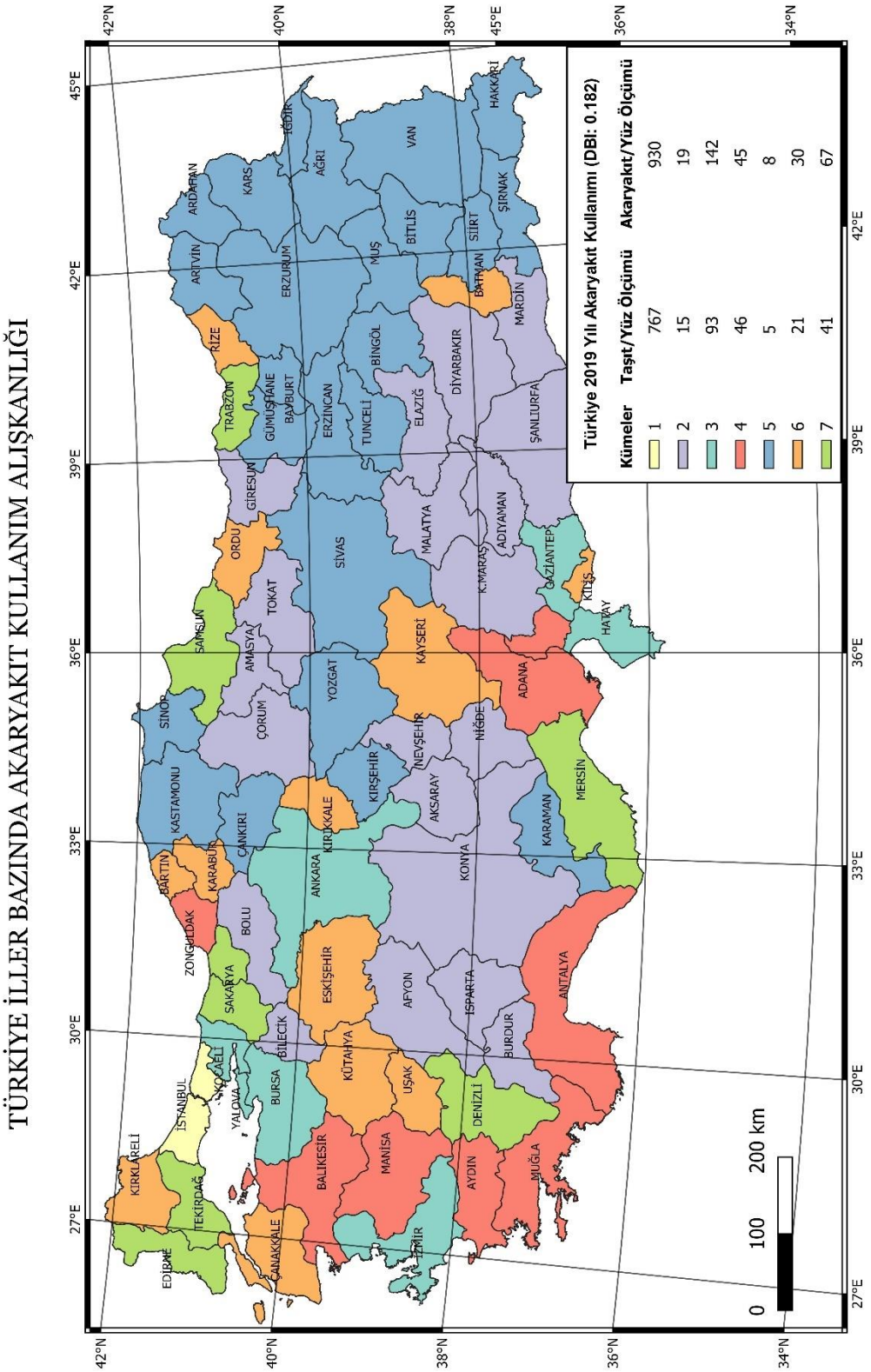
Ek3-c K-Means yöntemi kullanılarak yapılan kümeleme analizi sonucu elde edilen 5 kümeli tematik harita



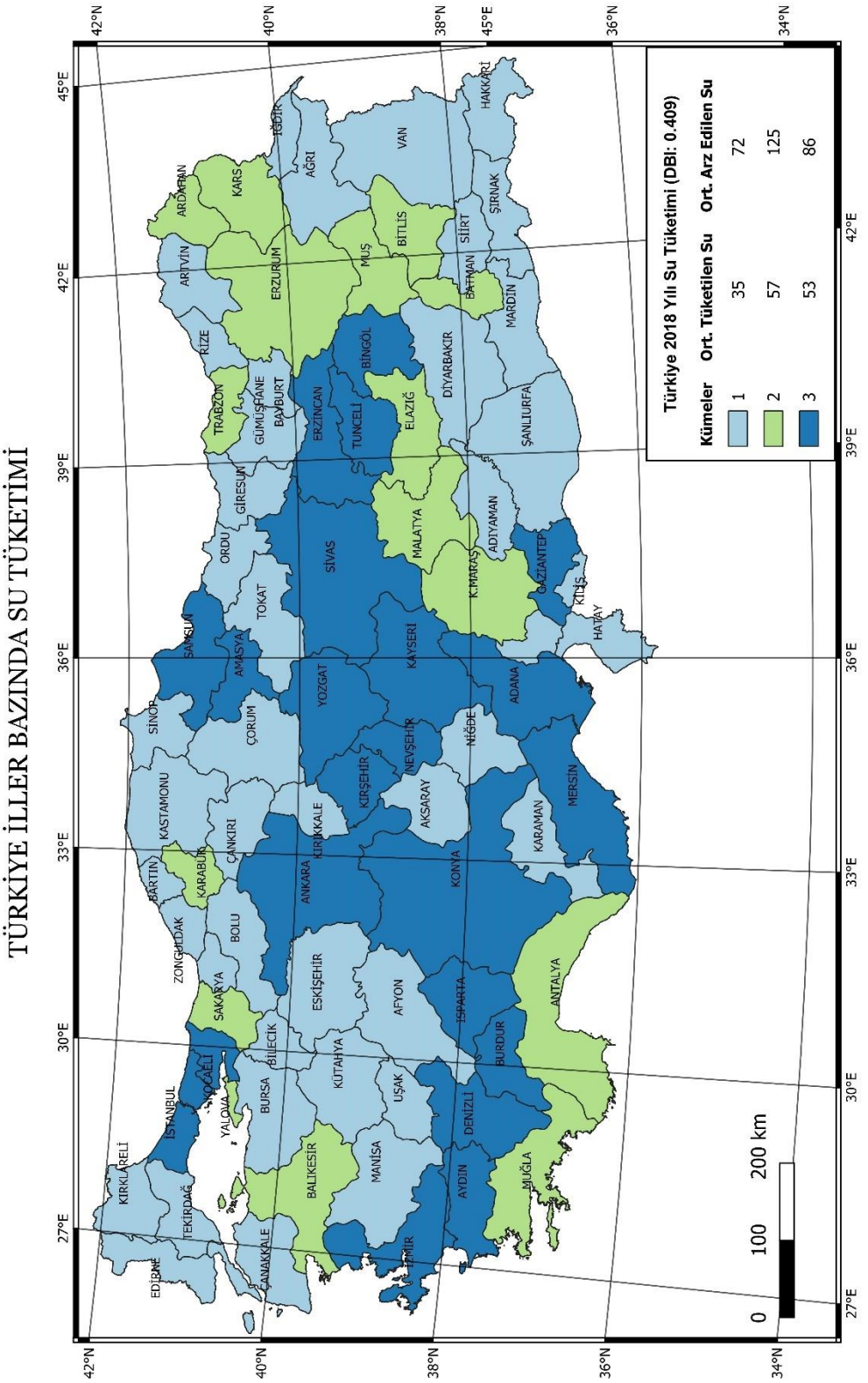
