T.C.
NECMETTİN ERBAKAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MEKANSAL VERİLERİN KÜMELEME ANALİZİ İLE DEĞERLENDİRİLMESİ

Burak ÇAĞLAR

YÜKSEK LİSANS TEZİ

Harita Mühendisliği Anabilim Dalı

Ocak-2018
KONYA
Her Hakkı Saklıdır

TEZ KABUL VE ONAYI

Burak ÇAĞLAR tarafından hazırlanan "MEKANSAL VERİLERİN KÜMELEME ANALİZİ İLE DEĞERLENDİRİLMESİ" adlı tez çalışması 08/01/2018 Tarihinde aşağıdaki jüri tarafından oy birliği ile Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü Harita Mühendisliği Anabilim Dalı'nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan	
Prof.Dr. İ. Öztuğ BİLDİRİCİ	…….
Danışman	
Yrd.Doç.Dr. Hüseyin Zahit SELVİ	…….
Üye	
Yrd.Doç.Dr. İlkay BUĞDAYCI	…….

Yukarıdaki sonucu onaylarım.

Prof. Dr. Ahmet COŞKUN

FBE 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.

Burak ÇAĞLAR

Tarih:

ÖZET

YÜKSEK LİSANS TEZİ

MEKANSAL VERİLERİN KÜMELEME ANALİZİ İLE DEĞERLENDİRİLMESİ

BURAK ÇAĞLAR

Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü

Harita Mühendisliği Anabilim Dalı

Danışman: Yrd.Doç.Dr. Hüseyin Zahit SELVİ

2018, 95 Sayfa

Jüri

Prof.Dr. İ. Öztuğ BİLDİRİCİ

Yrd.Doç.Dr. Hüseyin Zahit SELVİ

Yrd.Doç.Dr. İlkay BUĞDAYCI

Teknolojik gelişmeler sayesinde her geçen gün iş, toplum, bilim ve mühendislik, sağlık ve günlük hayatla ilgili her alandan sürekli olarak veriler toplanmakta ve bu veriler büyük kapasiteli veritabanlarında saklanmaktadır. Bu veritabanlarında yer alan verilerin insanoğlunun hayatında daha faydalı olabilmesi için çeşitli tekniklerle işlenerek anlam kazandırılması yani "bilgi"ye dönüştürülmesi gerekmektedir. Veri Madenciliği disiplini çeşitli algoritma ve teknikler kullanılarak büyük veritabanlarında yer alan veri yığınlarından anlamlı bilginin elde edilmesine imkân sağlamıştır.

Bu çalışma kapsamında "Veri Madenciliği" disiplini, veri madenciliğinin kullanım alanları ve veri madenciliği model ve teknikleri açıklanmıştır. Ayrıca mekânsal verilerin analizinde veri madenciliği tekniklerinin kullanımı üzerinde durulmuştur. Bu kapsamda Türkiye'deki 2011, 2012 ve 2013 yıllarına ait Trafik Kaza istatistik veri setleri üzerinde k-ortalama yöntemi, k-medoids yöntemi ve Birleştirici Hiyerarşik Kümeleme (AGNES) yöntemleri kullanılarak kümeleme analizi yapılmış ve kümeleme analizi sonuçları kullanılarak çok değişkenli haritalar üretilmiştir. Üretilen haritalar karşılaştırılarak bu haritaların risk yönetimi ve planlamada kullanılabilirliği tartışılmıştır. 2011, 2012 ve 2013 yıllarına ait verilerin AGNES kümeleme analizi sonuçlarıyla hazırlanan çok değişkenli haritaların birbirleriyle oldukça uyumlu olduğu görülmüştür. Bu sonuç AGNES yöntemiyle üretilen çok değişkenli haritaların risk yönetimi açısından da oldukça önemli olduğunu göstermiştir. k-ortalama ve k-medoids kümeleme analizleri sonuçlarıyla üretilen çok değişkenli haritalarda farklı küme sayıları için kümeleme sonuçları gözlemlenmiştir. Her iki algoritmanın da kümeleme performansları benzerlik gösterse de k-medoids algoritmasında kümelerin birbirinden daha iyi ayrıldığı gözlemlenmiştir.

Anahtar kelimeler: Veri, Bilgi, Veritabanı, Veri Madenciliği, Kümeleme Analizi, Çok Değişkenli Harita

ABSTRACT

MS THESIS

EVALUATION OF SPATIAL DATA WITH CLUSTERING ANALYSIS

BURAK ÇAĞLAR

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF NECMETTIN ERBAKAN UNIVERSITY

THE DEGREE OF MASTER OF SCIENCE

IN GEOMATICS ENGINEERING

Advisor: Asst.Prof.Dr. Hüseyin Zahit SELVİ

2018, 95 Pages

Jury

Prof.Dr. İ. Öztuğ BİLDİRİCİ

Asst.Prof.Dr. Hüseyin Zahit SELVİ

Asst.Prof.Dr İlkay BUĞDAYCI

Thanks to technological developments, data are constantly collected everyday from work, society, science and engineering, health and daily life, and these data are stored in large-capacity databases. In order to make the data contained in these databases useful in the human life, it has to be transformed into "information" by means of various techniques. The discipline of Data Mining has enabled the use of various algorithms and techniques to obtain meaningful information from data stacks in large databases.
In this study, "Data Mining" discipline, usage areas of data mining, data mining models and techniques were explained. In addition, the use of data mining techniques in the analysis of spatial data was examined. In this context, clustering analysis was carried out on the Traffic accident data sets for 2011, 2012 and 2013 in Turkey using k-means method, k-medoids method and Agglomerative and Divisive Hierarchical Clustering (AGNES) method and clustering analysis results were used to design multivariate maps. By comparing these maps, the usage possibilities in risk management and planning is discussed. The multivariate maps prepared with the results of the AGNES cluster analysis of the data for the years 2011, 2012 and 2013 were found to be very compatible with each other. This result was showed that the multivariate maps produced by AGNES method are also very important in terms of risk management. The clustering results for k-means and k-medoid clustering analyses were observed for different cluster numbers in the generated multivariate maps. Although the clustering performances of both algorithms are similar, it was observed that the k-Medoids algorithm has better separation of clusters.

Keywords: Data, Information, Database, Data Mining, Clustering Analysis, Multivariate Map

ÖNSÖZ

Bu tez çalışmasına beni yönlendirerek yüksek lisans eğitiminin amacı doğrultusunda öğrencisini araştırma yetilerini ortaya koyarak bilgi dağarcığını bir nokta daha ileriye taşımasına olanak veren, Necmettin Erbakan Üniversitesi öğretim üyesi danışman hocam, Sayın Yrd.Doç.Dr. Hüseyin Zahit SELVİ hocama teşekkürlerimi sunarım.

Her zaman yanımda olan eşim Nuriye ÇAĞLAR'a ve sevgisiyle beni besleyen kızım Feyza ÇAĞLAR'a çok teşekkür ederim.

Burak ÇAĞLAR
KONYA-2018

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

Kısaltmalar

AGNES	AGglomerativeNEsting
BIRCH	Balanced Iterative Reducing and Clustering Using Hierarchies
CART	Classification and Regression Trees
CHAID	Chi-Square Automatic Interaction Detector
CHAMELEON	Hierarchical Clustering Using Dynamic Modeling
CLIQUE	Clustering High-Dimensional Space
CRM	Customer Relations Management
CURE	Clustering Using REpresentatives
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DENCLUE	DENsity-based CLUstEring
DIANA	DIvisive ANAlysis
ID3	Iterative Dichotomiser 3
OLAP	On-Line Analytical Processing
OPTICS	Ordering Points To Identify the Clustering Structure
PAM	Partitioning Around Medoids
ROCK	RObust Clustering using linKS
SLIQ	Supervised Learning InQuest
SPRINT	Scalable PaRallelizable Induction of Decision Trees
STING	STatistical INformation Grid
VTBK	Veritabanlarında Bilgi Keşfi
VM	Veri Madenciliği
WaveCluster	Clustering Using Wavelet Transformation

ŞEKİL LİSTESİ

ÇİZELGE LİSTESİ

1. GİRİŞ

Veri, bilginin hammaddesi olup, bilginin temsil biçimidir. Bilginin hammaddesi olan veri bazı durumlarda tek başına bilgi özelliği de taşıyabilir. Bilgi verilerin toplamından oluşan bir küme olarak düşünülmemelidir. **Veri**, gerçek dünyada ki nesnelerin sembolojik gösterimi olarak ifade edilirken; **bilgi**, kullanıcı tarafından işlenerek anlaşılabilir formlara dönüştürülmüş veri seti olarak düşünülebilir (Yomralıoğlu, 2000). Yeryüzünde veya yakınında belirli bir anlama sahip olan doğal (*nehir, orman vb.*) ve yapay (*yol, bina vb.*) nitelikteki coğrafi verileri belirli bir referans sistemine göre yerini ve biçimini belirten vektörel ve raster verilere **mekansal veri** denilmektedir.

İnsanoğlu büyük miktardaki verilerin günlük olarak toplandığı Dünya'da yaşamaktadır. Bu gibi verilerin analiz edilmesi ve insanoğlu için yarar sağlayacak bilgiye dönüştürülmesi önemli bir ihtiyaçtır. Veri madenciliği, veriden bilgiyi elde etmek için çeşitli araçlar sağlayarak bu ihtiyacı karşılayabilmektedir.

Günümüzde her insanın kullandığı en popüler sözcüklerden birisi "Bilgi çağında yaşıyoruz" sözüdür. Acaba gerçekten bilgi çağında mı yaşıyoruz? Bu çalışmaya altlık teşkil eden bu konu incelediğinde gerçekte "Bilgi" çağında değil de iş, toplum, bilim ve mühendislik, sağlık ve günlük hayatla ilgili her alandan internet aracılığıyla ve çeşitli veri toplama cihazlarıyla toplanan terabaytlarca hatta petabaytlarca (1000 terabayt) verinin içerisinde yaşadığımız görülmektedir.

İşletmeler, Dünya çapında satış işlemlerini, hisse senedi alım-satım kayıtlarını, ürün tanımlamalarını, satış promosyonlarını, firma profillerini ve performanslarını, müşteri bildirimlerini vb. içeren devasa veri setleri oluşturmaktadır. Örneğin, dünya çapında şubeleri bulunan büyük mağazalar haftada yüz milyonlarca işlem gerçekleştirebilmektedir. Yine bilim ve mühendislik uygulamaları, sürekli olarak uzaktan algılama, ölçme işlemleri, bilimsel deneyler, mühendislik gözlemleri ve çevresel gözetimlerden yararlanarak petabaytların üstünde veri üretmektedirler. Küresel omurga telekomünikasyon (Global backbone telecommunication) ağları her gün petabaytlarca veri trafiği gerçekleştirmektedirler. Medikal ve sağlık sektörü medikal kayıtlardan, hasta takibinden, medikal görüntülemeden büyük miktarda veri üretmektedir. Arama motorları tarafından desteklenen milyarlarca Web araması günde petabaytlarca veri işlemektedir. Dijital resimler ve videolar, bloglar, çeşitli sosyal ağlar üreten topluluklar ve sosyal

medya önemli veri kaynakları haline gelmiştir. Büyük miktarda veri üreten kaynakların listesi daha da artırılabilir. Hızla büyüyen, yaygın olarak kullanılabilen ve devasa bir hale gelen veri, yaşadığımız çağımızı gerçek anlamda "veri çağı" yapmaktadır. Büyük miktardaki verilerden değerli bilgileri otomatik olarak ortaya çıkarmak için güçlü ve çok yönlü araçlara ihtiyaç vardır. Bu ihtiyaç "veri madenciliğinin" doğmasına neden olmuştur (Han ve ark., 2011).

Veri toplama araçları ve veritabanlarına yaşanan teknolojik gelişmeler sayesinde yersel ölçme, fotogrametri, GPS, uzaktan algılama ve mevcut haritaların sayısallaştırılması yöntemleriyle elde edilen mekânsal verilerin veritabanlarında depolanması kolaylaşmış ve tam anlamıyla mekânsal veri patlaması yaşanmaktadır. Klasik mekânsal analiz (bindirme analizi, tampon analizi, ağ analizi) yöntemleriyle sınırlandırılmış ve aynı türden (homojen) veriler arasında bilgi çıkarımı yapılmaktadır. Veri Madenciliği disiplini kullanıcılara büyük veri tabanlarında yer alan farklı türdeki (heterojen) mekânsal veriler arasındaki gizli özellikler ve ilişkiler keşfetmeye olanak sağlamaktadır. Mekânsal verilerin, veri madenciliği disiplini ile analizi sonu elde edilecek bilgiler çevresel yönetim, ulaşım, halk sağlığı, tarım, endüstri, ulusal savunma, risk yönetimi vb. konularda doğru kararların alınmasını kolaylaştırmaktadır.

Kümeleme analizi, veri madenciliği tekniklerinden en yaygın olarak kullanılanıdır. Kümeleme analizindeki temel amaç, bir veri setindeki elemanların aralarındaki uzaklıkların en az olacak şekilde kümelenmesi, özellikleri birbirinden çok farklı olan elemanlar arasındaki uzaklıkların en çok olacak şekilde kümeler oluşturulmasıdır. Kümeleme analizi tekniğinin kullanımı konusunda muhtelif çalışmalar bulunmaktadır. Karpat ve Yılmaz (2002), Türkiye'deki trafik kaza oluş şekillerinin, kazanın olduğu yerdeki trafik, aydınlatma ve yol durumlarına göre nasıl kümelenme gösterdiklerini, hiyerarşik olmayan k-means algoritması kullanarak araştırmışlardır. Çakmak, Uzgören ve Keçek (2005), Türkiye'deki 73 ilin kültürel yapılarına göre nasıl kümelenme gösterdiklerini, hiyerarşik kümelenme yöntemlerini kullanarak araştırmışlardır. Yılmaz ve Temurlenk (2005), Türkiye'deki 'Düzey' ve 'Düzey 2' istatistik bölgelerini kişi başına düşen gelir açısından nasıl kümelenme gösterdiklerini, hiyerarşik olmayan kümeleme yöntemlerinden k-means algoritması ve hiyerarşik kümeleme yöntemlerinden tek bağlantı (en yakın komşu) metodu kullanarak araştırmışlardır. Akat (2007), 52 ülkenin askeri yapıları ve bu yapıyı temel olarak

etkileyen değişkenlerini kullanılarak nasıl kümelenme gösterdiklerini, hiyerarşik olan kümeleme yöntemlerinde Ward metodu ve hiyerarşik olmayan kümeleme yöntemlerinden K-means algoritması kullanarak araştırmıştır. Çetinkaya (2008), İstanbul'daki binaların kat, bodrum kat, taşıyıcı sistem, kullanım amacı ve binanın zemininin jeolojik özelliklerine nasıl kümelenme gösterdiklerini, yoğunluğu dayalı kümeleme yöntemlerinden DBSCAN kümeleme metodu kullanarak araştırmıştır. Şekerler ve Murat (2009), Denizli iline ait 2004, 2005 ve 2006 yıllarına ait trafik kaza verilerini kullanarak trafik kazalarının daha yoğun olduğu noktaların kara nokta olarak belirlenmesi yönünde, k-means ve fuzzy-c algoritmaları kullanarak trafik kaza noktalarının nasıl kümelendiğini araştırmışlardır. Atalay ve Tortum (2010), Türkiye'deki illerin 1997-2006 yılları arasındaki trafik kazalarına göre nasıl kümelenme gösterdiklerini, k-means ve fuzzy-c algoritmaları kullanarak araştırmışlardır. Sarıman (2011), "Flags" veri seti kullanılarak ükelerin nasıl kümelenme gösterdiklerini, hiyerarşik olmayan k-means ve k-medoids kümeleme algoritmaları kullanarak araştırmıştır. Alkan (2012), Bingöl, Elazığ, Malatya ve Tunceli il ve ilçe merkezlerindeki hanelerin yıllık elektrik tüketim değerleri dikkate alarak bu yerleşim yerlerinin nasıl kümelenme gösterdiklerini, hiyerarşik kümeleme yöntemleri kullanarak araştırmıştır. Çelik (2013), Türkiye'deki 81 ilin sağlık göstergelerini göre nasıl kümelenme gösterdiklerini, hiyerarşik olmayan kümeleme yöntemlerinde k-means algoritması kullanarak araştırmıştır. Çiçekdağı (2013), Kütahya ve etrafındaki 250 kilometrelik yarıçaplı çemberi kapsayan bölgede 1900'lü yıllardan günümüze gelene kadar meydana gelen depremlerin büyüklüğü, zamanı, derinliği, sıcak su kaynağı üzerinde olup olmaması, toprak çeşidi ve Kütahya merkeze olan uzaklığı değişkenlerini kullanarak nasıl kümelenme gösterdiklerini, hiyerarşik olmayan kümeleme yöntemlerinde k-ortalama kümeleme metodu kullanarak araştırmıştır.

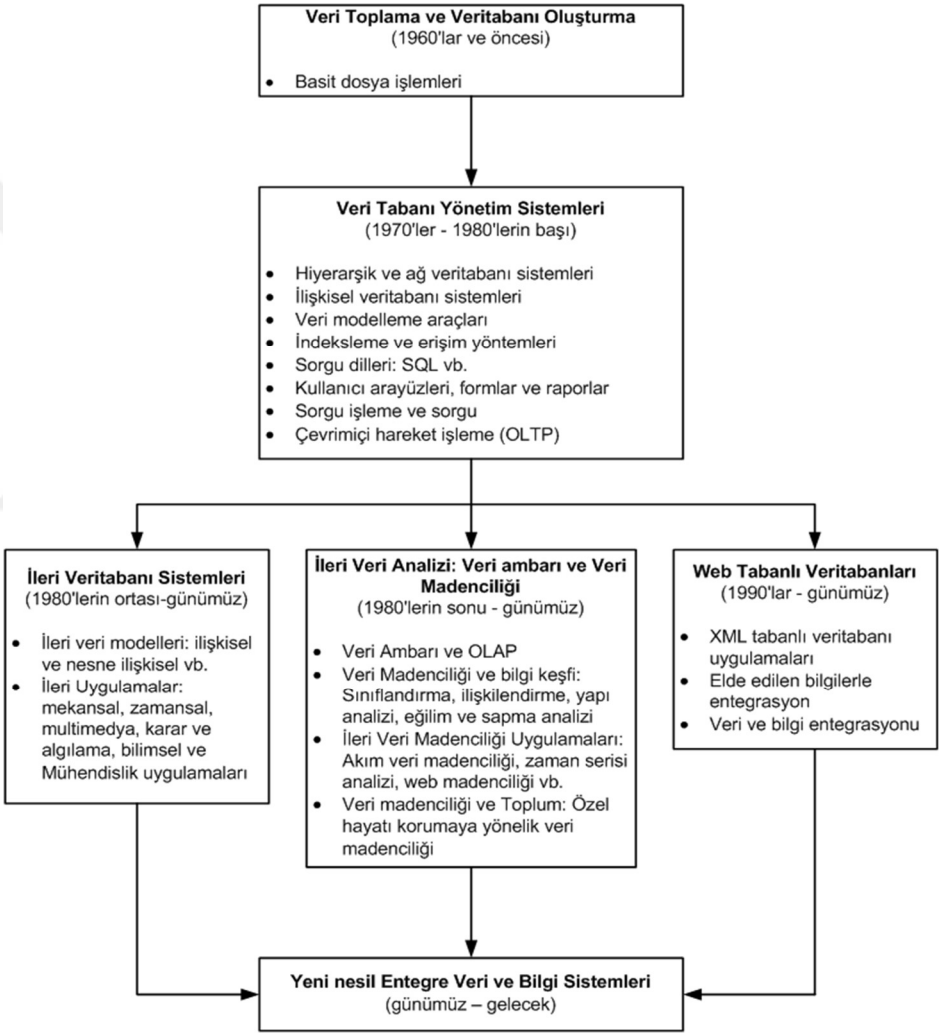
Mekânsal objelerin birden fazla özelliği çok değişkenli haritalama (multivariate mapping) ile harita üzerinde gösterilebilir. Bu yöntem harita kullanıcısına mekânsal objelerin farklı özelliklerini eş zamanlı olarak karşılaştırmasına olanak sağlamaktadır (Buckley, 2008). Slocum ve ark. (2009), New York şehrine ait veri setinde yer alan 1990-2000 yılları nüfus değişimi, işsizlik oranı, Afrikalı – Amerikalı oranı ve bebek ölüm oranı gibi değişkenleri kullanarak hiyerarşik kümeleme yöntemiyle kümeleme analizi gerçekleştirmiş ve kümeleme sonuçlarını çok değişkenli harita gösterimiyle yorumlamışlardır.

Tezin ikinci bölümünde büyük miktardaki veri setlerinin analiz edilerek insanoğlunun yararlandığı anlamlı bilgiyi elde etme de yararlanılan "Veri Madenciliği" disiplini, "Veri ambarları ve OLAP", veri madenciliğinin kullanım alanları, hakkında bilgi verilmiştir. Veri madenciliği sürecinde veri setlerinin işlenerek anlamlı bilginin elde edilmesini sağlayan "Veri Madenciliği Model ve Teknikleri" hakkında kısa bilgi verilmiş ve bu tez çalışmasında kullanılan hiyerarşik olamayan kümeleme analizi yöntemlerinden k-ortalama ve k-medoids algoritmaları ile hiyerarşik kümeleme analizi yöntemi olan AGNES-DIANA algoritmaları hakkında bilgi verilmiştir. Ayrıca hiyerarşik olmayan kümeleme yöntemlerinde küme sayısının belirlenmesinde kullanılan Dunn ve Davies-Bouldin indeksleri hakkında bilgi verilmiştir. Üçüncü bölümünde bu çalışma kapsamında kullanılan materyal ve metotlar anlatılmıştır. Dördüncü bölümde, Türkiye'deki 81 ilde 2011, 2012 ve 2013 yıllarına ait il bazlı motorlu kara taşıtı sayısı, ölümlü ve yaralanmalı trafik kaza sayıları, ölü sayıları ve yaralı sayıları verileri kullanılarak (4 farklı değer); k-Ortalama, k-Medoids ve Birleştirici Hiyerarşik Kümeleme (AGNES) yöntemleriyle kümeleme analizi yapılarak ve kümeleme analizi sonuçlarına göre çok değişkenli haritalar üretilmiştir. Sonuç bölümünde de üç farklı yöntemle 2011, 2012 ve 2013 yıllarına ait verilerden kümeleme analizi sonucu elde edilen sonuçlar değerlendirilmiş, kümeleme analizi sonuçlarına göre üretilen çok değişkenli haritalar karşılaştırılmış ve kümeleme başarısı açısından hangi yöntemin daha uygun olduğu değerlendirilmiştir.

2. KAYNAK ARAŞTIRMASI

2.1. Geçmişten Günümüze Veri Madenciliği

Veri madenciliği bilgi teknolojilerinin doğal evriminin bir sonucu olarak da nitelendirilebilir. Veritabanı sistemleri Şekil 2.1'de görülen evrimsel yolu izleyerek *veri toplama ve veritabanı oluşturma*, *veri yönetimi (veri saklama ve geri erişim dahil)* ve *gelişmiş veri analizi* (veri ambarı ve veri madenciliğini içeren) aşamalarından geçerek günümüze gelmiştir (Han ve Kamber, 2006).



Şekil 2.1 Veritabanı Sistemi Teknolojisinin Gelişimi (Han ve Kamber, 2006)

1960'lı yıllardan itibaren veritabanı ve bilgi teknolojileri basit dosya işlemlerinden özel tasarlanmış ve güçlü veritabanı sistemlerine sistematiksel olarak gelişmektedir. 1970'li yıllardan itibaren veritabanı sistemlerindeki araştırma ve gelişme ilk hiyerarşik ve ağ veritabanı yapılarından ilişkisel veritabanı sistemlerine (verinin ilişkili tablo yapılarında saklandığı sistemdir. Ayrıntılı bilgi için "Data Mining Concepts

and Techniques, Han&Kamber" bkz.), veri modelleme araçlarına ve indeksleme ve erişim metotlarına geçişi sağlamıştır. Bunun yanında, kullanıcılar, sorgulama dilleri, kullanıcı ara yüzleri, sorgu optimizasyonu ve işlem yönetimi sayesinde kullanışlı ve esnek veri erişimine sahip olmuşlardır (Kocabaş, 2010; Han ve Kamber, 2006).

Veritabanı yönetim sistemleri kurulduktan sonra, veritabanı teknolojisi *gelişmiş veritabanı sistemlerinin* geliştirilmesine, *veri ambarlarına* ve *gelişmiş veri analizi* ve *web tabanlı veriler için veri madenciliğine* doğru yönelmiştir. 1980'li yılların ortasından itibaren gelişmiş veritabanı sistemleri üzerinde durulmuştur. Bu sistemler, genişletilmiş-ilişkisel, nesne-yönelik, nesne-ilişkisel ve tümdengelim modelleri gibi yeni ve güçlü veri modellerini birleştirmiştir. Günümüzde mekânsal, zamansal, multimedya, bilimsel ve mühendislik veritabanları, bilgi tabanları ve ofis bilgi tabanları içeren uygulama odaklı veritabanı sistemleri gelişmektedir. Verinin dağıtımı, çeşitlendirilmesi ve paylaşımı ile ilgili konularda yoğun çalışmalar yapılmaktadır (Kocabaş, 2010; Han ve Kamber, 2006).

Son otuz yılda bilgisayar donanım teknolojisinin istikrarlı ve göz kamaştırıcı ilerleyişi güçlü ve uygun fiyatlı bilgisayarları, veri toplama cihazlarını ortaya çıkarmıştır. Bu teknoloji, veritabanı ve bilgi teknolojisine büyük destek sağlamıştır. Ayrıca bu teknoloji, işlem yönetimi, bilgi erişimi ve veri analizi için kullanılabilir veritabanları ve bilgi havuzlarına olanak sağlamıştır. Günümüzde veri, veritabanlarında ve bilgi havuzlarında birçok farklı formatta saklanabilmektedir (Han ve Kamber, 2006).

2.2. Veri Madenciliği Nedir?

Veri bolluğunun güçlü veri analiz araçları ihtiyaçlarıyla birleştiği durum, *veri zengini fakat bilgi yoksunu* durum olarak tanımlanmaktadır (Şekil 2.2). Çok miktarda verinin ve arşivin olduğu böyle bir ortamda önemli kararlar bu verilere göre değil karar vericilerin sezgilerine göre alınmaktadır. Zira karar alıcıların çok yüksek miktardaki bu verinin içinde gömülü değerli bilgiyi çıkarmak için araçlara ihtiyaçları vardır. İşte bu veri ve bilgi arasındaki açığı kapatacak olan yaklaşım veri madenciliği ve veri madenciliği araçlarıdır (Han ve Kamber, 2006).

Veri madenciliği için çeşitli tanımlamalar yapılmıştır. Bunlardan bir kısmı aşağıdaki gibidir:

Veri madenciliği geniş veri tabanlarında bilinmeyen ve beklenmeyen bilgi örüntülerini araştıran bir karar destek sürecidir (Friedman,1997).

Şekil 2.2 Veri Zengini Fakat Bilgi Fakiriyiz (Han ve Kamber, 2006)

Veri madenciliği otomatik öğrenme, örüntü tanıma, istatistik, veritabanı ve görselleştirme tekniklerini bir araya getirerek büyük veritabanlarından bilgi çıkarmaya yarayan bir ara disiplin alanıdır (Cabena ve ark., 1998).

Gartner Grup tarafından yapılan bir diğer tanımlama ise şöyledir: Veri madenciliği büyük veri kümelerinin, önceden akla gelmeyen ilişkileri bulmak ve veriyi hem anlaşılır hem de kullanılabilir hale getirecek biçimde özetlemek için analiz edilmesidir (Han ve ark., 2001).

Veri madenciliği, büyük miktarlardaki verilerde var olan anlamlı örüntü ve kuralların otomatik ve yarı otomatik araçlarla incelenmesi ve analiz edilmesi sürecidir (Berry and Linoff, 2004).

Veri madenciliği örüntü tanıma (pattern recognition) teknolojilerinin yanı sıra istatistiksel ve matematiksel teknikleri kullanarak veri havuzunda depolanan büyük miktardaki verileri dikkatle inceleyerek anlamlı yeni ilişkileri, örüntüleri ve trendleri keşfetme sürecidir (Larose, 2005).

Veri madenciliği, çeşitli mimarilerde depolanmış olan büyük miktarlardaki verilerden ilgi çekici bilginin keşfedilmesi sürecidir (Han ve Kamber, 2006).

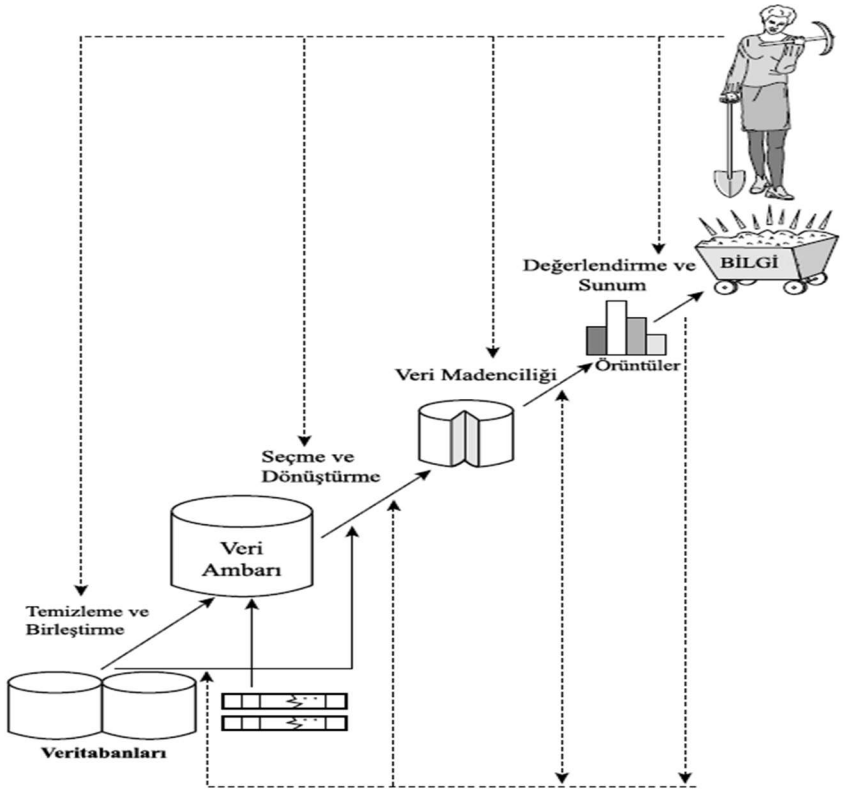
Veri madenciliği deyimi bilim adamları tarafından farklı isimlerle de literatürde kullanılmıştır. Bunlar; Veritabanlarında Bilgi Keşfi (VTBK), Bilgi Çıkarımı, Veri ve

Örüntü analizi, Veri Tarama, Bilgi Keşfi, Veri Avcılığı, Bilgi Üretimidir. Veri madenciliği deyimi yerine en çok kullanılan isim VTBK olmuştur.

VTBK aslında, veriler arasından yararlı bilgileri keşfetme sürecidir ve veri madenciliği VTBK sürecinin önemli bir adımıdır. VTBK süreci ise veri hazırlama, veri seçme, veri temizleme ve veri madenciliği sonucu çıkan sonuçların yorumlanması gibi ek adımlarla birlikte veriden türetilen yararlı bilginin elde edilmesi demektir. Şekil 2.3 VTBK adımlarını göstermektedir (Han ve Kamber, 2006).

Han'ın sunduğu modeldeki VTBK sürecinde yer alan adımlar açıklamaları ile aşağıda belirtilmiştir (Han ve Kamber, 2006);

Veri Temizleme: Gerçek hayatta kullanılan veritabanları içindeki veriler bozulmaya, eksik hale gelmeye ve tutarsız olmaya eğilimlidirler. Bu nedenle verilerin kullanılmadan önce bazı ön işlemlerden geçmeleri gerekir. Ön işlemlerden geçen veriler üzerinde veri madenciliği algoritmalarının uygulanması ile daha kaliteli sonuçlar elde edilir. Bu ön işlemlerden biri veri temizlemedir. Veri temizleme ile veritabanlarındaki eksik, tutarsız ve bozulmuş veriler giderilir.



Şekil 2.3 Veritabanlarında Bilgi Keşfinin Adımları (Han ve Kamber, 2006)

Veri Birleştirme: Veri temizlemeden sonra veri birleştirme işlemi uygulanır. Veri birleştirme, çeşitli kaynaklardan gelen verilerin tek bir veri ambarı altında toplanmasıdır.

Veri Seçme: Veritabanlarında üzerinde işlem yapılacak olan veriler seçilir ve veri türleri bu aşamada belirlenir.

Veri Dönüştürme: Bu aşamada veriler veri madenciliği algoritmalarının uygulanabilmesi için uygun bir formata dönüştürülür. Veri dönüştürme işlemi veri düzeltme, birleştirme, genelleştirme ve normalleştirme gibi işlemlerin bir ya da birkaçını içerebilir.

Veri Madenciliği: Bu aşamada, anlamlı örüntüler elde edebilmek için veri üzerinde veri madenciliği algoritmaları uygulanır. Sınıflandırma, kümeleme algoritmaları gibi veri madenciliği algoritmaları kullanılarak yararlı bilgi keşfedilmesi sağlanır.

Örüntü Değerlendirme: Elde edilmiş olan bilginin basitlik, geçerlilik, yararlılık ve yenilik gibi bazı kriterlere göre değerlendirildiği aşamadır.

Bilgi Sunumu: Bu aşamada, çeşitli görselleştirme ve bilgi sunum araçları kullanılarak elde edilmiş olan bilginin kullanıcıya sunumu gerçekleştirilir.

2.3. Veri Madenciliği Ne Değildir?

İdeal durumda tüm kurumlar faaliyetleri sonucunda elde ettikleri verileri değerlendirerek, kullanılabilir sonuçlar elde etmeyi hedeflemelidirler. Ancak uygulamalara bakıldığında kurumların önemli bir kısmının verileri toplamanın ötesine geçmedikleri gözlenmektedir. Gelişim çizgisine bakıldığında verilerin toplanması (doğru şekilde toplanması) başlangıç noktasıdır. Verilerden yapılan sorgulamalar ve detaylı analizler ile elde edilen sonuçları, veri madenciliği olarak değerlendirmemek gerekir. Bir ölçüde bunlar da veri madenciliğidir ancak daha doğru tanımı veri düzenlemeciliği olarak adlandırılabilir (Argüden ve Erşahin, 2008).

Veri madenciliği; veri toplamak, mevcut verilerden sorgulamalar yapmak veya gelişmiş analiz teknikleri kullanmanın ötesinde bir noktadır. Bir restoran zincirinde; hangi şubelerin ne kadar ciro yaptığı, hangi ürünlerin hangi noktalarda daha fazla satıldığı, hangi saatlerde yoğunluk yaşandığı, gibi analizler veya bir satış şirketinde; hangi müşterilerin devamlılık gösterdikleri, hangi bölgelerde performans düşüklüğü yaşadıklarını belirlemek veri madenciliği değildir. Gelir ile yaş ilişkisinin incelendiği bir değişken, bir sonuç ve az sayıda veriden oluşan bir modeli tanımlayarak, yaşa göre gelir

tahmini yapmak da veri madenciliği değildir. Yüzlerce değişkenin, değişkenler arasında sadece rakamsal değerlerin değil, sıralı (örnek: yüksek-orta-düşük) veya sırasız (örnek: evli-bekar-dul) kategorilerin olduğu, milyonlarca veriye sahip ancak doğru algoritmalar ve güçlü bir bilgisayar ile sonuca ulaşmanın mümkün olduğu modelleri kurmak veri madenciliğidir (Argüden ve Erşahin, 2008).

Gorunescu (2011) tarafından Veri madenciliğinin ne olmadığı ve ne olması gerektiği Çizelge 2.1'de birkaç örnekle gösterilmiştir.

Çizelge 2.1 Veri Madenciliği Ne Değildir? Ne Olmalıdır?

NE DEĞİLDİR	NE OLMALIDIR
İnternetten ayrıntılı bilgi araştırmak	İnternette aynı içerikteki benzer bilgileri gruplamak
Aynı hastalığa sahip hasta kayıtlarını sorgulamak	Benzer semptomlar görülen aynı hastalığa sahip hastaları gruplamak
Yer listesinden termal otellerin yerini sorgulamak	Termal otelleri, hangi hastalığın tedavisi ile ilgili olduğuna göre gruplamak
Şirketlerin finansal raporlarından tabloları analiz etmek	Şirketlerin satış ile ilgili veri tabanlarından müşteri profillerini ortaya çıkarmak

2.4. Veri Ambarları ve OLAP

Veri ambarı gelişmekte olan veri havuzu mimarisidir. Veri ambarı teknolojisi veri temizleme, veri entegrasyonu ve çevrimiçi analitik işlemi (OLAP) içerir. OLAP sayesinde veri analizi, özetleme, birleştirme ve entegrasyon bileşenleri ile çok boyutlu bir şekilde yapılabilir. Tüm bunlara rağmen sınıflandırma, kümeleme, veri niteliğinin zamanla değişimini gözlemleme gibi ayrıntılı analiz yapmak için ek olarak veri analiz araçları gerekmektedir (Han ve Kamber, 2006).

2.4.1. Veri Ambarları

Temel olarak veri madenciliği çalışmaları için ***veri*** ve ***veritabanı*** gerekmektedir; ancak işletmelerde kullanılan işlemsel veritabanları (*transactional database)* doğrudan veri madenciliği uygulamalarında kullanılamaz; bu verilerin veri madenciliği amacıyla kullanılabilmesi için uygun hale getirilmesi gereklidir. İşte belirli bir döneme ait, yapılacak çalışmaya göre konu odaklı olarak düzenlenmiş, birleştirilmiş ve sabitlenmiş veritabanlarına ***veri ambarları*** denilir. Tanımda verilen veri ambarının taşıması gereken özellikler aşağıda kısaca açıklanmaktadır (Silahtaroğlu,2013);

Konu Odaklı: Aynı olayı veya varlığı ilgilendiren veriler birbirlerine bağlanmıştır. Örneğin bir veri ambarı müşteri, ürün vs. gibi varlıklar ya da satış, sipariş alma veya teslimat gibi olaylara yönelik düzenlenmiş olabilir.

Bütünleşiktir: Birden fazla veritabanı bir araya getirilmiş veya veritabanlarına düz dosyalar, İnternet sayfaları vs. gibi kaynaklardaki bilgiler de aktarılmış ve veritabanıyla bütünleştirilmiştir. Bunun yanı sıra, tekrarlanan yer, kişi adları tek bir alanda toplanmış, gerekli dönüştürme ve normalizasyon uygulamaları yapılmıştır.

Belirli Bir Döneme ve Zaman Dilimine Aittir: Barındırılan bilgiler, örneğin son 5 yıllık veya son 10 yıllık dönemlere aittir. Veriler zaman içindeki değişimi gösterecek raporlamaya uygun bir haldedir. Her veri bir şekilde dolaylı veya doğrudan zaman değişkeniyle ilişkilendirilmiştir.

Geçici ve Uçucu Değildir: Veri ambarlarındaki veriler silinmez ve yeni veri eklenmez. Yani veri giriş çıkışı yoktur. İşlemsel veritabanlarından, düz dosyalardan vs. elde edilmiştir; geçmişte belirli bir döneme aittir ve yeni veri giriş çıkışına uygun bir mimaride değildir. Veriler sadece okunabilir bir yapıda tutulmaktadır. Veri ambarında "verilerin yüklenmesi" ve "veriye erişim" olmak üzere sadece iki tür işlemden söz edilmektedir (Silahtaroğlu, 2013).

Özkan (2008)'e göre veri ambarı şu verileri içermelidir:

Metaveri: Veriye ilişkin veri olarak tanımlanabilen metaveri veri ambarlarının en önemli bileşenlerinden biridir. Metaveri karar destek sistemleri analistlerine yardım etmek üzere yaratılan bir dizindir ve veri ambarı içeriğinde neler olduğunu belirtmektedir. İşlemsel çevreden veri ambarına dönüştürülen verinin konumları hakkında bilgileri ve verilerin hangi algoritmaya göre düşük ya da yüksek seviyede özetlendiğini içeren bir kılavuz niteliğindedir.

Ayrıntı veri: Veri ambarında en son olayları içeren ve henüz işlenmediği için diğerlerine oranla daha büyük hacimli ve disk üzerinde saklandığından erişimleri ve yönetimleri pahalı olan verilerdir.

Eski ayrıntı veri: Ayrıntı verinin dışında kalan ve daha eski tarihe ait olan verilerdir. Ayrıntılı veriye göre daha düşük bir ayrıntı düzeyine indirgenerek saklanmaktadır.

Düşük düzeyde özetlenmiş veri: Ayrıntı veriden süzülerek elde edilen düşük seviyede özetlenmiş verilerdir. Veri ambarının tasarımı esnasında hangi verinin özetleneceği ve

özetleme işleminin ne düzeyde olacağı belirlenmelidir.