Ek3-d K-Means yöntemi kullanılarak yapılan kümeleme analizi sonucu elde edilen 6 kümeli tematik harita



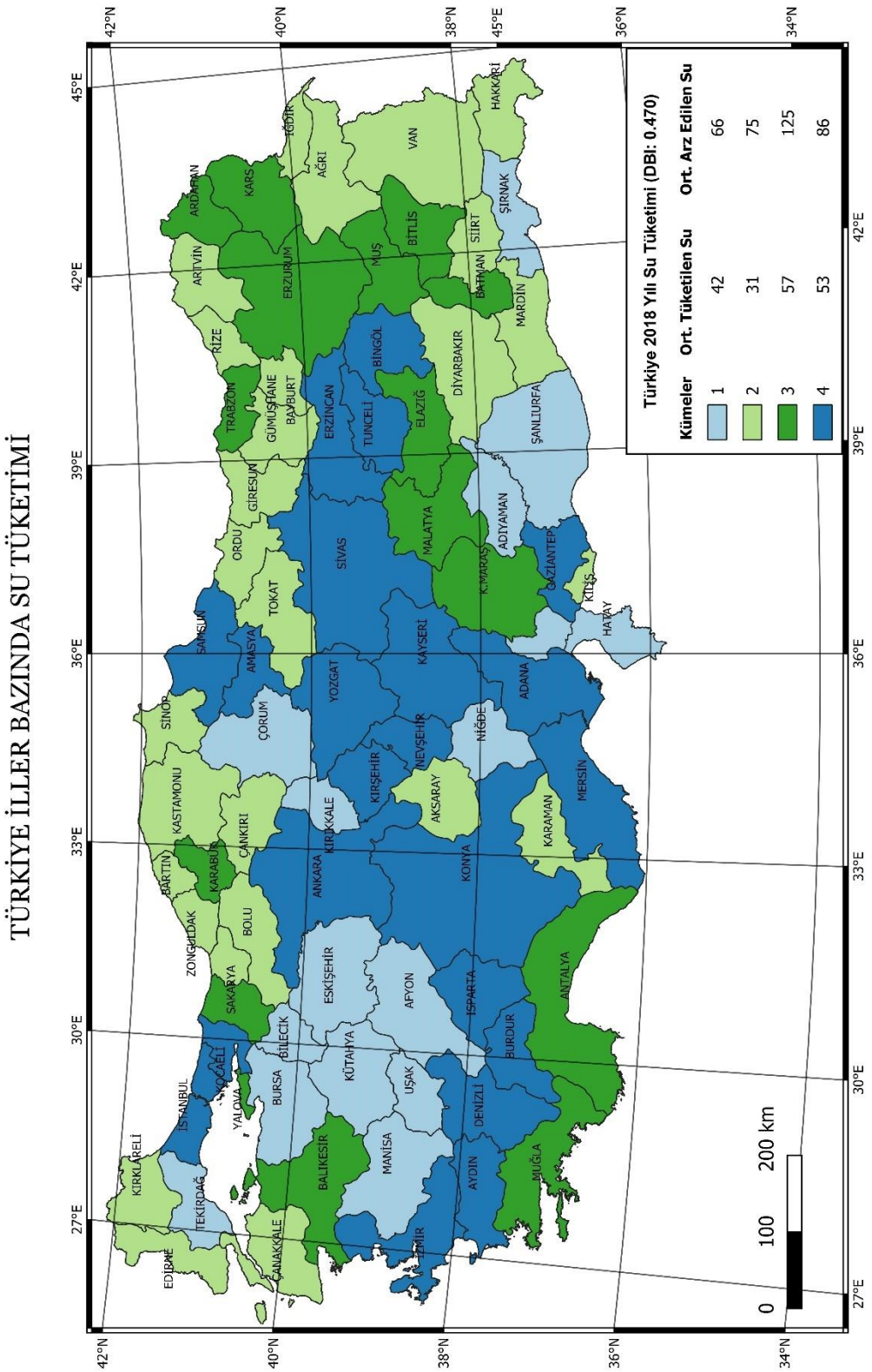
Ek3-e K-Means yöntemi kullanılarak yapılan kümeleme analizi sonucu elde edilen 7 kümeli tematik harita