Yüksek düzeyde özetlenmiş veri: Ayrıntı veri daha yüksek düzeyde özetlenerek, kolayca erişilebilir hale getirilebilir. Bu tür veriler de veri ambarının bir bileşeni olarak yer alabilir.

2.4.2. OLAP (Çevrimiçi Analitik İşleme)

Veri ambarları üzerinde, çeşitli taktik ve stratejik konular hakkında karar vermeye yardımcı olacak veri analizi ve sorgulama işlemlerine ***OLAP*** *(On-Line Analytical Processing)* denilir. Sorgulama, kullanılan tüm işlemsel veritabanları üzerinde de yapılabilir; ancak OLAP sorgulamaları bu tür sorgulamalardan farklıdır. OLAP işlemleri kısaca bilgisayar üzerinde akıl yürüterek işlem yapma olarak tanımlanabilir (Silahtaroğlu, 2013).

OLAP veritabanları üzerinde çeşitli stratejik kararlar almaya yardımcı olacak analiz ve sorgu işlemleridir. Geleneksel sorgu ve raporlama araçları, veritabanında "Ne?" sorusuna yanıt almaya çalışırken OLAP bir kademe daha ilerisine yönelir ve "Niçin?" sorusunu ispatlamak için kullanılır. Örneğin bir analist kredi borcu ödeme güçlüğüne sebep olan risk faktörlerini belirlemek istiyor olsun. Öncelikle düşük gelirli kişilerin kredi riskinin yüksek olacağı şeklinde bir hipotez ileri sürebilir ve veritabanını bunun doğruluğunu göstermek için analiz edebilir. Eğer doğruluğunu ispat edemezse hipotezini değiştirir. Yüksek borç sahibi olmanın risk faktörü olduğunu düşünerek bunu doğrulamaya çalışır. Eğer bunu da doğrulayamazsa her iki faktörün birlikte kredi riskinde etkili olduğu tezini araştırabilir. Yani analist örüntü ve ilişkilerle ilgili bir seri hipotez üretir ve bunların doğruluğunu veya yanlışlığını ispat etmeye çalışır. Bu yüzden OLAP tümdengelimsel bir işlemdir. Ancak incelenmesi gerekli değişken ve parametre sayısı düzinelerce yüzlerce olduğu zaman etkili hipotezler ileri sürmek ve bunları OLAP ile doğrulamak çok daha zorlaşır (Kocabaş, 2010; Silahtaroğlu, 2008).

Veri madenciliği bu açıdan OLAP'dan farklıdır. Çünkü hipotez ileri sürerek bunu doğrulamaya çalışmak yerine doğrudan veriyi bu tip örüntüleri ve ilişkileri açığa çıkarmak için kullanır. Esas olarak Veri Madenciliği tümevarımsal bir yöntemdir. Örneğin, analistin kredi ödeme borcu ödeme güçlüğüne sebep olan risk faktörlerini belirlemek için Veri Madenciliği programı kullandığını varsayalım. Veri Madenciliği programı yüksek borçlu ve düşük gelirli insanların kredi riskinin yüksek olduğunu bulabilir. Ancak daha da fazlasını, analistin hiç hesaba katmadığı bir faktörü, örneğin yaş

faktörünün belirleyici bir faktör olduğunu ortaya çıkarabilir. İşte bu noktada Veri Madenciliği ve OLAP birbirlerini tamamlarlar. Ayrıca OLAP bilgi keşif sürecinin ilk safhalarında tamamlayıcı bir rol oynar. Çünkü verinin araştırılmasına, önemli değişkenlere odaklanarak keşfedilmesine, etkileşimleri bulmaya yardımcı olur (Kocabaş, 2010; Luan, 2002).

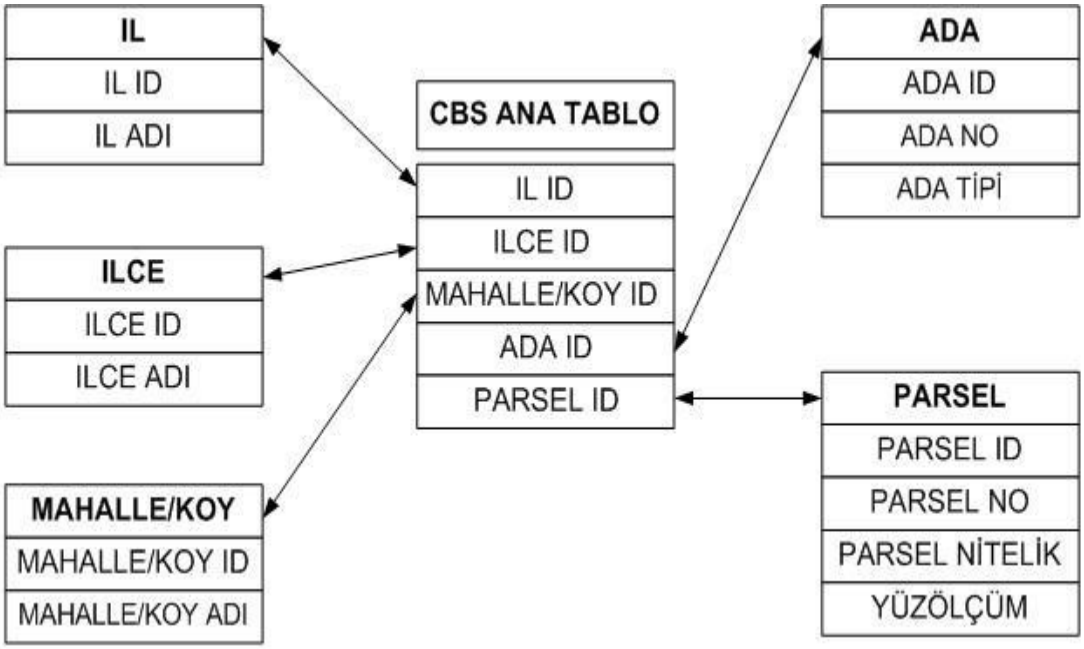
Çizelge 2.2'de görüldüğü üzere OLAP geçmişe ait bilgilendirici veriler verirken Veri Madenciliği geleceğe dönük tahminler üzerinde yoğunlaşmaktadır (Kocabaş, 2010; Apte ve ark., 2002; SPSS, 1999).

Çizelge 2.2 OLAP ile Veri Madenciliği Kavramlarının Karşılaştırılması

OLAP	**Veri Madenciliği**
Postalarımıza geri dönüş oranı nedir?	Gelecekteki postalarımıza yanıt verme potansiyeline sahip müşteri profili nedir?
Yeni ürünümüzden mevcut müşterilerimize ne kadar sattık?	Yeni ürünümüzü hangi müşterilerimiz alma eğilimine sahiptir?
Geçen ay hangi müşterilerimiz poliçelerini yenilemedi?	Önümüzdeki 6 ayda hangi müşterilerimiz rakip firmalara gidebilir?
Geçen yılki en iyi 10 müşterim kimlerdi?	Hangi 10 müşteri en büyük kar profili potansiyeline sahiptir?
Hangi müşteriler geçen yıl borçlarını ödemedi?	Bu müşteri ödeme riskine sahip bir müşteri midir?
Son çeyrekte bölgedeki satış cirosu ne kadardı?	Gelecek yıl bölgedeki satış cirosu tahmini nedir?
Dün üretilen parçaların yüzde kaçı hatalı idi?	Arızalı parçaları azaltmak ve iş çıkarma yeteneğini artırmak için ne yapabilirim?

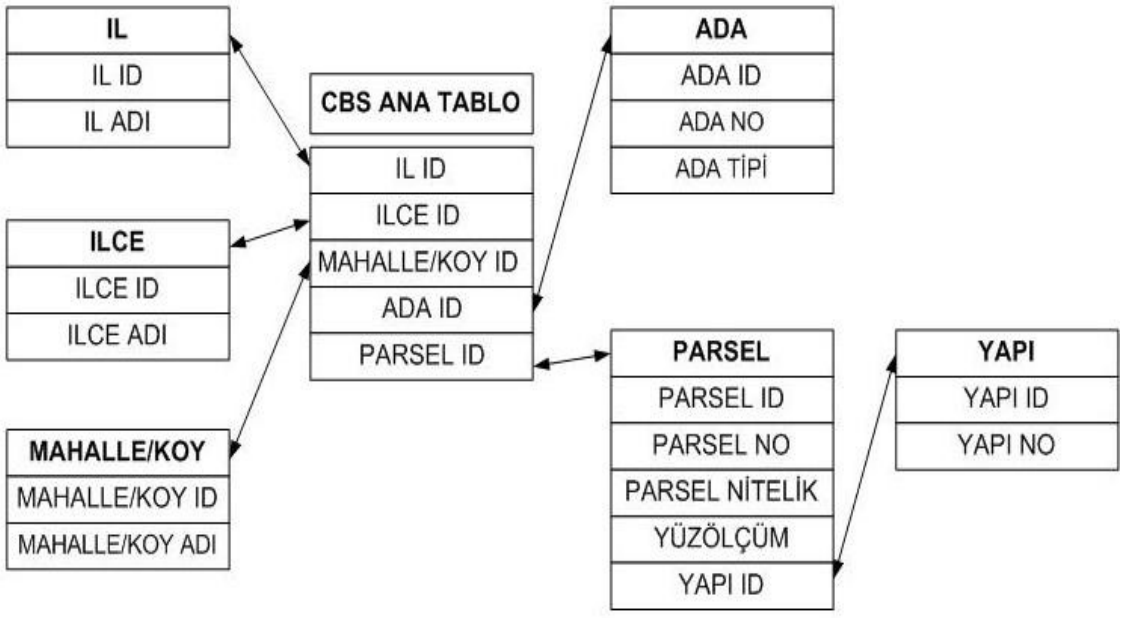
Veri ambarı mimarisi temel olarak üç değişik veri ambarı şemasını kullanır. Bunlar; ***yıldız***, ***kartanesi*** ve ***anatablo*** birliğidir (Silahtaroğlu, 2013).

Yıldız şema türünde, ortada bir ana tablo ve etrafında veri ambarının boyutlarını oluşturduğunu söyleyebileceğimiz boyut tabloları bulunur.



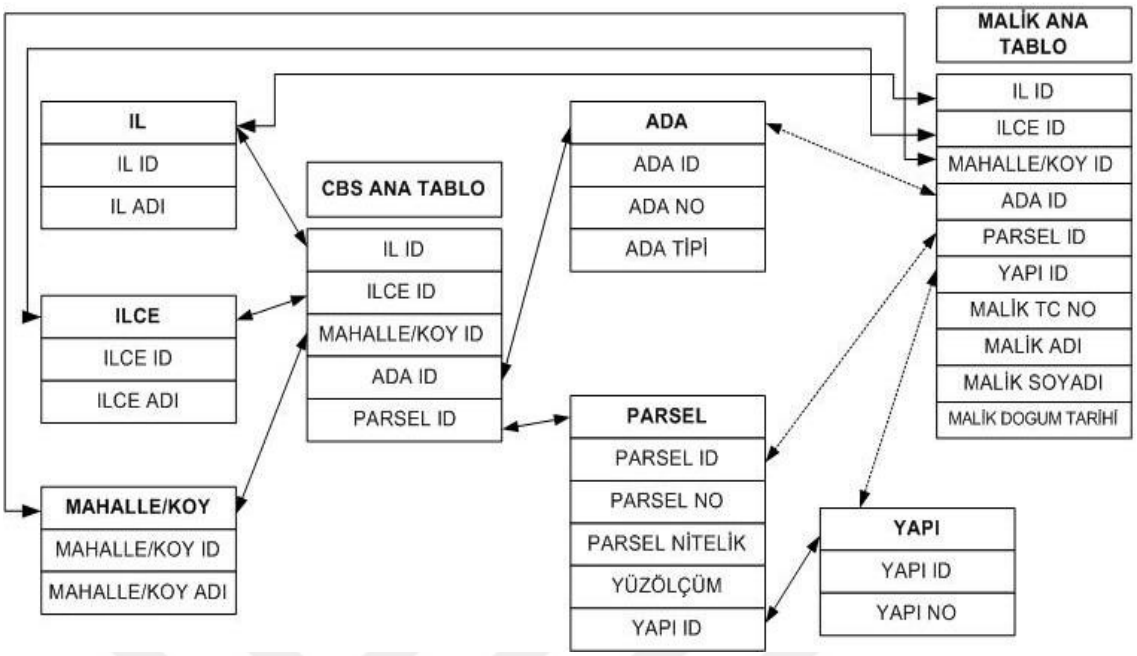
Şekil 2.4 Veri Ambarı Yıldız Mimarisi Örneği (Silahtaroğlu, 2013)

Kartanesi şema türünde, yıldız şema türünden farklı olarak boyut tabloları, normalize edilmiş diğer boyut tablolarına bağlanmıştır. Yine ortada ana tablo durmaktadır.



Şekil 2.5 Veri Ambarı Kartanesi Mimarisi Örneği (Silahtaroğlu, 2013)

Anatablolar birliğindeyse birden fazla ana tablo mevcut boyut tablolarını ortak olarak kullanır ve görünümde birden fazla yıldız şema iç içe monte edilmiş gibidir.



Şekil 2.6 Veri Ambarı Anatablolar Birliği Mimarisi Örneği (Silahtaroğlu, 2013)

2.5. Veri Madenciliğinin Kullanım Alanları

Veri madenciliğinin birçok uygulama alanı vardır. Bu uygulama alanlarından başlıcaları aşağıdaki sıralanmıştır (Akpınar, 2000; Silahtaroğlu, 2013; Han ve Kamber, 2006).

Pazarlama

- Müşterilerin satın alma örüntülerinin belirlenmesi
- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması
- Posta kampanyalarında cevap verme oranının artırılması
- Pazar sepeti analizi (Market Basket Analysis)
- Müşteri ilişkileri yönetimi (CRM: Customer Relations Management)
- Müşteri değerlendirme
- Satış tahmini
- Müşteri dağılımında
- Çeşitli pazarlama kampanyalarında
- Mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerinin oluşturulmasında
- Çapraz satış analizleri
- Çeşitli müşteri analizlerinde

Banka ve Sigortacılık

- Farklı finansal göstergeler arasında gizli korelasyonların bulunması
- Kredi kartı ve sigorta dolandırıcılıklarının tespiti
- Kredi taleplerinin değerlendirilmesi.
- Müşteri dağılımında
- Usulsüzlük tespiti
- Yeni poliçe talep edecek müşterilerin tahmininde
- Riskli müşterilerin örüntülerinin belirlenmesinde

Perakendecilik

- Satış noktası veri analizleri
- Alış-veriş sepeti analizleri
- Tedarik ve mağaza yerleşim optimizasyonu
- Hisse senedi fiyat tahmini
- Genel piyasa analizleri
- Alım-satım stratejilerinin optimizasyonu

Telekomünikasyon

- Kalite ve iyileştirme analizleri
- Hatların yoğunluk tahminleri

Biyoloji, Tıp, Genetik ve Kimya

- Bitki türlerinin ıslahı
- Gen haritasının analizi ve genetik hastalıkların tespiti
- Kanserli hücrelerin tespiti
- Yeni virüs türlerinin keşfi ve sınıflandırılması
- Yeni kimyasal moleküllerin keşfi ve sınıflandırılması
- Yeni ilaç türlerinin keşfinde

Endüstri

- Kalite kontrol analizleri
- Lojistik
- Üretim süreçlerinin optimizasyonu

Yüzey Analizi ve Coğrafi Bilgi Sistemleri

- Bölgelerin coğrafi özelliklerine göre sınıflandırılması
- Kentlerde yerleşim yerlerinin belirlenmesi
- Kentlerde suç oranı, zenginlik-yoksulluk, köken belirlemede
- Kentlere yerleştirilecek posta kutusu, otomatik para makineleri, otobüs durakları gibi hizmetlerin konumlarının tespitinde

Navigasyon Uygulamaları

- Yaya navigasyonu uygulamalarında kullanıcı profilinin belirlenmesinde

Görüntü Tanıma ve Robot Görüş Sistemleri

- Çeşitli algılayıcılar aracılığı ile tespit edilen görüntülerden yola çıkarak engel tanıma, yol tanıma, yüz tanıma, parmak izi tanıma gibi tekniklerde

Uzay Bilimleri ve Teknolojisi

- Gezegen yüzey şekillerinin ve gezegen yerleşimleri, yeni galaksiler keşfinde
- Yıldızların konumlarına göre gruplandırılmasında

Meteoroloji ve Atmosfer Bilimleri

- Bölgesel iklim, yağış haritalarını oluşturmada
- Hava tahminleri, ozon tabakası deliklerinin tespitinde
- Çeşitli okyanus hareketlerinin belirlenmesinde

Metin Madenciliği

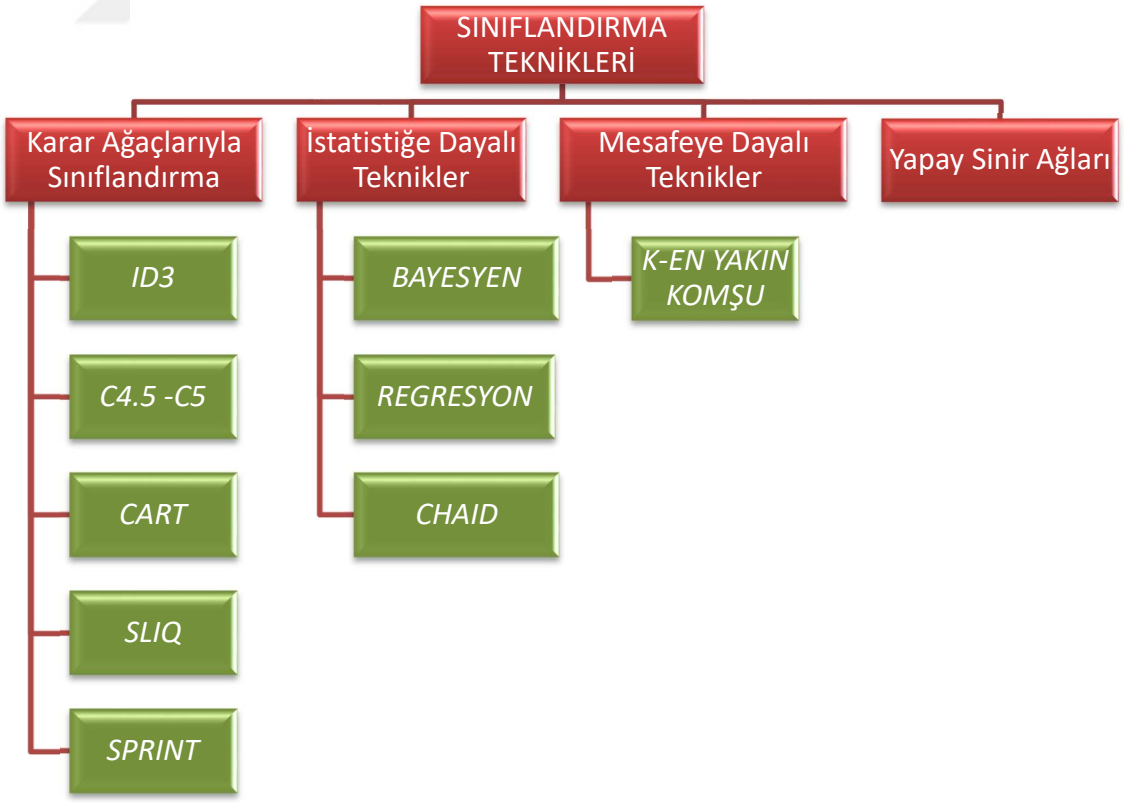
Çok büyük ve anlamsız metin yığınları arasından anlamlı ilişkiler elde etmekte kullanılmaktadır.

2.6. Veri Madenciliği Modelleri ve Teknikleri

Veri madenciliği modelleri veri madenciliğinin görevlerinde bulunan örüntülerin çeşitlerini belirlemekte kullanılır. Genellikle veri madenciliği modelleri tahmin edici (predictive) ve tanımlayıcı (descriptive) olmak üzere iki ana başlık altında toplanır. Tanımlayıcı (Descriptive) madencilik modelleri veritabanındaki verilerin genel özelliklerini ortaya çıkarırken, Tahmin edici (Predictive) madencilik modelleri tahminlerde bulunmak için geçerli olan veriler üzerinde sonuçlar çıkarmaya çalışır (Han ve Kamber, 2006).

Veri madenciliği modelleri gördükleri işlevlere göre üç ana başlık altında toplanır;

Sınıflama ve Regresyon Modelleri: En yaygın uygulanan Veri Madenciliği (VM) tekniklerinden biri olan sınıflama, sınıfı tanımlanmış mevcut verilerden yararlanarak sınıfı belli olmayan verilerin sınıfını tahmin etmektir. Sınıflandırma bir öğrenme algoritmasına dayanmaktadır ve öğrenmenin amacı bir sınıflandırma modelinin yaratılmasıdır. Öğrenme sırasında tüm veriler kullanılmamaktadır. Sınıflandırma öğrenme ve sınıflara atama olmak üzere iki aşamadan oluşmaktadır. Öncelikle bir adet bağımlı (sınıf ya da hedef değişken olarak da adlandırılır) ve birden çok bağımsız değişkenden oluşan veri kümesi, öğrenme ve test kümesi olmak üzere ikiye ayrılır. İlk aşamada algoritma öğrenme kümesi üzerinde çalışır ve öğrenme işlemini gerçekleştirir. İkinci aşamada ise test kümesini kullanarak sınıflandırma kuralları belirlenir ve yaratılan sınıflandırma modeli test verilerine uygulanarak, doğruluk oranına göre modelin doğru sınıflandırma yapıp yapmadığı sınanır. Eğer doğruluk oranı kabul edilebilir ise elde edilen model yeni veri kümelerinin sınıflandırılmasında kullanılabilmektedir (Han ve Kamber, 2006; Silahtaroğlu, 2013; Özkan, 2008). Literatürde yer alan başlıca sınıflandırma teknikleri Şekil 2.7'de gösterilmektedir;



Şekil 2.7 Sınıflandırma Teknikleri

Birliktelik Kuralları ve İlişki Analizi: İş, bilim, mühendislik, sağlık vb. sektörlerin veritabanlarındaki bilgi miktarındaki artışı bu sektörlerin sahip oldukları bilgi arasındaki ilişkiyi ortaya çıkarmaya yönlendirmiştir. Bu şekildeki büyük bilgi yığınları arasından elde edilecek ilişkiler sektörler için kıymetli sonuçlar doğurabilecek bu sektörlerde alınacak kararlarda önemli rol oynayacaktır. İlişki analizi ile veritabanında yer alan bir bilginin diğer kayıtlı bilgilerle olan bağlantısını açıklar. Örneğin; bir müşterinin bir marketten bir ürün satın alırken, bu ürünle birlikte diğer ürün veya ürünleri satın alınması yönündeki bağlantıyı ortaya koyar. Bu tür ilişkilerin ortaya çıkarılması ve bunun kural olarak ortaya konması birliktelik kuralları ve ilişki analizi konusuna girer. Bu tür çalışmalara çeşitli literatürde "pazar sepeti analizi" denilir (Silahtaroğlu, 2013). İlişki analizlerinin yapılıp birliktelik kurallarının ortaya çıkarılmasında en çok bilinen ve kullanılan algoritma Apriori algoritmasıdır.

Kümeleme Analizi: Kümeleme analizinden nesneler "küme içi benzerlikleri artır, kümeler arası benzerlikleri azalt" prensibine dayalı olarak kümelere ayrılmaktadır. Böylece aynı küme içindeki benzerlikler maksimum, farklı kümeler arası benzerlikler ise minimum olacaktır.

Sınıflama ve regresyon modelleri **tahmin edici (predictive),** kümeleme ve birliktelik kuralları modelleri **tanımlayıcı (descriptive)** modellerdir (Akpınar, 2000).

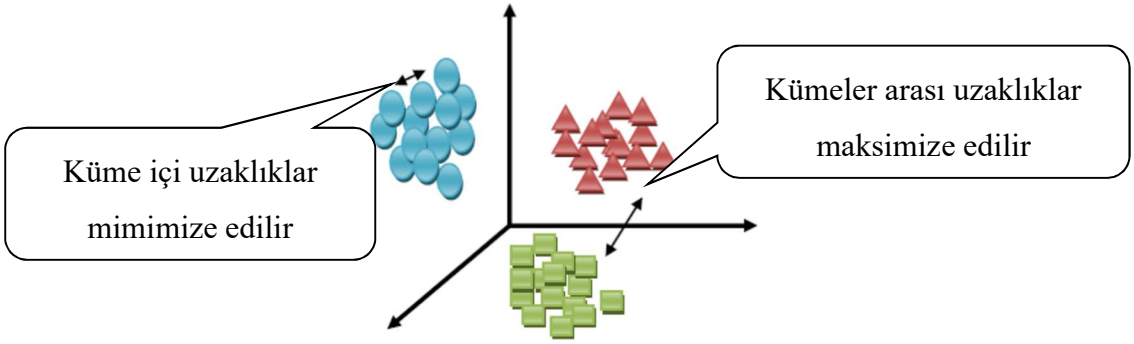
2.7. Kümeleme Analizi

Bir veri kümesindeki nesnelerin belirli yakınlık kriterine göre gruplara ayırma işlemine Kümeleme analizi denilmektedir. Kümeleme analizi işlemi oluşma her bir gruba da "küme" denilmektedir. Kümeleme analizine kısaca kümeleme denilmektedir. Kümeleme yabancı literatürde kaynaklarda *clustering* ya da *segmentation* olarak adlandırılmaktadır. Kümeleme en basit tanımıyla benzer özellik gösteren veri elemanlarının kendi aralarında gruplara ayrılmasıdır. Aynı küme içindeki elemanların benzerliği fazla, kümeler arası benzerlik ise az olmalıdır (Dinçer, 2006).

Sınıflandırma işleminde, sınıflar önceden belirli olduğu için bu yöntem gözetimli sınıflandırma yöntemidir. Kümeleme yönteminde ise sınıflar önceden belirli olmadığı için gözetimsiz sınıflama yöntemidir. Verilerin hangi kümelere, kaç değişik gruba ayrılacağı eldeki verilerin birbirlerine olan benzerliğine göre belirlenir.

Kümeleme modellerinde amaç; özellikleri birbirine benzeyen verilerin

aralarındaki uzaklıklar en az olacak şekilde kümelenmesi, özellikleri birbirinde çok farklı olan verilerin aralarındaki uzaklık en çok olacak şekilde kümeler oluşturulması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir.



Şekil 2.8 Kümelemenin Amacı

Kümeleme analizinin belli başlı özellikleri aşağıda açıklaması ile birlikte verilmiştir (Han ve Kamber, 2006).

- **Ölçeklenebilir olmalıdır.** Kümeleme algoritmaları 200' den az veri nesnelerini içeren küçük veri kümelerinde iyi bir şekilde çalışırken, büyük veri kümeleri üzerinde çok iyi bir şekilde çalışmayabilir. Bu gibi durumlarda ölçeklendirme algoritmalarına ihtiyaç vardır.
- **Farklı nesne tiplerine göre çalışabilmelidir**. Kümeleme algoritmaları sayısal verilerle çalışmak içi geliştirilmiş olsa da sayısal olmayan, ikili (binary) ve kategorik veri tipleriyle çalışabilmelidir.
- **Düzgün şekilli olmayan kümeler de bulabilmelidir.** Birçok kümeleme algoritması Manhattan ve Öklit uzaklık ölçümlerine göre kümelere karar vermektedir. Uzaklık ölçümlerine dayalı olan algoritmalar benzer boyut ve yoğunlukta olan küresel kümeler bulmaya eğilimlidirler. Buna rağmen kümeler herhangi bir şekilde olabilirler. Düzgün şekillerde olmayan kümeleri bulabilen algoritmaları geliştirmek önemlidir.
- **En az miktarda giriş değişkeni gerektirmelidir.** Kümeleme algoritmaları ideal bir kümeleme işlemini gerçekleştirmesi için mümkün olduğunca kullanıcıdan bağımsız olması ve minimum sayışa giriş parametresi gerektirmelidir.
- **Gürültü içeren verileri de kullanılabilmelidir.** Gerçek hayatta kullanılan birçok veritabanı eksik, tanımlanmamış ve aykırı veriler içerir. Kümeleme algoritmaları

bu tür verilere karşı oldukça duyarlıdır ve bu tür veriler zayıf kalitede kümeler üretilmesine sebep olabilirler.

- **Verilen parametrelerin sırasına duyarsız olmalıdır.** Kümeleme işlemi veritabanındaki hangi veriden başlanırsa başlansın aynı kümeleme sonucunu vermelidir.
- **Çok boyutlu veritabanları ile çalışabilmelidir.** Veritabanı veya veri ambarları birçok boyut ve nitelik içerebilirler. Birçok kümeleme algoritması düşük boyutlu veriyi kullanmakta iyidir. İnsan gözü en çok 3 boyutlu veriyi anlayabilecek yapıdadır. Fakat kümeleme algoritması daha fazla boyutta çalışabilmelidir.
- **Veri kümesinin sahip olduğu kısıtlamalar dikkate alınmalıdır.** Gerçek dünya uygulamaları çeşitli kısıtlamalar altında kümeleme işlemini yapılabilmesine ihtiyaç duyar. Örneğin; belirli sayıda yeni ATM makineleri için yerleri seçmemiz gerektiğini düşünelim. Bu yerlere karar vermek için, yol ağları ve her bölgenin müşteri gereksinimleri gibi kısıtlamaları dikkate almak gereklidir. Burada yapılması gereken, belirtilen kısıtlamaları tatmin eden iyi bir kümeleme yaparak verinin gruplarını bulmaktır.
- Kolay yorumlanabilen ve kullanılabilen sonuçlar üretebilmelidir.

Mevcut kümeleme algoritmaları ideal bir kümeleme algoritmasından istenen bu özelliklerin tamamına sahiptir değildir. Kümeleme analizi ile ilgili çalışmalar devam etmektedir ve bu özelliklerin olabildiğince tamamını içinde barındırabilecek algoritmaların geliştirileceği umulmaktadır (Han ve Kamber, 2006).

2.7.1. Kümeleme Analizi Veri Türleri

Kümeleme analizinde veri yapısı matris formundadır. Kümeleme işleminde kullanılan matrisler iki temel gruba ayrılır (Han ve Kamber, 2006):

- **Veri Matrisi (Data Matrix):** Bu tip veri yapısında *n* tane nesne, *p* tane değişken olur. Örneğin nesneler insanları temsil ediyorsa, değişkenler; bir insanın ağırlık, boy ve yaşını temsil etmektedir. Bu matris aşağıdaki denklem de gösterilmiştir.

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

- **Farklılık Matrisi (Dissimilarity Matrix):** Nesnelerin diğer nesnelerle olan uzaklık bilgilerinin tutulduğu nxn boyutunda bir matristir. Bu matrisin genel ifadesi aşağıdaki denklem de gösterilmiştir.

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

Veri madenciliği uygulamalarında çoğunlukla farklılık matrisi kullanılır. Farklılık matrisi tek modlu matris olarak bilinir. Çünkü nesneler arası uzaklık fonksiyonu değişme özelliğine sahip olduğundan farklılık matrisinin asal köşegeni altında ve üstünde kalan değerler simetriktir. Veri matrisi ise bu değişme özelliğine sahip olmadığından iki modlu matris olarak bilinir (Han ve Kamber, 2006).

2.7.1.1. Aralık ölçekli değişkenler (interval-scaled variables)

Tam olarak kesin belirlenmiş değerlerden çok, belli bir aralık seklinde belirlenen verilerde geçerlidir. En sık kullanılan ağırlık ölçekli değişkenler boy, ağırlık, genişlik ve uzunluk verileridir. Ölçümde kullanılan birim çok önemlidir. Birimin değişmesi, analizin sonucunu etkiler. Sonucun kafa karıştırıcı olmaması için analize giren verilerin de standart olması gerekir. Standartlaştırmadan sonra farklılık matrisi ile analiz yapılır (Han ve Kamber, 2006].

Aralık ölçekli veriler için uzaklık ölçümlerini hesaplamada ***Öklid (Euclidean)***, ***Manhattan*** ve ***Minkowski*** formülleri kullanılır (Han ve Kamber, 2006):

Öklid Uzaklığı: En sık kullanılan yöntemdir. İki ya da daha çok boyutlu düzlemde kolaylıkla kullanılabilir ve

$$\mathrm{d(i,j)} = \sqrt{\left(\mathrm{x_{i1}} - \mathrm{x_{j1}}\right)^2 + \left(\mathrm{x_{i2}} - \mathrm{x_{j2}}\right)^2 + \cdots + \left(\mathrm{x_{in}} - \mathrm{x_{jn}}\right)^2} \qquad (2.1)$$

ifadesi ile verilmektedir. Burada *i* ve *j* ifadeleri *p* boyutlu veri nesnelerini temsil etmektedir.

Manhattan Uzaklığı: *p* boyutlu uzayda herhangi iki noktanın karşılıklı her bir koordinat değerinin farkı alınarak bulunur ve bu ifade

$$\mathrm{d(i,j)} = \left|\mathrm{x_{i1}} - \mathrm{x_{j1}}\right| + \left|\mathrm{x_{i2}} - \mathrm{x_{j2}}\right| + \cdots + \left|\mathrm{x_{ip}} - \mathrm{x_{jp}}\right| \qquad (2.2)$$

olarak verilmektedir.

Minkowski Uzaklığı: Öklid ve Manhattan uzaklığının genelleştirilmiş hali olarak,

$$d(i,j) = \left(\left|x_{i1} - x_{j1}\right|^q + \left|x_{i2} - x_{j2}\right|^q + \cdots + \left|x_{ip} - x_{jp}\right|^q\right)^{1/q} \qquad (2.3)$$

şeklinde ifade edilir. *q* bir pozitif tam sayı olmak üzere bu ifade *q*=1 için Manhattan uzaklığını, *q*=2 için Öklid uzaklığını belirtir. *q* değişkeninin değeri artırıldıkça daha hassas uzaklık ölçüm ifadeleri elde edilir (Han ve Kamber, 2006).

2.7.1.2. İkili değişkenler (Binary variables)

Bir ikili değişkenin "0" ve "1" olmak üzere iki durumu vardır. "0" yok, "1" var anlamında kullanılır. Aralık ölçeklinin tersine, kesin ve net sonuçların olduğu analizlerde kullanılır. Örneğin; bir kişinin sigara içip içmediğine yönelik sorulan sorunun karşılığı; eğer içiyorsa "1", içmiyorsa "0" dır. Örnekte de görüldüğü gibi cevap olarak bir aralık çıkmamakta ve kesin bir cevap alınmaktadır (Han ve Kamber, 2006).

İkili değişkenler verisi için olasılık tablosu aşağıda gösterilmiştir:

Çizelge 2.3 İkili Değişkenler İçin Olasılık Tablosu

		Nesne j		
		1	*0*	*toplam*
Nesne i	*1*	*Q*	*r*	*q+r*
	0	*S*	*t*	*s+t*
	toplam	*q+s*	*r+t*	*p*

q, Ortak olan "1" lerin sayısını belirtmektedir. *r*, ilk nesne için "1", ikinci nesne için "0" olanların sayısını belirtmektedir. *s*, ilk nesne için "0", ikinci nesne için "1" olanların sayısını ifade etmektedir. *t*, ortak olan "0"ların sayısını ifade etmektedir.

Simetrik ikili değişkenler için uzaklık ölçüsü:

$$d(i,j) = \frac{r+s}{q+r+s+t} \qquad (2.4)$$

Asimetrik ikili değişkenler için uzaklık ölçüsü:

$$d(i,j) = \frac{r+s}{q+r+s} \tag{2.5}$$

Jaccard katsayısı, asimetrik ikili değişkenler için benzerlik ölçüsüdür.

$$sim_{jaccard}(i,j) = \frac{q}{q+r+s} \tag{2.6}$$

2.7.1.3. Kategorik, ordinal ve oran değişkenler

Kategorik (Categorical) değişkenler: Kategorik değişkenler, ikili değişkenlere benzeyen ve çok sayıda seçeneği olan değişkenlerdir. Örneğin renk değişkeni kategorik bir değişkense kırmızı, yeşil, mavi, pembe ve sarı durumlarına sahip olduğunu düşünebiliriz (Han ve Kamber, 2006).

Kategorik değişkenin durumlarının sayısı M olsun. Durumlar; 1, 2, ..., M gibi tamsayı kümesi, sembol ve harflerle ifade edilebilir. Tamsayılar özel bir sıralama olmadan veriyi kontrol etmek için kullanılır. Örneğin map color kategorik değişkenini oluşturmak için, yukarıda listeden her bir renk için bir ikili değişkeni yaratılabilir. Sarı rengine sahip bir nesne için, sarı değişkeni "1" e ayarlanır, kalan 4 değişkende "0"a ayarlanır. Kategorik değişkenler olarak tanımlanan nesneler arasında farklılığın hesaplanması için aşağıdaki formül kullanılır:

$$d(i,j) = \frac{p-m}{p} \tag{2.7}$$

Formüldeki *p* değişkeni *i* ve *j* nesnelerinin sahip olduğu toplam özellik sayısını, *m* değişkeni *i* ve *j* değişkenlerinde aynı anda yer almış olan özellik sayısını ifade eder (Han ve Kamber, 2006).

Ordinal değişkenler: Ordinal değişkenler, kategorik değişkenlerde olduğu gibi sonlu sayıda farklı durum içerir. Kategorik değişkenlerden farklı olarak ordinal değişkenlerde sıra önemlidir. Örneğin yarışmalarda en yüksek dereceye sahip olan yarışmacıya altın, daha sonrakine gümüş ve üçüncü olan yarışmacıya da bronz madalya verilir (Han ve Kamber, 2006).

Ordinal değişkenler arası farklılığı hesaplamak için, ordinal değişkenlerin alabileceği değerleri [0-1] aralığında sayı değerleri alabilecek şekilde standartlaştırıp aralık ölçekli değişkenlerde kullanılan mesafe yöntemleri kullanılır (Han ve Kamber,

2006).

Oran Ölçekli (ratio-scaled) Değişkenler: Üstel olarak artan verilerin benzerliğinin bulunmasında kullanılır. Oran ölçekli değişkenlere bakteri popülasyonlarında büyüme ve radyoaktif elementin yarı ömrünün ölçüm sonuçları örnek olarak verilebilir. Oran ölçekli değişkenlerin genel yapısı aşağıdaki gibidir:

$$Ae^{Bt} \text{ yada } Ae^{-Bt} \tag{2.8}$$

Denklemdeki *A* ve *B* pozitif sabitlerdir. Oran ölçekli değişkenlerde nesneler arasındaki farklılığı hesaplamak için üç farklı metot vardır:

a) Bu yöntemde; oran ölçekli değişkenler aralık ölçekli değişkenler gibi davranırlar. Bu yöntem iyi bir seçim değildir. Çünkü ölçülen aralık doğrusal olmadığından ölçümün hatalı olması olasıdır.

b) Oran ölçekli değişkenlere logaritmik ölçümler uygulanabilir.

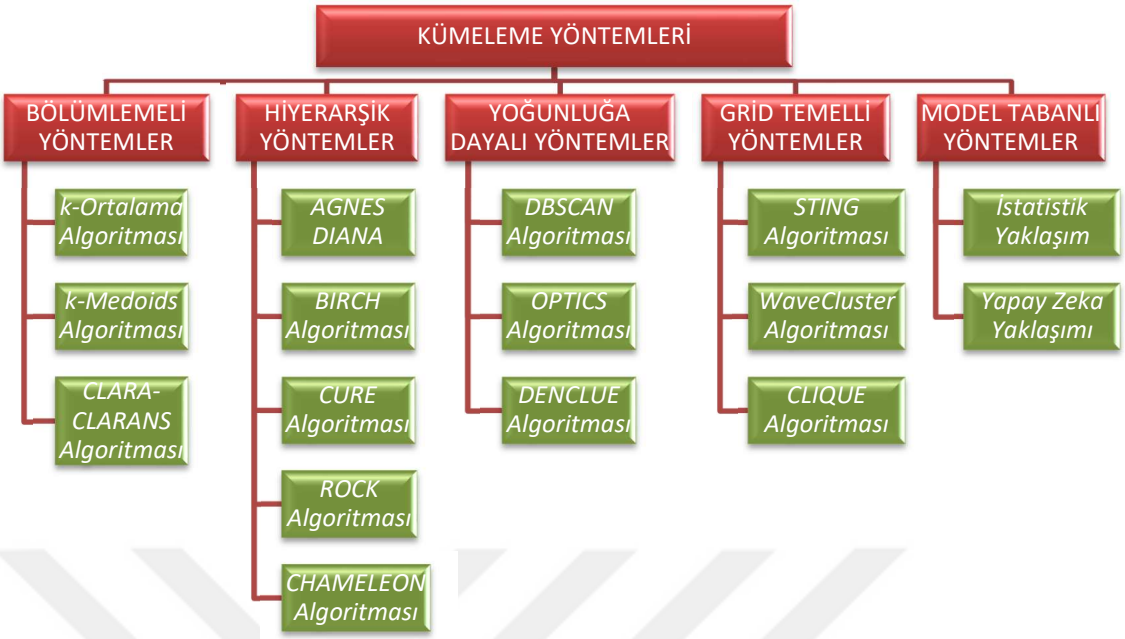
$$y_{if} = \log(x_{if}) \tag{2.9}$$

Bu formül kullanıldığında elde edilen y_{if} değeri ile aralık ölçekli değişken olarak işlem yapılabilir.

c) Bu metotta, oran ölçekli değişkenler sürekli ordinal değişkenler olarak düşünür ve ordinal değişkenlerdeki uzaklık hesaplamaları kullanılır (Han ve Kamber, 2006).