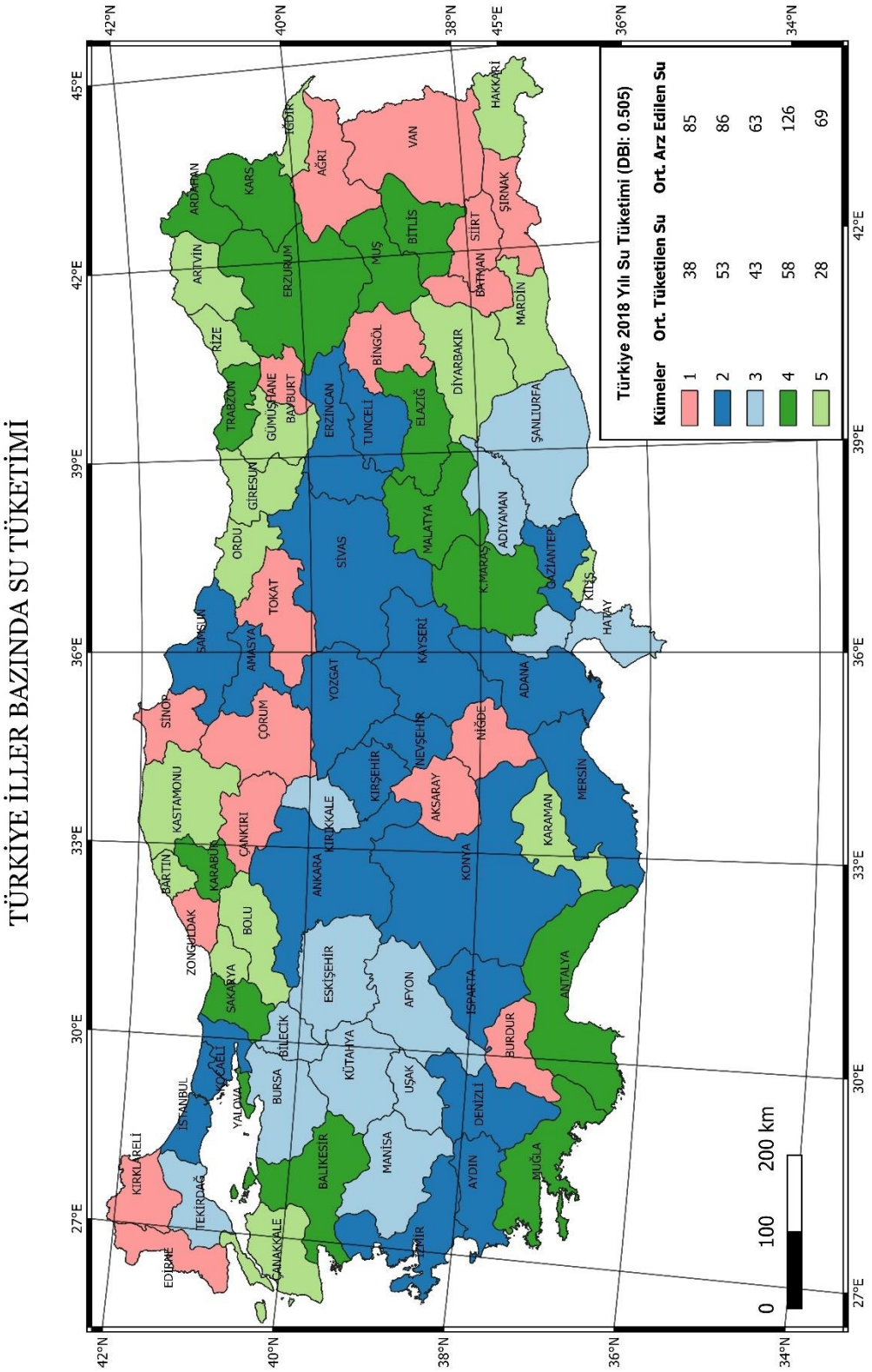
Ek4-a K-Means yöntemi kullanılarak yapılan kümeleme analizi sonucu elde edilen 3 kümeli tematik harita



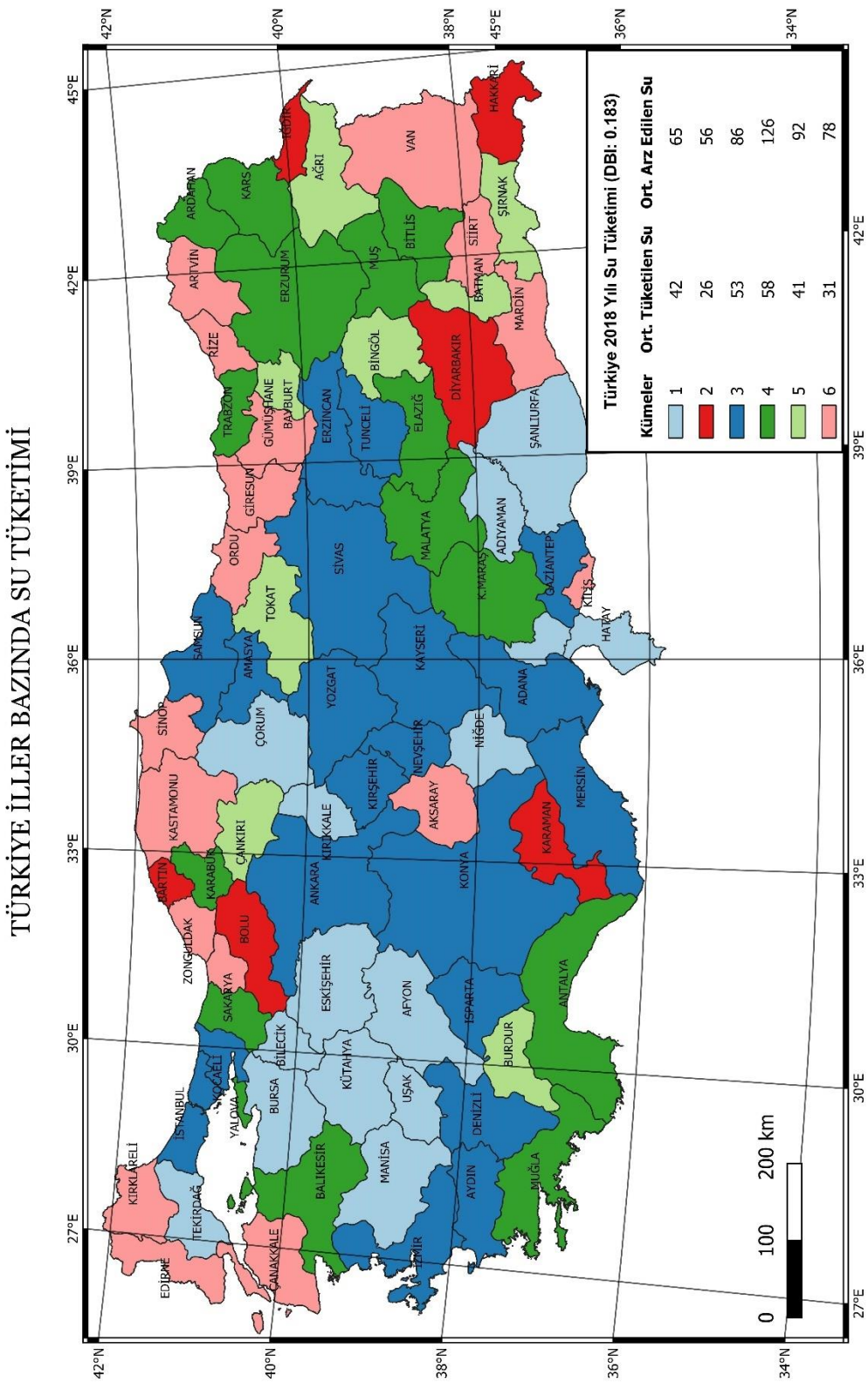
Ek4-b K-Means yöntemi kullanılarak yapılan kümeleme analizi sonucu elde edilen 4 kümeli tematik harita



Ek4-c K-Means yöntemi kullanılarak yapılan kümeleme analizi sonucu elde edilen 5 kümeli tematik harita



Ek4-d K-Means yöntemi kullanılarak yapılan kümeleme analizi sonucu elde edilen 6 kümeli tematik harita



Ek4-e K-Means yöntemi kullanılarak yapılan kümeleme analizi sonucu elde edilen 7 kümeli tematik harita