2.7.2. Kümeleme Yöntemleri

Kümeleme yöntemlerinin sınıflandırılması literatürde kullanılan en genel ayrım *hiyerarşik* ve *hiyerarşik olmayan* kümeleme yöntemleri ayrımıdır. Kümeleme yöntemlerinin sınıflandırılması Şekil 2.9'da gösterilmiştir.



Şekil 2.9 Kümeleme Yöntemleri

2.7.2.1. Bölümlemeli Yöntemler

Bölümlemeli yöntemlerde *n* adet nokta önceden verilen *k* küme sayısına ($k<n$) göre kümelere ayrılır. Oluşturulacak küme sayısı önceden belirlidir. Kullanıcı algoritmaya kümeler arasındaki minimum/maksimum mesafeyi ve kümelerin iç benzerlik kriterlerini de vermek zorundadır (Silahtaroğlu, 2013).

2.7.2.1.1. k-Ortalama Algoritması

İlk olarak 1967 yılında Mac Quenn tarafından ortaya atılan bu algoritma sürekli olarak kümelerin yenilendiği ve en uygun çözüme ulaşana kadar devam eden döngüsel bir algoritmadır. k-Ortalama algoritmasının genel mantığı *n* adet veri nesnesinden oluşan bir veri kümesini, araştırmacının ön bilgisine ve tecrübesine dayanarak belirlenen *k* adet kümeye bölümlemektir. Amaç, gerçekleştirilen bölümleme işlemi sonunda elde edilen kümelerin küme içi benzerliklerini maksimum ve farklı kümeler arası benzerliklerin minimum olmasını sağlamaktır. Algoritmanın kaba kodu şu şekildedir (Silahtaroğlu, 2013);

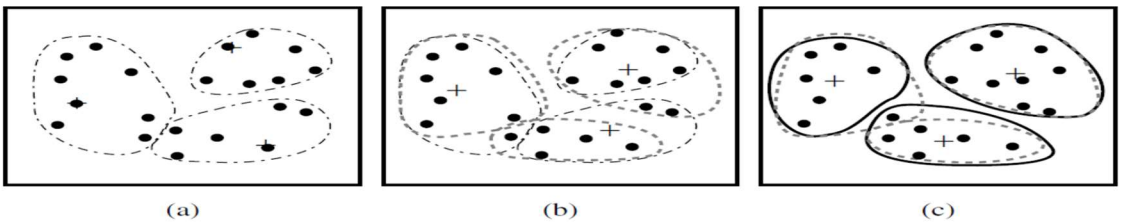
Girdiler: D = {t1, t2, ..., tn} // eldeki veritabanı

K // verilen küme sayısı

Adımlar:

1. Keyfi olarak m1, m2, ..., mk ortalama belirle (Eldeki veritabanından rastgele)
2. Her bir t_i yi en yakın olduğu $m_i$nin kümesine ata
3. Kümelere ait m1, m2, ..., mk değerlerini yeniden hesapla
4. Küme elemanlarında herhangi bir değişiklik yoksa dur
5. İlk adımdan itibaren tekrar et.

Çıktı K adet küme



Şekil 2.10 k-Ortalama Algoritması ile Kümeleme (Han ve Kamber, 2006)

Şekil 2.10'da bir nesne setinin k-Ortalama metodu ile kümelenmesi gösterilmiştir. Her bir kümenin orta değeri "+" ile işaretlenmiştir (Han ve Kamber, 2006)].

2.7.2.1.2. k-Medoids Algoritması

1990 yılında Kauffman ve Rousseuw tarafından geliştirilen bu algoritma k adet kümeyi bulmak için seçilen temsilcilerin (medoid) etrafına ana kümedeki tüm elemanları toplayarak ve her defasında bu temsilcileri değiştirerek kümeleme işlemini tamamlar.

Temsilci (medoid) seçiminden kasıt kümenin merkezine yakın mesafede bulunan noktanın belirlenmesidir. K adet küme için seçilen k adet temsilci belirlendikten sonra, veritabanındaki temsilci olmayan diğer noktalar (veriler) kendilerine en çok benzeyen temsilcinin etrafında toplanır (Silahtaroğlu, 2013).

K-Medoids algoritmasının işlem basamakları aşağıdaki gibidir (Akın,2008):

Adım 1: K küme sayısının belirlenmesi.

Adım 2: Başlangıç medoidleri olarak k nesnelerinin seçimi.

Adım 3: En yakın medoid x'e sahip kümeye, kalan nesneleri atamak

Adım 4: Amaç fonksiyonunu hesaplamak. (Hata kareler kriteri: en yakın medoidler için bütün nesnelerin uzaklıklarının toplamı)

Adım 5: Tesadüfî olarak medoid olmayan y noktasının seçimi.

Adım 6: Eğer x ile y'nin yer değiştirmesi amaç fonksiyonunu minimize edecekse

bu iki noktanın (x ile y) yerini değiştirmek.

Adım 7: Değişiklik olmayana kadar Adım 3 ile Adım 6 arası işlemler tekrarlanır.

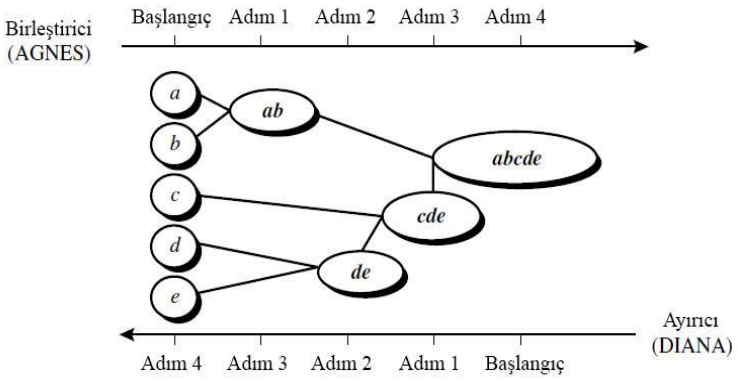
2.7.2.2. Hiyerarşik Yöntemler

Hiyerarşik kümeleme metodu veri nesnelerini ağaç yapısı içinde gruplamaya çalışır. Hiyerarşik kümeleme metotları hiyerarşik ayrışmanın aşağıdan yukarı (birleştirme) yada yukarıdan aşağıya (ayırma) formuna bağlı olarak *birleştirici* yada *ayırıcı* olmak üzere sınıflandırılırlar (Han ve Kamber, 2006).

2.7.2.2.1. AGNES - DIANA Hiyerarşik Kümeleme

Genellikle hiyerarşik kümeleme metotlarının iki türü vardır.

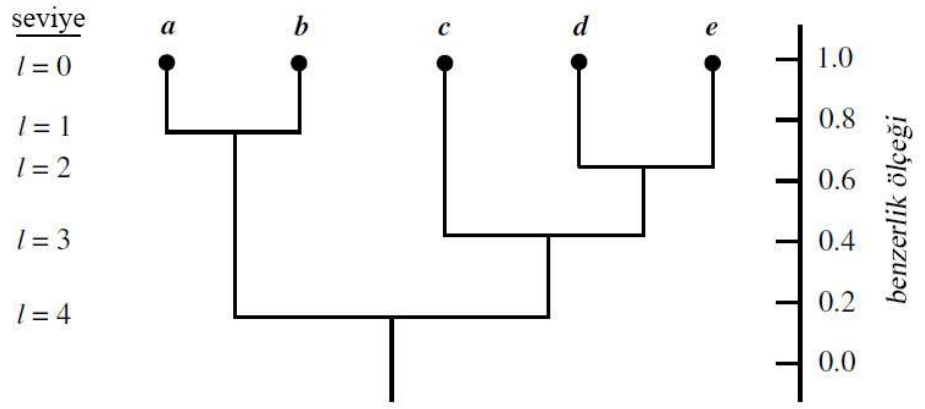
- **Birleştirici hiyerarşik kümeleme (AGNES):** Aşağıdan yukarıya doğru çalışan bir strateji izler. Başlangıçta her nesne ayrı bir küme olarak kabul edilir. Algoritmanın bir sonraki her adımında bu atomik kümelerden benzerlik gösterenler tek bir küme oluncaya kadar yada istenen özellikleri sağlayana kadar birleştirilirler. Hiyerarşik kümeleme metotlarının çoğu bu kategoride yer almaktadır. Diğerleri sadece küme içi benzerliklerin tanımlanmasında farklıdır (Han ve Kamber, 2006).
- **Ayırıcı hiyerarşik kümeleme (DIANA):** Yukarıdan aşağı çalışan bir strateji izler. Başlangıçta verilen nesnelerinin tümü bir küme olarak kabul edilir. Algoritmanın bir sonraki her adımında kendi aralarında benzerliklerin en yüksek olan nesneler bir araya getirilerek büyük küme daha küçük kümelere bölünür. Bu kümeleme işlemi her nesne kendi başına bir küme oluşturana kadar, istenen küme sayısı yada her kümenin çapının belirli bir eşik değerin altında olması gibi istenen özellikler elde edilinceye kadar devam eder (Han ve Kamber, 2006).



Şekil 2.11 {a,b,c,d,e} veri nesneleri üzerinde Birleştirici ve Ayırıcı Hiyerarşik Kümeleme (Han ve Kamber, 2006)

Şekil 2.11'de ***AGNES*** ve ***DIANA*** hiyerarşik kümele algoritmalarının uygulamaları 5 adet ***{a,b,c,d,e}*** nesnesi olan bir veri setinde gösterilmektedir. Başlangıçta, AGNES her nesneyi ayrı bir küme olarak kabul eder. Daha sonra kümeler bazı kriterle göre adım adım birleştirilir. DIANA' da ise tüm nesneler başlangıçta bir küme olarak kabul edilir. Daha sonra küme, küme içinde yakın komşuluk ilişki olan objeler arasından maksimum öklid uzaklığı gibi bazı kriterlere göre kümelere ayrılır. Küme bölme işlemi, her yeni küme sadece bir nesne içerene kadar devam eder (Han ve Kamber, 2006).

Hiyerarşik kümelemenin sürecini göstermek üzere ***dendrogram*** olarak adlandırılan bir ağaç yapı kullanılır. Dendrogram nesnelerin adım adım nasıl gruplandığını gösterir. Aşağıdaki şekilde 5 nesne için bir dendrogram gösterilmiştir. l=0 seviyesinde nesneler birer tekil küme olarak görülmektedir. l=1 seviyesinde a ve b nesneleri birleşerek ilk kümeyi oluşturmuşlardır ve sonraki seviyeler birlikte kalmışlardır. Ayrıca kümeler arasındaki benzerlik ölçeğini göstermek üzere dikey eksen kullanılabilir. Örneğin; *{a,b}* ve *{c,d,e}* nesne grupları arasındaki benzerlik yaklaşık 0.16 dır. Bu nesne grupları tek bir küme oluşturmak için bir araya getirilmiştir (Han ve Kamber, 2006).



Şekil 2.12 {a,b,c,d,e} Veri Nesnelerinin Hiyerarşik Kümelenmesi İçin Dendrogram Gösterim (Han ve Kamber, 2006)

Kümeler arasındaki uzaklık için aşağıdaki belirtilen dört yaygın ölçüt kullanılmaktadır;

$|p - p'|$: p ve p' nesneleri yada noktaları arasındaki uzaklık

$m_i ve m_j$: $C_i ve C_J$ kümeleri için ortalama

$n_i ve n_j$: $C_i ve C_J$ kümelerindeki nesne sayısı

olmak üzere,

Tek bağlantılı (Single Linkage) kümeleme metodunun uzaklık formülü;

$$\text{Minimum uzaklık: } d_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'| \quad (2.10)$$

Tam bağlantılı (Complete Linkage) kümeleme metodunun uzaklık formülü;

$$\text{Maksimum uzaklık: } d_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} |p - p'| \quad (2.11)$$

Merkez tabanlı kümeleme metodunun uzaklık formülü;

$$\text{Ortalama uzaklık: } d_{mean}(C_i, C_j) = |m_i - m_j| \quad (2.12)$$

Ortalama bağlantılı kümeleme metodunun uzaklık formülü;

$$\text{Ortalama uzaklık: } d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'| \quad (2.13)$$

şeklindedir (Han ve Kamber, 2006).

Hiyerarşik kümeleme yöntemi, basit olarak düşünülse de birleştirme veya bölme noktalarının seçimi konusunda zorluklarla karşılaşılır. Bu gibi bir karar önemlidir çünkü bir nesne grubu bir kez bölünmüş yada birleşmiştir ve bir sonraki adımdaki işlemler yeni oluşturulan kümeler üzerinden yürütülecektir. Ne önceki işlemleri ne de kümeler arasındaki nesneleri değiştirmek mümkündür. Bu yüzden birleştirme yada bölme kararları, bazı adımlarda seçilmezse düşük kaliteli kümelerin oluşmasına neden olacaktır. Ayrıca, metot ölçeklenebilirliği iyi değildir, bu yüzden her birleştirme ya da ayırma kararı küme yada nesne sayısının iyi bir incelenmesini ve değerlendirmesini gerektirir. Hiyerarşik kümeleme yöntemlerinin kalitesini artırmak için diğer kümeleme teknikleriyle hiyerarşik kümeleme tekniklerinin entegre edilmesi ve çok aşamalı kümelemeyle işlemin sonuçlandırılması gerekir (Han ve Kamber, 2006).

2.7.3. Küme Geçerliliği Teknikleri

Kümeleme Analizinin en kritik konusu küme sayısına karar vermektir. Kümeleme analizi uygulamalarında doğru küme sayısı çoğunlukla bilinemez. Kümeleme analizinin sonuçlarının kalitesini değerlendirmek için küme geçerliliği tekniklerine ihtiyaç vardır. Bu teknikler arasında en çok kullanılanları Dunn indeksi ve Davies-Bouldin indeksidir.

2.7.3.1. Dunn İndeksi

Dunn Geçerlilik indeksinin temel varsayımı, kümelerin yoğun ve iyi dağılmış olmalarıdır. Dunn Geçerlilik indeksi D katsayısı ile gösterilir, D katsayısı büyüdükçe,

küme kalitesi ve sayısı artmaktadır.

$$\mathbf{D} = \min_{1\le i\le n}\left\{\left\{\min_{\substack{1\le j\le n\\ i=j}}\frac{d(c_i,c_j)}{\max_{1\le k\le n}(d'(c_k))}\right\}\right\} \tag{2.14}$$

Denklemde $\boldsymbol{d(c_i, c_j)c_i}$ ve $\boldsymbol{c_j}$ kümeleri arasındaki mesafeyi temsil eder. $\mathbf{max}(\boldsymbol{d'(c_k)}k$ kümesinin noktaları arasındaki en uzak mesafeyi ve *n* küme sayısını belirtmektedir. Algoritmanın amacı kümeler arası mesafeyi en küçüğe çekerken, küme içi mesafeyi maksimumda tutmaktır. Bu sayede elde edilecek D değeri yükseldikçe optimum küme sayısına yaklaşılmış olacaktır (Silahtaroğlu, 2013).

2.7.3.2. Davies-Bouldin İndeksi

Davies-Bouldin Geçerlilik indeksi DB katsayısıyla gösterilmektedir.

$$\mathbf{DB} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{max}\left\{\frac{S_n(Q_i)+S_n(Q_j)}{S_n(Q_i,Q_j)}\right\} \tag{2.15}$$

Denklemdeki *n* küme sayısını, S_n kümenin elemanlarının küme merkezine olan uzaklıklarının ortalamasını ve $S_n(Q_i, Q_j)$ iki küme merkesi arasındaki uzaklığı temsil etmektedir. Bu durumda DB değerinin düşük olması kümelerin kendi içinde homojen ve kümelerin birbirlerinden uzak olduğunu belirtir.

Dunn değerinin Yüksek, Davies-Bouldin değerinin küçük olması küme kalitesinin iyi olduğunu göstermektedir. Değerin büyük ya da küçük olduğunu söyleyebilmek için en az iki senaryo halinde kümeleme yapılması ve her bir senaryo için bu indeks değerlerinin hesaplanması gerekir (Silahtaroğlu, 2013).

3. MATERYAL VE YÖNTEM

Bu tez çalışmasında mekânsal verilerin analizinde veri madenciliği tekniklerinde bir olan kümeleme analizi yönteminin kullanılabilirliğinin incelenmesi amaçlanmıştır. Mekânsal veri olarak Türkiye'nin 81 iline ait trafik kaza istatistik verileri kullanılacaktır. Mekânsal verilerin kümeleme analizi AGNES, k-means ve k-medoids algoritmalarıyla gerçekleştirilecek ve SPSS, RapidMiner ve MultiDendrograms yazılımlarından yararlanılacaktır. ArcGIS Desktop yazılımında kümeleme analizi sonuçları kullanılarak çok değişkenli tematik harita üretimi gerçekleştirilecek ve çok değişkenli haritalar incelenerek kümeleme başarısı açısından AGNES, k-means ve k-medoids algoritmalarından hangi yöntemin daha uygun olduğu değerlendirilecektir.

3.1. SPSS (Statistical Package for the Social Sciences)

SPSS yazılımı ilk sürümü 1968 yılında kullanıma sürülmüş istatiksel analize yönelik bilgisayar programıdır. Uzun bir dönem ABD asıllı SPSS Inc. firması tarafından hazırlanıp kullanıma sürülen yazılım 2009 yılında IBM şirketine satılmıştır. Ağustos 2010 yılından itibaren IBM SPSS Statistics isimlendirilmektedir. SPSS özellikle “Sosyal Bilimler” dalında istatistiksel analiz için kullanılmaktadır. Pazar araştırmacıları, sağlık araştırmacıları, anket şirketleri, devlet kurumları, eğitim araştırmacıları, veri madencileri vb. tarafından pratik olarak kullanılan bir yazılımdır. Kullanımı kullanıcı dostu grafiksel arayüz ile kolaylaştırılmıştır. Ayrıca makro dilleri yardımıyla kullanıcı kendi istekleri doğrultusunda programı yönlendirebilmektedir. SPSS programının Windows, Mac OS X ve Linux işletim sistemleri için farklı sürümleri mevcuttur (URL1, URL2).

3.2. RapidMiner

RapidMiner, eski adıyla YALE (Yet Another Learning Enviroment) ilk kez 2001 yılında Ralf Klinkenberg, Ingo Mierswa ve Smon Fischer tarafından Dortmund Teknik Üniversitesi, yapay zekâ biriminde geliştirilmiştir. 2006 yılında itibaren, Ingo Mierswa ve Ralf Klinkenberg tarafından kurulan Rapid-I firması tarafından geliştirilmeye başlanmıştır. 2007 yılında yazılımın adı YALE’den RapidMiner’a çevrilmiştir. RapidMiner makine öğrenmesi, veri madenciliği, metin madenciliği, tahmin edici analiz ve iş analizi amaçlarına yönelik olarak geliştirilmiş yazılım platformudur. Kullanıcı dostu grafik bir ara yüze sahiptir. Veri Analizi, Ön İşleme, Sınıflama, Kümeleme, Birliktelik Kuralları Çıkarımı, Nitelik Seçimi işlevlerine sahiptir. Oracle, MS SQL Server, PostgreSQL, MySQL, JDBC, Sybase, Access, IBM DB2 veritabanlarını ve metin

dosyalarını desteklemektedir. Windows, Mac OS X ve Linux işletim sistemlerinde çalışabilmektedir (URL3, URL4).

3.3. MultiDendrograms

MultiDendrograms yazılımı ROVIRA i VIRGILI Üniversitesi'nde Sergio GOMEZ ve arkadaşları tarafından geliştirilen ve açık kaynak lisansı altında kullanıcılarına sunulan, gerçek verileri Hiyerarşik Kümeleme yapmak için kullanılan basit ama güçlü bir programdır. MultiDendrograms yazılımı bir Uzaklık (veya benzerlik) matrisinden başlar, en yaygın Birleştirici (Agglomerative) Hiyerarşik Kümeleme algoritmalarını kullanarak dendrogramını hesaplar, birçok grafik gösterim parametresinin ayarlanmasını sağlar ve sonuçları kolayca dosyaya aktarabilir. Kullanıcı dostu grafik ara yüze sahiptir. Yazılım Java yazılım dilinde geliştirilmiş olup Windows, Mac OS ve Linux işletim sistemlerinde çalışabilmektedir (URL5).

3.4. ARCGIS Desktop

ARCGIS Desktop (ArcInfo, ArcView ve ArcEditor) içerisinde bütünleşik olarak gelen ArcMap, ArcCatalog, ArcToolbox, ArcGlobe ve Model Builder ara yüzleri ile haritalama, coğrafi analizler, veri güncelleme, veri yönetimi ve görüntüleme işlemlerini gerçekleştirilebildiği ESRI tarafında geliştirilen entegre bir coğrafi bilgi sistemi yazılımıdır. ARCGIS Desktop Extensions (Modüller) kullanılarak bütün yazılımlara yeni yetenekler eklenebilir. ArcObjects (ARCGIS yazılım bileşenleri kütüphanesi) kullanılarak özel modüller geliştirilebilir. Ayrıca Visual Basic, .Net, Java, Visual C++ gibi programlama ara yüzleri kullanılarak yeni modüller ve özel araçlar da geliştirilebilir. Yazılım Windows ve Linux işletim sistemlerinde çalışabilmektedir (URL6).

3.5. Çok Değişkenli Haritalar (Multivariate Mapping)

Çok değişkenli haritalama mekânsal objelerin birden fazla özelliğinin harita üzerinde gösterilmesidir. Birden fazla özelliğin eş zamanlı gösterimi mekânsal objelerin farklı özellikleri dikkate alınarak karşılaştırılmalarına olanak sağlar. Çok değişkenli haritalama için kartografik gösterimi sağlamak amacıyla farklı bilgisayar destekli yöntemler geliştirilmiştir (Buckley, 2008).

Çok değişkenli haritalamada her bir özellik için ayrı bir harita mı yapılacağına yoksa birçok özelliğin aynı haritada mı gösterileceğine karar vermek gerekir. Her bir özellik için ayrı harita yapılması, çeşitli özelliklere sahip iki objenin karşılaştırılmasını

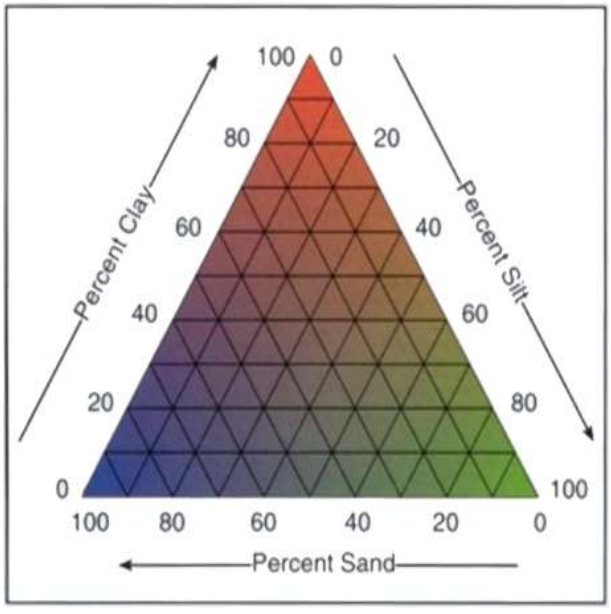
zorlaştırabilir. Bu nedenle birçok özelliğin aynı haritada gösterildiği yöntemler daha çok tercih edilmektedir. Birçok özelliği aynı harita üzerinde gösterilmesinde kullanılan bazı haritalama yöntemleri;

- Üç Değişkenli Koroplet Haritalar (Trivariate Choropleth Maps)
- Çok Değişkenli Nokta Haritalar (Multivariate Dot Maps)
- Çok Değişkenli Noktasal İşaret Haritalar (Multivariate Point Symbol Maps)
- Farklı İşaretlerin Birleştirilmesi (Combining Different Symbols)
- Ayrılabilir ve Bütünleyici İşaretler (Seperable and Integral Symbols)

şeklindedir (Slocum ve ark., 2009).

3.5.1. Üç Değişkenli Koroplet Haritalar

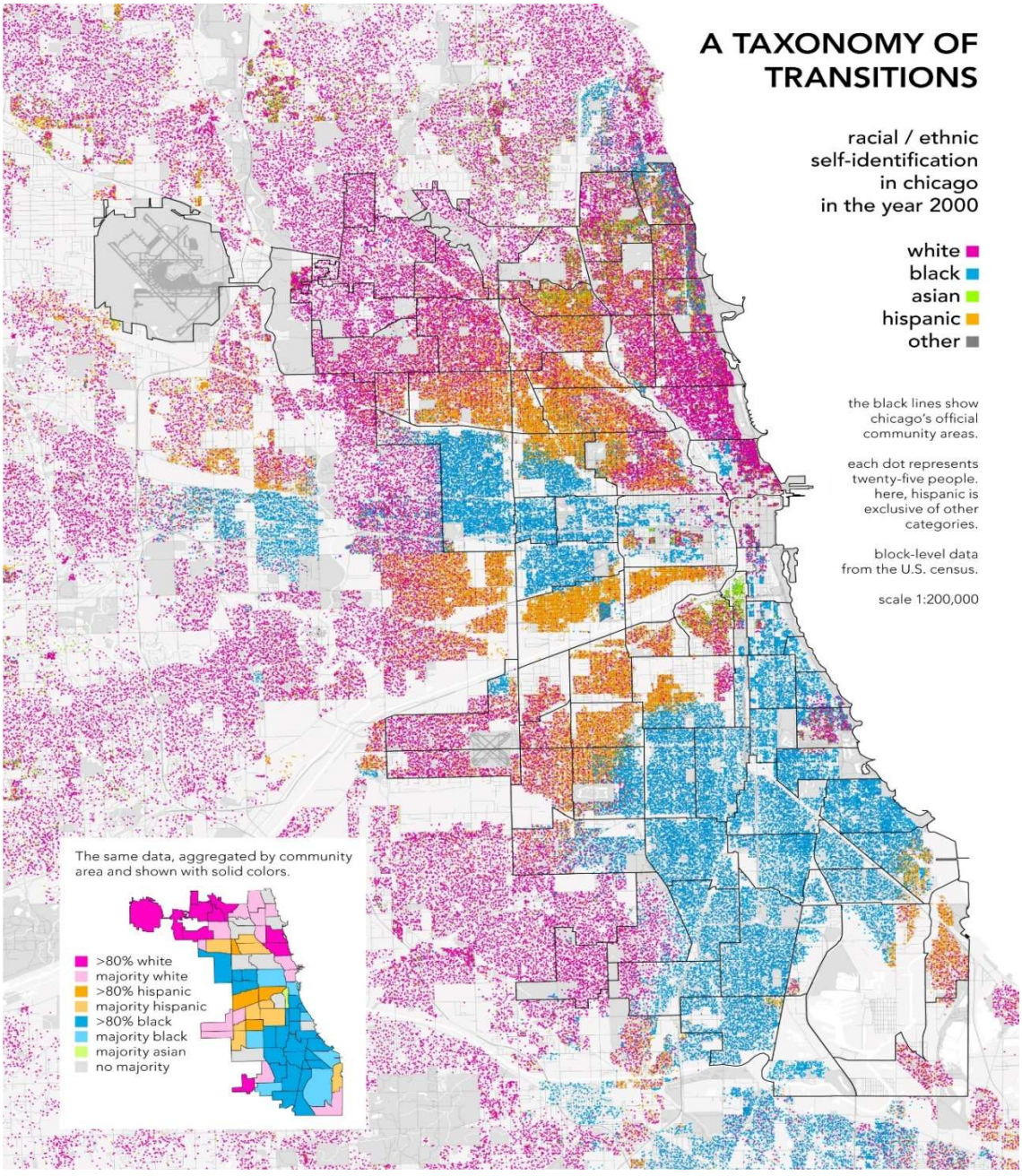
Üç değişkenli koroplet harita, iki değişkenli koroplet haritaların üst üste çakıştırılmasıyla elde edilmektedir. Ancak bu yaklaşım ideal olarak yüzde yüze tamamlayan üç özellik için kullanılmalıdır. Örneğin; bir toprağın kum, alüvyon ve kil yapısal özellikleri yada bir ülkede seçime giren Cumhuriyetçi, Demokrat ve Bağımsız siyasal partilerin seçim sonuçları yüzdesel olarak tanımlanabilir. Üç özelliğe renk atama da CMY, RGB yada Kırmızı-Mavi-Sarı Ana renk yaklaşımları kullanılabilir. Bir toprağın yapısal özelliği tanımlayan üç özelliğe RGB yaklaşımıyla renk atanmasının sonuçları Şekil 3.1'de görülmektedir (Slocum ve ark., 2009).



Şekil 3.1 Üç Değişkenli Koroplet Harita Oluşturmada RGB Renk Şeması (Slocum ve ark., 2009)

3.5.2. Çok Değişkenli Nokta Haritalar

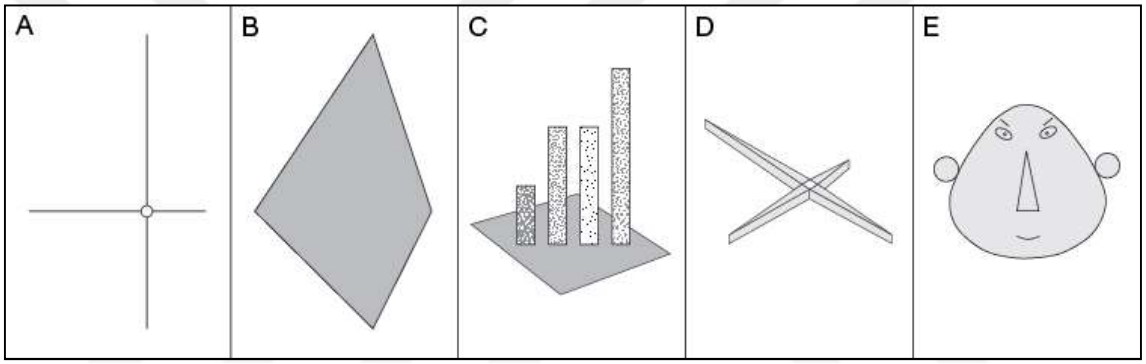
Çok değişkenli nokta haritalar da her bir özellik için belirli bir renk veya şekildeki işaret kullanılmaktadır (Jenks, 1953). Örneğin; Şekil 3.2'deki çok değişkenli nokta haritasında her bir nokta ile 25 kişi gösterimi yapılırken bu kişilerin ırk/etnik özellikleri farklı renk tonlarıyla ifade edilmiştir. Bu haritadan da görüleceği üzere çok değişkenli nokta haritalama yöntemiyle birbirleriyle ilişkili olan değişkenlerin yada özelliklerin mekânsal dağılımları başarılı bir şekilde gösterilmektedir.



Şekil 3.2 Çok Değişkenli Nokta Haritalama Örneği (Rankin, 2009)

3.5.3. Çok Değişkenli Noktasal İşaret Haritaları

Çok değişkenli verinin bir noktasal işaret kullanılarak gösterildiği yönteme Çok Değişkenli Noktasal İşaretleme denilmektedir. Çok değişkenli noktasal işaretler her ne kadar noktasal olayların anlatımı için kullanılsa da, çok değişkenli alansal sembolleri üretmenin zorluğundan dolayı alansal olayların anlatımı için de kullanılmaktadır. Bir bütünü tamamlayan ve aynı birimle ifade edilen birden fazla ilişkili özelliği (related attributes) anlatmada pasta grafik işareti kullanımı yaygındır. Aynı birimle ifade edilmeyen ve bir bütünün parçası olmayan birden fazla ilişkisiz özelliği (Nonrelated attributes) anlatmada ise Şekil 3.3'de görülen ve "Glif" adı verilen çok değişkenli noktasal işaretleri kullanılmaktadır.



Şekil 3.3 Çok Değişkenli Noktasal İşaret Örnekleri (Slocum ve ark., 2009)

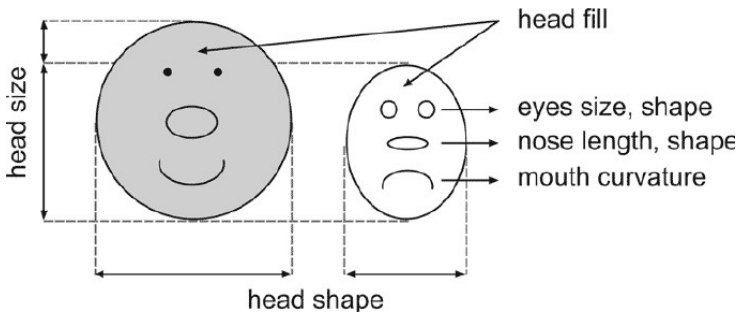
Şekil 3.3A. Çok değişkenli ışın glifi yada yıldız işareti, bir iç çemberden her bir özellik için orantılı olarak uzayan ışınlarla oluşturulmaktadır.

Şekil 34.3B Işın glifinin yada yıldız işaretinin uç noktaları bir çokgenle bağlanırsa çok köşeli glif yada kartanesi işareti oluşturulur.

Donna Cox ve meslektaşları tarafından 1990 yılında Illinois Üniversitesi'nde birçok yeni çok değişkenli noktasal işaretleri geliştirilmiştir. Şekil 3.3C'de görülen üç boyutlu bar grafik işareti bunlardan biridir. Üç boyutlu bar grafik işaretinde barları yüksekliği çeşitli özelliklerin büyüklüğüyle orantılı yapılırken, Cox'un uygulamasında her bir barın farklı desenler kullanılarak farklı renklerde gösterilmiştir.

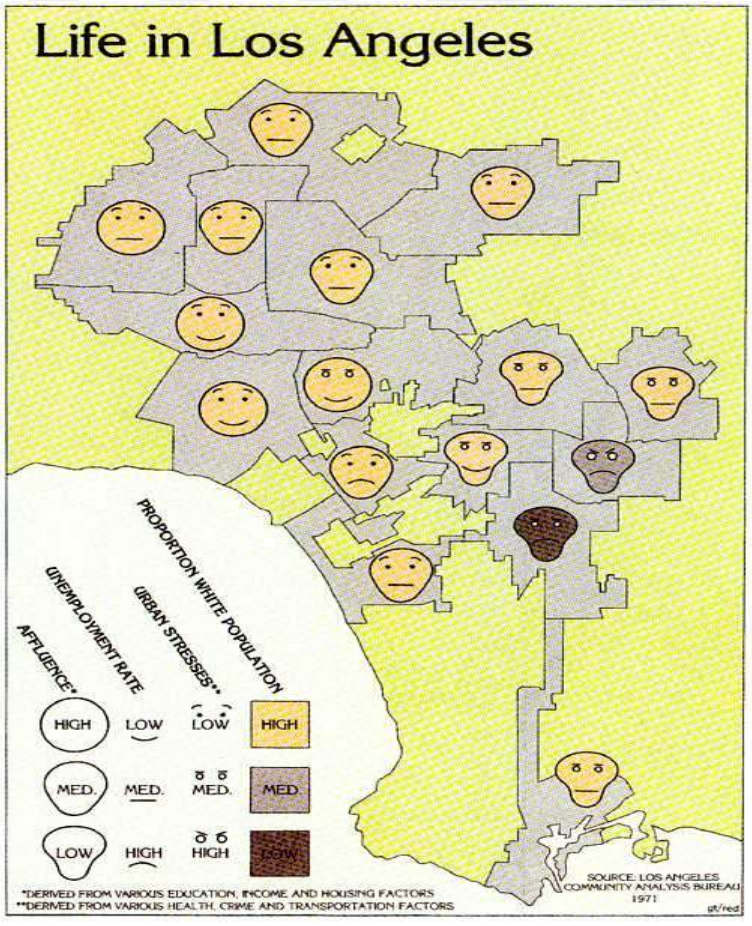
Ellson tarafından geliştirilen bir diğer noktasal işaret de Şekil 3.3D'de görülen Veri Jakı'dır. Veri Jakı, kare bir merkezi alandan her bir niteliğin büyüklüğü ile orantılı olarak uzayan üçgenlerden oluşmaktadır. Bar grafiklerde olduğu gibi veri jaklarının sivri uçları farklı renklerde gösterilirse kolaylıkla ayırt edilebilir.

Çok değişkenli noktasal işaretlerde en dikkat çekicisi Şekil 3.3E'de görülen, Hermann Chernoff tarafından ilk olarak 1973 yılında tasarlanan ve farklı yüz özelliklerinin çeşitli niteliklerle ilişkilendirildiği Chernoff yüzüdür (Reyes, 2009).



Şekil 3.4 Chernoff Yüzünün Altı Değişkeni (Reyes, 2009)

Chernoff yüzü Reyes'ın araştırmalarına göre; yüzün şekli, yüzün büyüklüğü, yüzün dolgusu, göz büyüklüğü ve şekli, burun şekli ve uzunluğu ve ağız şekli olmak üzere en fazla altı değişkenle kullanılmaktadır (Şekil 3.4). Şekil 3.5'de Chernoff yüzünün "Life in Los Angeles, 1970" haritasında uygulaması görülmektedir.



Şekil 3.5 Çok Değişkenli "Life in Los Angeles, 1970" Haritası (Reyes, 2009)

3.5.4. Farklı İşaretlerin Birleştirilmesi

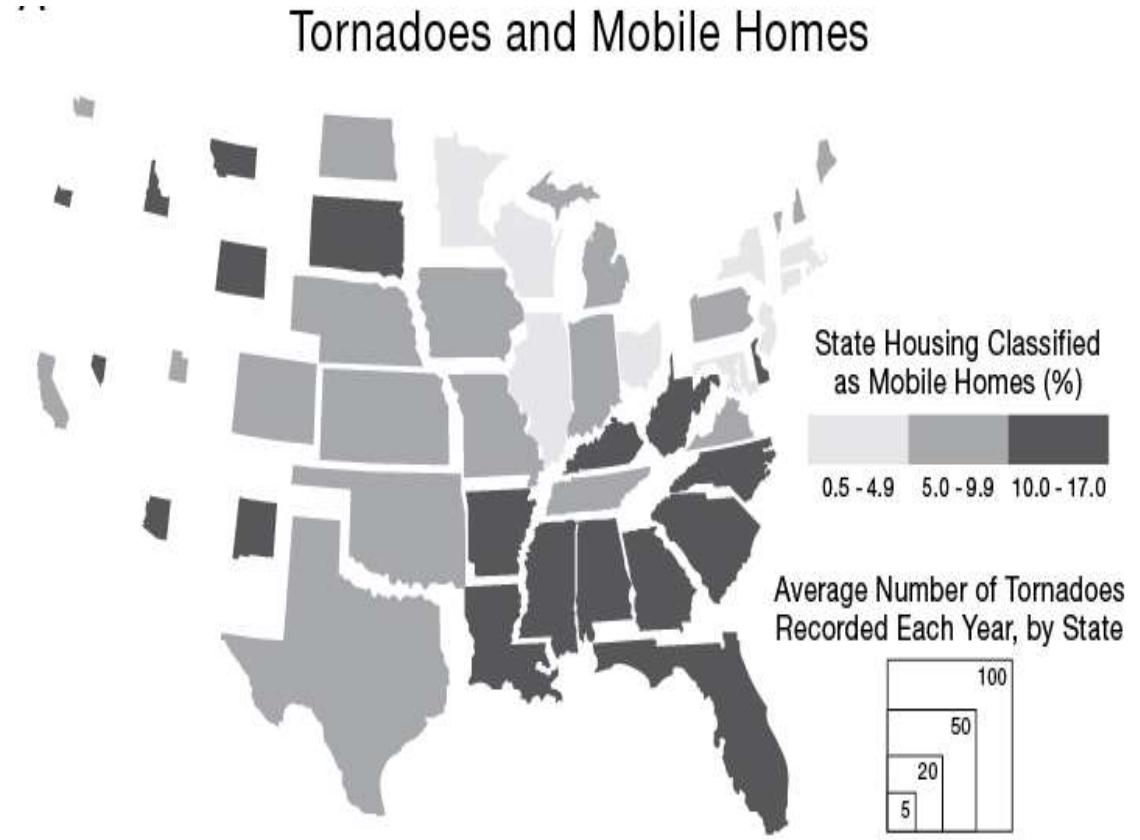
Farklı işaretlerin birleştirilmesi yöntemi, çok değişkenli verinin çeşitli işaretlerin birleştirilmesiyle gösterilmesidir (DiBiase, 1994). Bu yönteme en iyi örnek David DiBiase ve meslektaşlarının iklim modellerinden üretilen verileri keşfetmekte kullandıkları SLCViewer yazılımıdır. SLCViewer yazılımı analiste dört iklim özelliğini küçük katmanlar halinde görüntülemesine ya da çok değişkenli bir harita oluşturmak için nokta, çizgi ve alan işaretlerinin eklenmesine izin vermektedir (DiBiase ve ark., 1994).

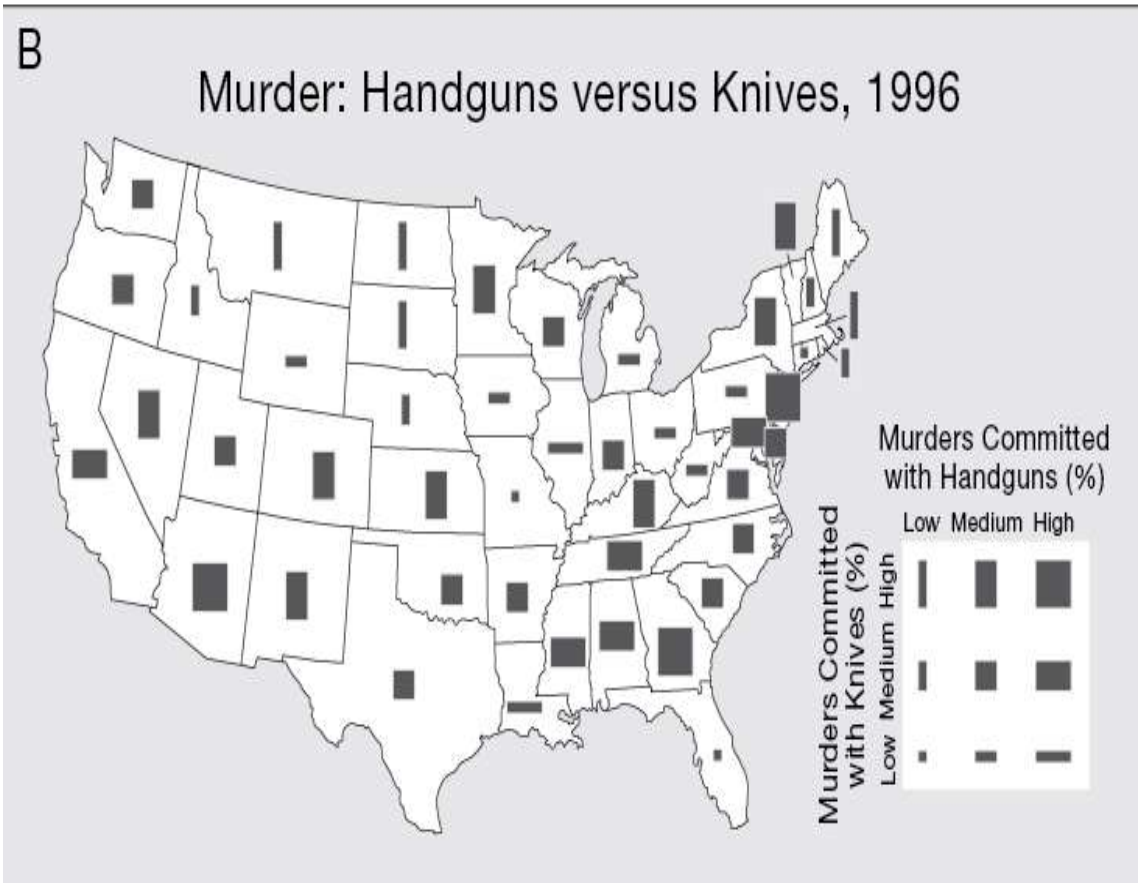
3.5.5. Ayrılabilir (Seperable) ve Bütünleyici (Integral) İşaretler

Çok değişkenli veriler tek bir haritada gösterildiğinde, her bir özelliğe ait işaretlerin ayrılabilir veya bütünleyici olup olmadığı sorusu ortaya çıkmaktadır. **Ayrılabilir işaretler,** haritada tek olarak gösterilebilirler ve böylelikle harita kullanıcısına ayrı özelliklere odaklanma imkânı vermektedir. Kartografya farklı büyüklükteki ve farklı renkteki işaretler ayrılabilir işaret olarak düşünülmektedir. **Bütünleyici işaretler,** harita kullanıcılarının konuyla bütünleşmesi sağlayan işaretlerdir. Bu işaretler haritalarda tek başlarına gösterilemezler. Bütünleyici işaretler haritalarda özellikler arasındaki ilişkiyi incelemek için kullanılırlar (Slocum ve ark., 2009).

Ayrılabilir ve Bütünleyici işaretler kavramı Psikoloji alanında geliştirilmiş olsa da, Şekil 3.6'da görüldüğü üzere 2000 yılında Kartograf Elisabeth Nelson tarafından desteklenmiş ve kullanılmıştır (Nelson, 2000).

Şekil 3.6'da görülen çalışmalar da Nelson'un iki değişkenli haritalar üzerinde yoğunlaştığı görülsede, örnekler incelendiğinde bu kavramın çok değişkenli haritalarda da kullanılabileceği görülmektedir (Slocum ve ark., 2009).





Şekil 3.6 Ayrılabilir (A) ve Bütünleyici (B) İşaret Örnekleri (Nelson, 2000)

4. UYGULAMA

Trafik kazaları sonucu oluşan ölümler, yaralanmalar ve maddi hasarlar Türkiye'nin en önemli sorunlarındandır. Son 5 yıla ait veriler incelendiğinde her yıl 1.000.000'dan fazla trafik kazası meydana geldiği, bunların ortalama 145.000'nin ölümlü ve yaralanmalı, yaklaşık 1.060.000'nin de maddi hasarlı kaza olduğu görülmektedir. Bu kazalarda ortalama yılda 4.000 kişi hayatını kaybetmekte, yaklaşık 250.000 kişi yaralanmaktadır. Türkiye'de trafik güvenliğinin sağlanabilmesi amacıyla alınması gereken önlemler ve yapılacak yatırımların belirlenmesi için birden çok mevcut trafik kaza verisinden yararlanarak, hangi illerdeki kaza verilerinin benzerlik taşıdığının belirlenmesi oldukça önemlidir. Bu amaçla bu tez çalışmasında Türkiye İstatistik Kurumu (TUİK) tarafından hazırlanan 2011, 2012 ve 2013 yıllarına ait il bazlı motorlu kara taşıtı sayısı, ölümlü ve yaralanmalı trafik kaza sayıları, ölü ve yaralı sayıları verileri kullanılarak (4 farklı değer) 3 farklı yöntemle kümeleme analizi yapılmış ve kümeleme analizi sonuçlarına göre çok değişkenli haritalar üretilmiştir. 3 farklı yöntemle her üç yıla ait üretilen haritalar karşılaştırılarak çok değişkenli haritalama ve kümeleme başarısı açısından hangi yöntemin daha uygun olduğu değerlendirilmiştir.

4.1. Veri Setinin Elde Edilmesi

Bu tez çalışmasında Emniyet Genel Müdürlüğü (EGM) ve Türkiye İstatistik Kurumu (TUİK) tarafından hazırlanan 2011, 2012 ve 2013 yılı Trafik Kaza İstatistikleri Karayolu verileri kullanılacaktır. Türkiye İstatistik Kurumu (TUİK) sitesinden elde edilen motorlu karata taşıtı sayısı, ölümlü yaralanmalı trafik kaza sayısı, ölü sayısı ve yaralı sayısı verilerini içeren veri setleri Ek1-a, Ek1-b ve Ek1-c'de verilmiştir.

4.2. Verilerin Hazırlanması

Kümeleme analizinde kullanılacak verilerin ilk olarak veri hazırlama sürecinden geçmesi gereklidir. Veri hazırlama sürecinde ilk olarak veriler arasında birim farklılıkları varsa birimsel farklılıkların giderilmesi gerekir. Bu tez çalışmasında kullanılan verilerin dört özniteliği de (motorlu kara taşıtı sayısı, ölümlü yaralanmalı trafik kaza sayısı, ölü sayısı, yaralı sayısı) aynı birimde olduğundan verilerin birim bütünlüğü için birimsel standardize işlemine ihtiyaç duyulmamıştır. Ancak veri seti incelendiğinde Motorlu Kara Taşıtı Sayısı özniteliğindeki veri değerlerinin diğer üç öznitelikteki veri değerlerinden çok büyük olduğu tespit edilmiştir. Bu durumun kümeleme analizi işlemini olumsuz etkilememesi için verilere değersel standardize işlemi uygulanmıştır. Bu işlemde veri

setindeki her bir öznitelik altındaki veri değerleri [-1,1] aralığına standartlaştırılmıştır. Standardize edilmiş veri setleri Ek2-a, Ek2-b ve Ek2-c'de verilmiştir.

Kümeleme analizi yapılmadan önce kontrolü yapılacak işlemlerden birisi de kullanılacak verilerin birbiriyle ilişkili (korelasyonlu) olup olmadığının test edilmesidir. Çünkü her türlü veri birbiriyle ilişkili olsun veya olmasın kümeleme algoritmalarında analiz edilerek sonuç alınabilir. Gerçekte birbiriyle ilişkili olmayan verilerin kümelenmesi ile istenmeyen sonuçlar ortaya çıkabilir. Çalışmadaki verilerin birbiriyle ilişkili olup olmadığını belirlemek için verilerin birbirleriyle korelasyon katsayıları (4.1) numaralı eşitliğe göre hesaplanmıştır.

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (4.1)$$

Eşitlikte r iki değişken arasındaki korelasyon katsayısını, X_i ve Y_i değişkenleri, $\bar{X}$ ve $\bar{Y}$ değişkenlerin ortalama değerlerini göstermektedir. Hesaplamalar sonucu 2011, 2012 ve 2013 yılları için aşağıdaki değerler bulunmuştur:

2011 yılına ait veri seti üzerinde yapılan uygulamada;

Motorlu taşıt sayısı-Ölümlü yaralanmalı kaza sayısı korelasyonu r_1=0,96

Motorlu taşıt sayısı-Ölü sayısı korelasyonu r_2=0,83

Motorlu taşıt sayısı-Yaralı sayısı korelasyonu r_3=0,94

2012 yılına ait veri seti üzerinde yapılan uygulamada;

Motorlu taşıt sayısı-Ölümlü yaralanmalı kaza sayısı korelasyonu r_1=0,95

Motorlu taşıt sayısı-Ölü sayısı korelasyonu r_2=0,88

Motorlu taşıt sayısı-Yaralı sayısı korelasyonu r_3=0,93

2013 yılına ait veri seti üzerinde yapılan uygulamada;

Motorlu taşıt sayısı-Ölümlü yaralanmalı kaza sayısı korelasyonu r_1=0,94

Motorlu taşıt sayısı-Ölü sayısı korelasyonu r_2=0,89

Motorlu taşıt sayısı-Yaralı sayısı korelasyonu r_3=0,93

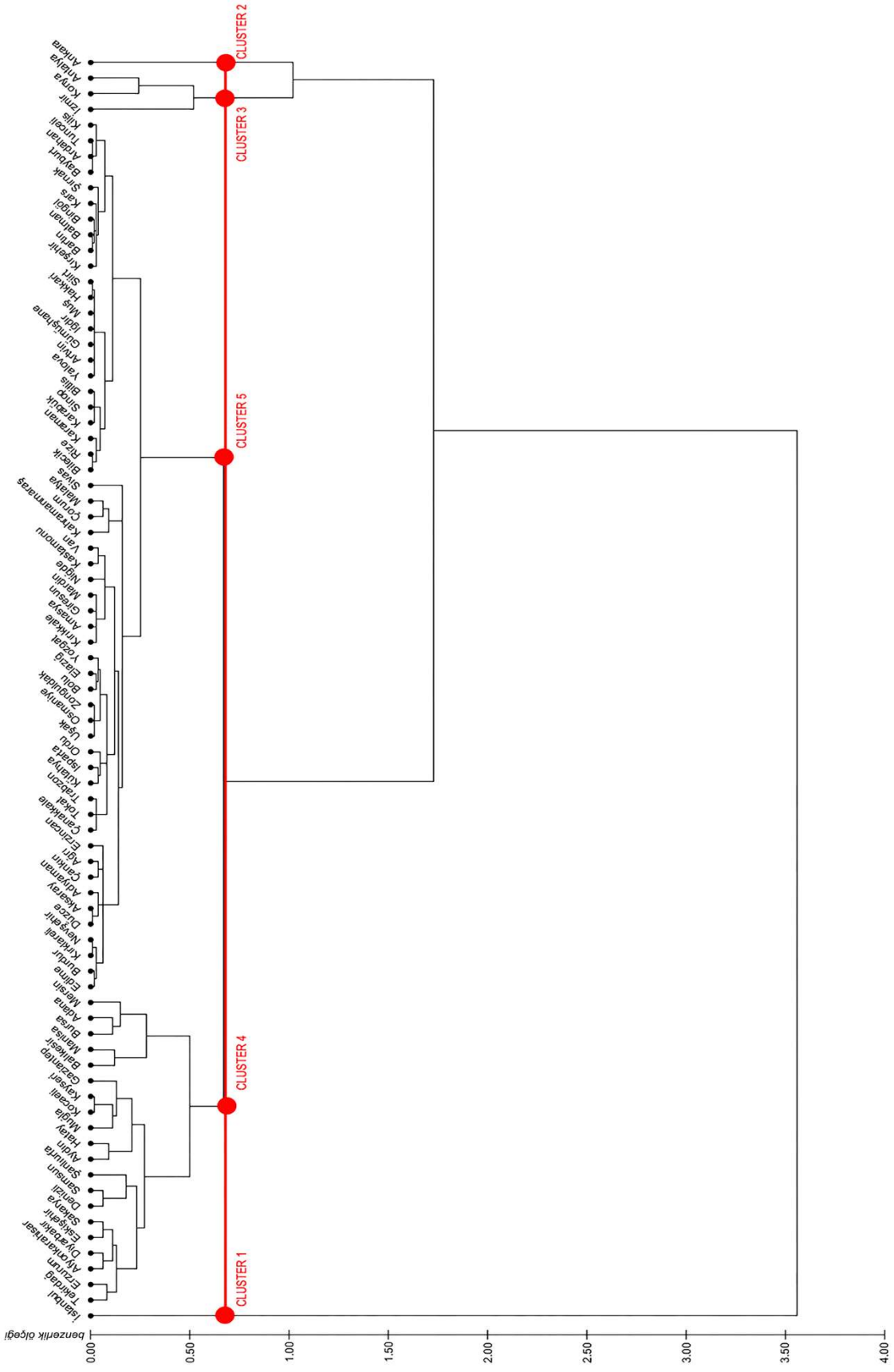
Romesburg (1984) kümeleme çalışmalarında kullanılacak verilerin ilişkili olması için korelasyon katsayısının 0,80 ve üzeri olması gerektiğini belirtmiştir. Bu görüş dikkate alındığında çalışmada kullanılan veri setindeki değişkenlerin birbiriyle ilişkili olduğu görülmektedir. Bu şekilde kümeleme analizi öncesi yapılması gereken iki aşama tamamlanmış ve kümeleme aşamasına geçilmiştir.

4.3. Veri Setlerinin Kümelenmesi

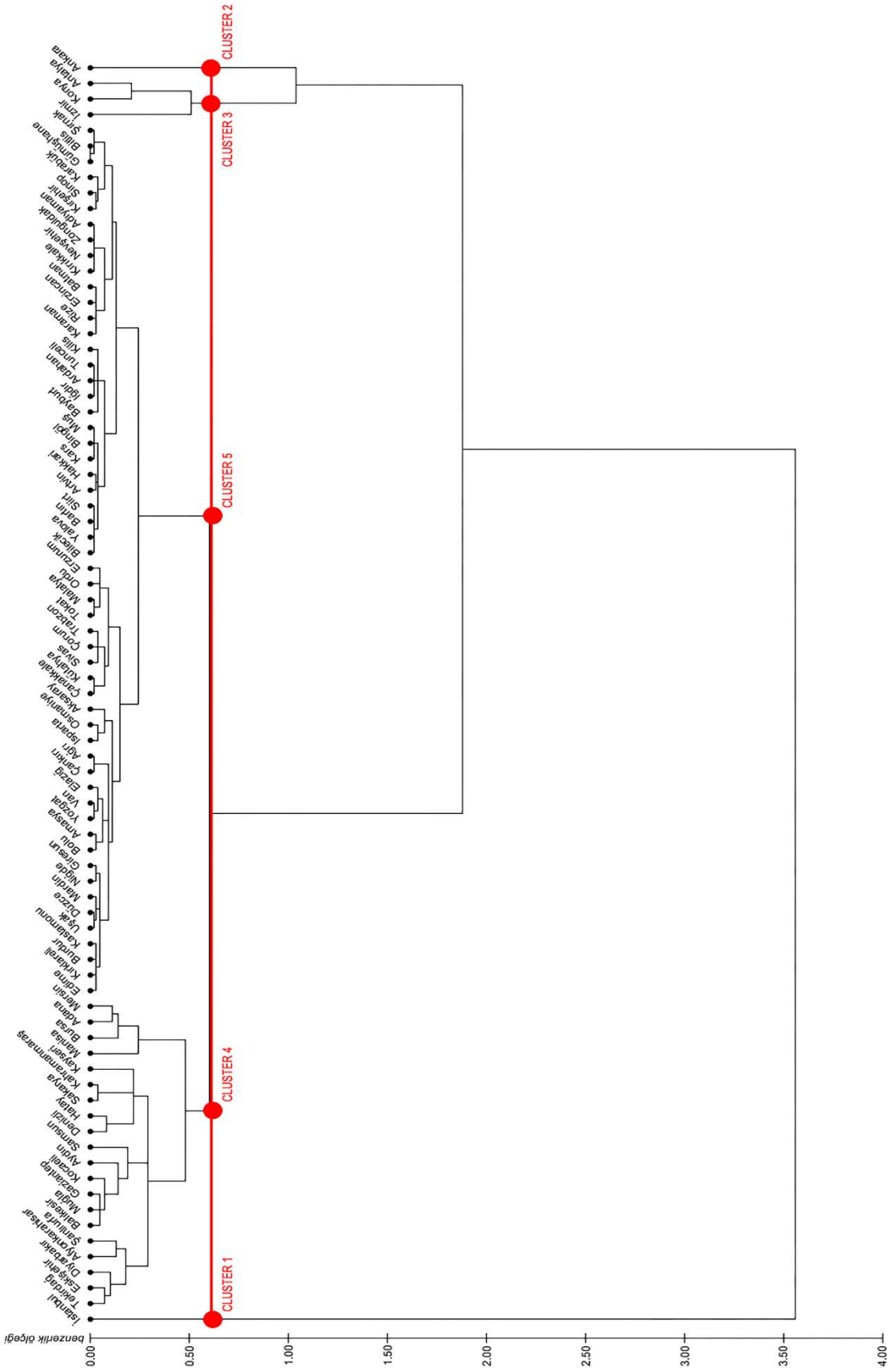
4.3.1. Birleştirici Hiyerarşik Kümeleme Yöntemiyle (AGNES) Veri Setinin Kümelenmesi

AGNES yönteminin uygulanmasında kümeyi oluşturacak i ve j elemanlarının birbirlerine olan benzerliklerinin belirlenmesinde birbirlerine olan öklid mesafeleri kullanılmıştır. Kümeyi oluşturacak iki elemanın birbirlerine olan öklid mesafesi (2.1) eşitliği ile hesaplanmıştır.

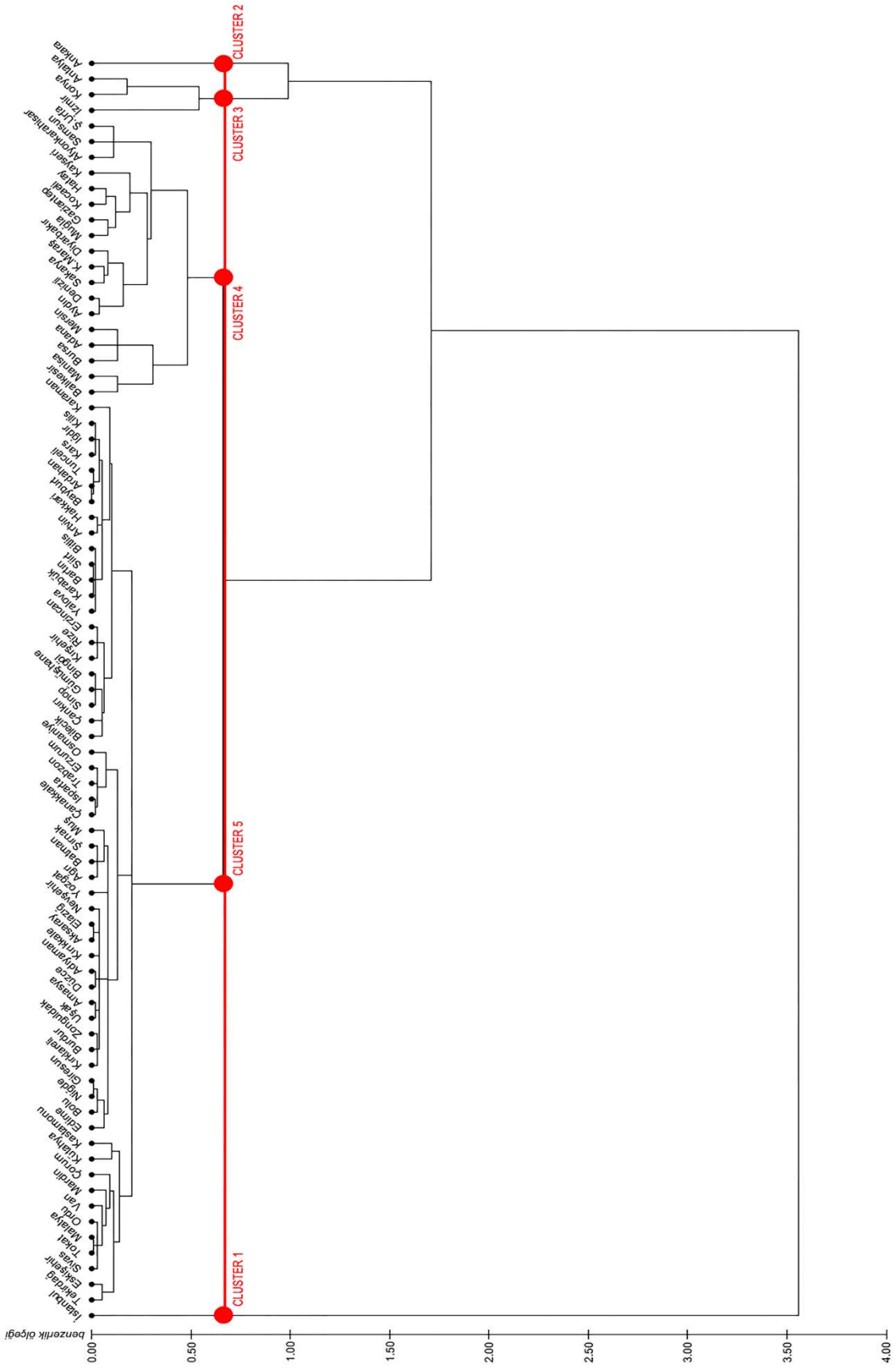
AGNES metodunda oluşturulacak en uygun küme sayısının belirlenmesinde dendrogram adı verilen ağaç yapıdan yararlanılmaktadır. Dendrogram üzerinde gerçekleşen büyük sıçramalar bize oluşturulacak kümeleri göstermektedir. Ancak kümelemede amaç gerçekten homojen ve birbirlerinden farklı gruplar oluşturmak olduğundan dendrogramın yardımcı bir eksen ile incelenmesi ve homojen grupların belirlenmesi gerekmektedir. AGNES Hiyerarşik Kümeleme işleminde SPSS ve MultiDendrograms 4.1 yazılımlarından yararlanılmıştır. Kümeleme işleminin gerçekleşmesinde kullanılan benzeşmezlik matrisi SPSS yazılımıyla, küme sayısının ve küme elemanlarının belirlenmesinde kullanılacak Dendrogram çizelgesi de MultiDendrograms 4.1 yazılımı yardımıyla elde edilmiştir. Yardımcı eksen ile dendrogram üzerinde ağaç yapının kesiştiği her bir düğüm bir kümeyi ifade etmektedir. Bu şekilde 2011, 2012 ve 2013 verileri için 5 küme belirlenmiştir (Şekil 4.1, Şekil 4.2 ve Şekil 4.3).



Şekil 4.1 2011 Yılı Trafik Kaza Verilerinin Dendrogram Üzerinde Gösterilmesi



Şekil 4.2 2012 Yılı Kaza Verilerinin Dendrogram Üzerinde Gösterilmesi



Şekil 4.3 2013 Yılı Kaza Verilerinin Dendrogram Üzerinde Gösterilmesi

Kümeleme işlemi sonucu oluşturulan kümelerin sonucunu yorumlanabilmesi amacıyla her bir kümenin her öznitelik verisi için ortalama z skoru hesaplanmıştır. Z-skoru yani Standart skor;

$$\boldsymbol{\mu} = \frac{\sum \boldsymbol{x}}{\boldsymbol{n}} \tag{4.2}$$

$$\boldsymbol{\sigma} = \sqrt{\frac{\sum(\boldsymbol{x}-\boldsymbol{\mu})^2}{\boldsymbol{n}}} \tag{4.3}$$

$$\boldsymbol{z\ skoru} = \frac{(\boldsymbol{x}-\boldsymbol{\mu})}{\boldsymbol{\sigma}} \tag{4.4}$$

formülleriyle hesaplanmıştır. Burada x oluşturulan her kümenin her bir öznitelik için aritmetik ortalama, **μ** kullanılan veri setindeki her bir öznitelik için aritmetik ortalama ve $\boldsymbol{\sigma}$ kullanılan veri setindeki her öznitelik için standart sapmadır. Elde edilen kümelere ait ortalama z skoru Çizelge 4.1, Çizelge 4.2 ve Çizelge 4.3'de verilmiştir.

Çizelge 4.1 2011 Yılı Verileri İçin Her Kümenin Ortalama z Skoru Tablosu

KÜME NO	KÜME ELEMAN SAYISI	MOTORLU KARA TAŞITI SAYISI	ÖLÜMLÜ YARALANMALI TRAFİK KAZA SAYILARI	ÖLÜ SAYILARI	YARALI SAYILARI
	ORTALAMA z SKORU				
1	1	7,22	5,61	4,26	5,36
2	1	3,09	3,97	2,78	4,16
3	3	1,49	2,10	2,39	2,04
4	20	0,24	0,42	0,78	0,47
5	56	-0,35	-0,43	-0,53	-0,45

Çizelge 4.2 2012 Yılı Verileri İçin Her Kümenin Ortalama z Skoru Tablosu

KÜME NO	KÜME ELEMAN SAYISI	MOTORLU KARA TAŞITI SAYISI	ÖLÜMLÜ YARALANMALI TRAFİK KAZA SAYILARI	ÖLÜ SAYILARI	YARALI SAYILARI
	ORTALAMA z SKORU				
1	1	7,20	5,28	4,62	5,04
2	1	3,09	3,95	3,65	4,18
3	3	1,50	2,29	2,62	2,22
4	20	0,25	0,46	0,47	0,50
5	56	-0,36	-0,45	-0,45	-0,46

Çizelge 4.3 2013 Yılı Verileri İçin Her Kümenin Ortalama z Skoru Tablosu

KÜME NO	KÜME ELEMAN SAYISI	MOTORLU KARA TAŞITI SAYISI	ÖLÜMLÜ YARALANMALI TRAFİK KAZA SAYILARI	ÖLÜ SAYILARI	YARALI SAYILARI
	ORTALAMA z SKORU				
1	1	7,22	5,20	5,19	4,94
2	1	3,09	3,89	2,77	4,13
3	3	1,48	2,26	2,12	2,22
4	18	0,29	0,54	0,68	0,59
5	58	-0,34	-0,44	-0,46	-0,45

Kümeleme analizinin sonuçlarını yorumlarken dikkatli olunmalıdır çünkü çoklu öznitelikler için herhangi bir veri setinin hatta bir sayı dizesinin kümelenmesi mümkündür. Bu bakımdan uygulanmadan elde edilen kümelemelerin uygunluğunun belirlenmesi gereklidir. AGNES uygulamasından elde edilen kümelerin uygunluğunun belirlenmesinde kullanılan temel yaklaşımlardan biri Cophenetic korelasyon katsayısının belirlenmesidir. Cophenetic korelasyon katsayısı, X veri kümesi için hesaplanan benzerlik matrisi $P=\{p_{ij}\}$ ile hiyerarşik kümeleme yöntemine göre elde edilen ağaç diyagramında veri gözlem çiftlerinin ilk defa aynı kümede gruplandığı yakınlık seviyeleri q_{ij} değerlerinden oluşan Cophenetic matrisi $Q=\{q_{ij}\}$ arasındaki benzerliğin bir ölçüsüdür. $\mu_p$ve μ_q sırasıyla $P=\{p_{ij}\}$ve $Q=\{q_{ij}\}$'nun ortalamaları,

$$\mu_P = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_{ij} \tag{4.5}$$

$$\mu_Q = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} q_{ij} \tag{4.6}$$

$$M = \frac{n(n-1)}{7} \tag{4.7}$$

şeklinde ifade edilir. Cophenetic korelasyon katsayısı, *CCC* kısaltmasıyla gösterilmektedir,

$$\mathbf{CCC} = \frac{\frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=1}^{n} p_{ij}q_{ij} - \mu_P\mu_Q}{\sqrt{\left(\frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=1}^{n} p_{ij}^2 - \mu_P^2\right)}\sqrt{\left(\frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=1}^{n} q_{ij}^2 - \mu_Q^2\right)}} \quad (4.8)$$

formülü ile hesaplanır. Burada *CCC*, [-1,1] arasında değer alır. *CCC* değerinin 1 olması verideki hiyerarşi yapıdan elde edilen $P=\{p_{ij}\}$ve $Q=\{q_{ij}\}$arasında yüksek derecede benzerlik olduğunu gösterir.

2011, 2012 ve 2013 yıllarına ait verilerden AGNES uygulamasıyla elde edilen kümelerin uygunluğunun belirleyecek Cophenetic Korelasyon Katsayısı (CCC) MultiDendrograms 4.1 yazılımı ile hesaplanmıştır. 2011,2012 ve 2013 yıllarına ait verilerde bulunan Cophenetic Korelasyon Katsayısı sırasıyla 0.939683, 0.943563 ve 0.937549 olarak hesaplanmıştır. Bu değerler kümeleme sonuçlarımızın kullanılabilir olduğunu gösterir, çünkü Romesburg (1984) Cophenetic değerin 0.80 yada daha büyük olduğunda kullanılan değişkenlerin birbirleriyle ilişkili ve kabul edilebilir olduğunu belirtmiştir.

4.3.2. K-Ortalama Yöntemiyle Veri Setinin Kümelenmesi

K-Ortalama yönteminin uygulanmasında RapidMiner yazılımı kullanılmıştır. K-Ortalama yönteminde küme sayısı kullanıcı tarafından belirlenir. Ancak veri madenciliğinde küme sayısı önemlidir. K-Ortalama algoritması gibi algoritmalar başlangıç küme sayısının kullanıcının girmesini ister. Bu durumda, kullanıcı ya deneme yanılmalarla en optimum küme sayısını belirleyecek yada her kümeleme sonrası bir takım testler yapıp, örneğin 2,3 yada 4 ayrı küme oluşturmanın hangisinin daha verimli sonuç verdiğini hesaplayacaktır. Bu uygulamada küme sayısının belirlenmesinde *Dunn Geçerlilik indeksi* ve testlerinden yararlanılmıştır.

Bu tez çalışmasında K-Ortalama yönteminin uygulanmasında kullanılacak *k* sayısını belirleme adına *k=2, k=3, k=4…, k=7* senaryoları izlenmiştir. Her senaryo için Dunn ve Davies Bouldin indeksleri hesaplanmıştır. 2011, 2012 ve 2013 veri setleri için hesaplanan Dunn ve Davies Bouldin indeksleri Çizelge 4.4'de gösterilmiştir.

Çizelge 4.4 K-Ortalama Algoritması İçin k Katsayısının Belirlenmesi

	2011 Yılı Veri Seti		**2012 Yılı Veri Seti**		**2013 Yılı Veri Seti**	
	Dunn	Davies-Bouldin	Dunn	Davies-Bouldin	Dunn	Davies-Bouldin
k=2	0,027	0,764	0,245	0,543	0,131	0,601
k=3	0,035	0,657	0,087	0,660	0,059	0,632
k=4	0,040	0,642	0,088	0,589	0,089	0,520
k=5	0,034	0,695	0,028	0,659	0,106	0,558
k=6	0,024	0,751	0,036	0,694	0,050	0,635
k=7	0,024	0,721	0,036	0,623	0,043	0,634

K-Ortalama yönteminin uygulanmasında 2011 verileri için küme sayısı olarak k=4, 2012 verileri için k=4 ve 2013 yılı verileri için k=5 seçilmiştir. K-Ortalama algoritmasının bir kez çalışması esnasında en fazla yapılacak iterasyon sayısı 100, algoritmanın maksimum dönüş sayısı 50 olarak alınmıştır. 2011, 2012 ve 2013 verileri için yöntem ayrı ayrı uygulanmıştır. Kümeleme işlemi sonucu oluşturulan kümelere ait ortalama z skoru tabloları Çizelge 4.5, Çizelge 4.6 ve Çizelge 4.7'de verilmiştir.

Çizelge 4.5 2011 Yılı Verileri için K-Ortalama Yöntemiyle Oluşturulan Kümelerin Ortalama z-Skoru Tablosu

KÜME NO	**KÜME ELEMAN SAYISI**	**MOTORLU KARA TAŞITI SAYISI**	**ÖLÜMLÜ YARALANMALI TRAFİK KAZA SAYILARI**	**ÖLÜ SAYILARI**	**YARALI SAYILARI**
	ORTALAMA z SKORU				
1	2	5,15	4,79	3,52	4,76
2	7	1,07	1,51	1,80	1,50
3	18	0,08	0,22	0,56	0,29
4	54	-0,36	-0,45	-0,55	-0,47

Çizelge 4.6 2012 Yılı Verileri İçin K-Ortalama Yöntemiyle Oluşturulan Kümelerin Ortalama z-Skoru Tablosu

KÜME NO	KÜME ELEMAN SAYISI	MOTORLU KARA TAŞITI SAYISI	ÖLÜMLÜ YARALANMALI TRAFİK KAZA SAYILARI	ÖLÜ SAYILARI	YARALI SAYILARI
	ORTALAMA z SKORU				
1	2	5,15	4,62	4,14	4,61
2	3	1,50	2,29	2,62	2,22
3	17	0,31	0,53	0,57	0,57
4	59	-0,34	-0,42	-0,44	-0,43

Çizelge 4.7 2013 Yılı Verileri İçin K-Ortalama Yöntemiyle Oluşturulan Kümelerin Ortalama z-Skoru Tablosu

KÜME NO	KÜME ELEMAN SAYISI	MOTORLU KARA TAŞITI SAYISI	ÖLÜMLÜ YARALANMALI TRAFİK KAZA SAYILARI	ÖLÜ SAYILARI	YARALI SAYILARI
	ORTALAMA z SKORU				
1	1	7,22	5,20	5,19	4,94
2	2	2,60	3,46	2,45	3,43
3	6	0,89	1,40	1,52	1,42
4	17	0,12	0,28	0,47	0,33
5	55	-0,36	-0,46	-0,49	-0,47

4.3.3. K-Medoids Yöntemiyle Veri Setinin Kümelenmesi

K-Medoids yönteminin uygulanmasında da RapidMiner yazılımından yararlanılmıştır. K-ortalamalar algoritmasından farklı olarak k-means işlem operatörü yerine k-medoids işlem operatörü kullanılmıştır. K-Medoids yönteminin uygulanmasında kullanılacak *k* sayısını belirleme adına *k=2, k=3, k=4…, k=7* senaryoları izlenmiştir. Her senaryo için Dunn ve Davies Bouldin indeksleri hesaplanmıştır. 2011, 2012 ve 2013 veri setleri için hesaplanan Dunn ve Davies Bouldin indeksleri Çizelge 4.8'de gösterilmiştir.

Çizelge 4.8 K-Medoids Algoritması İçin k Katsayısının Belirlenmesi

	2011 Yılı Veri Seti		**2012 Yılı Veri Seti**		**2013 Yılı Veri Seti**	
	Dunn	Davies-Bouldin	Dunn	Davies-Bouldin	Dunn	Davies-Bouldin
k=2	0,005	1,064	0,008	1,250	0,005	1,272
k=3	0,009	0,756	0,017	0,769	0,003	1,117
k=4	0,009	0,746	0,025	0,775	0,002	0,961
k=5	0,023	0,776	0,008	0,807	0,002	1,047
k=6	0,017	0,875	0,008	0,943	0,004	0,904
k=7	0,006	0,911	0,008	0,841	0,004	0,971

K-Medoids yönteminin uygulanmasında 2011 verileri için küme sayısı olarak k=5, 2012 verileri için k=4 ve 2013 yılı verileri için k=6 seçilmiştir. K-Medoids algoritmasının bir kez çalışması esnasında en fazla yapılacak iterasyon sayısı 100, algoritmanın maksimum dönüş sayısı 50 olarak alınmıştır. 2011, 2012 ve 2013 verileri için yöntem ayrı ayrı uygulanmıştır. Kümeleme işlemi sonucu oluşturulan kümelere ait ortalama z skoru tabloları Çizelge 4.9, Çizelge 4.10 ve Çizelge 4.11'de verilmiştir.

Çizelge 4.9 2011 Yılı Verileri İçin K-Medoids Yöntemiyle Oluşturulan Kümelerin Ortalama z-Skoru Tablosu

KÜME NO	**KÜME ELEMAN SAYISI**	**MOTORLU KARA TAŞITI SAYISI**	**ÖLÜMLÜ YARALANMALI TRAFİK KAZA SAYILARI**	**ÖLÜ SAYILARI**	**YARALI SAYILARI**
	ORTALAMA z SKORU				
1	3	4,16	4,13	3,06	4,02
2	9	0,65	0,97	1,56	1,02
3	14	0,08	0,22	0,46	0,28
4	31	-0,28	-0,33	-0,33	-0,32
5	24	-0,44	-0,58	-0,82	-0,64

Çizelge 4.10 2012 Yılı Verileri İçin K-Medoids Yöntemiyle Oluşturulan Kümelerin Ortalama z-Skoru Tablosu

KÜME NO	KÜME ELEMAN SAYISI	MOTORLU KARA TAŞITI SAYISI	ÖLÜMLÜ YARALANMALI TRAFİK KAZA SAYILARI	ÖLÜ SAYILARI	YARALI SAYILARI
	ORTALAMA z SKORU				
1	3	4,15	4,07	3,61	3,98
2	14	0,54	0,88	0,95	0,91
3	24	-0,16	-0,15	-0,02	-0,10
4	40	-0,40	-0,53	-0,59	-0,56

Çizelge 4.11 2013 Yılı Verileri İçin K-Medoids Yöntemiyle Oluşturulan Kümelerin Ortalama z-Skoru Tablosu

KÜME NO	KÜME ELEMAN SAYISI	MOTORLU KARA TAŞITI SAYISI	ÖLÜMLÜ YARALANMALI TRAFİK KAZA SAYILARI	ÖLÜ SAYILARI	YARALI SAYILARI
	ORTALAMA z SKORU				
1	3	4,14	4,04	3,36	3,93
2	12	0,61	1,01	1,10	1,05
3	5	0,12	0,21	0,66	0,27
4	20	-0,20	-0,19	-0,06	-0,14
5	21	-0,34	-0,44	-0,41	-0,46
6	20	-0,45	-0,61	-0,84	-0,67

4.4. Kümeleme Analizi Sonuçlarının Haritalarla Gösterimi

Bu tez çalışmasında Türkiye'nin 81 iline ait öznitelik verileri kullanılmış ve kümeleme analizi sonuçları elde edilmiştir. Bu sonuçların harita üzerinde gösterimi de 81 ilin tamamının yer aldığı Türkiye haritası üzerinde gerçekleştirilmiştir. Ülkemizin coğrafi konumu incelendiğinde 26°-45° doğu boylamları ve 36°-42° kuzey enlemleri arasında yer aldığını görülmektedir. Türkiye gibi doğu-batı yönünde genişleyen, yani boylam farkının büyük olduğu alanlarda J.H.Lambert tarafından geliştirilen Lambert Konform Konik projeksiyonun kullanımı daha uygundur. Bu nedenle tez kapsamında hazırlanan haritalarda Lambert Konform Konik projeksiyonu ve aşağıdaki projeksiyon parametreleri kullanılmıştır;

Projeksiyon: Lambert Konform Konik

Standart Paralel 1: 40° 40'

Standart Paralel 2: 43° 20'

Orta Meridyen: 34°

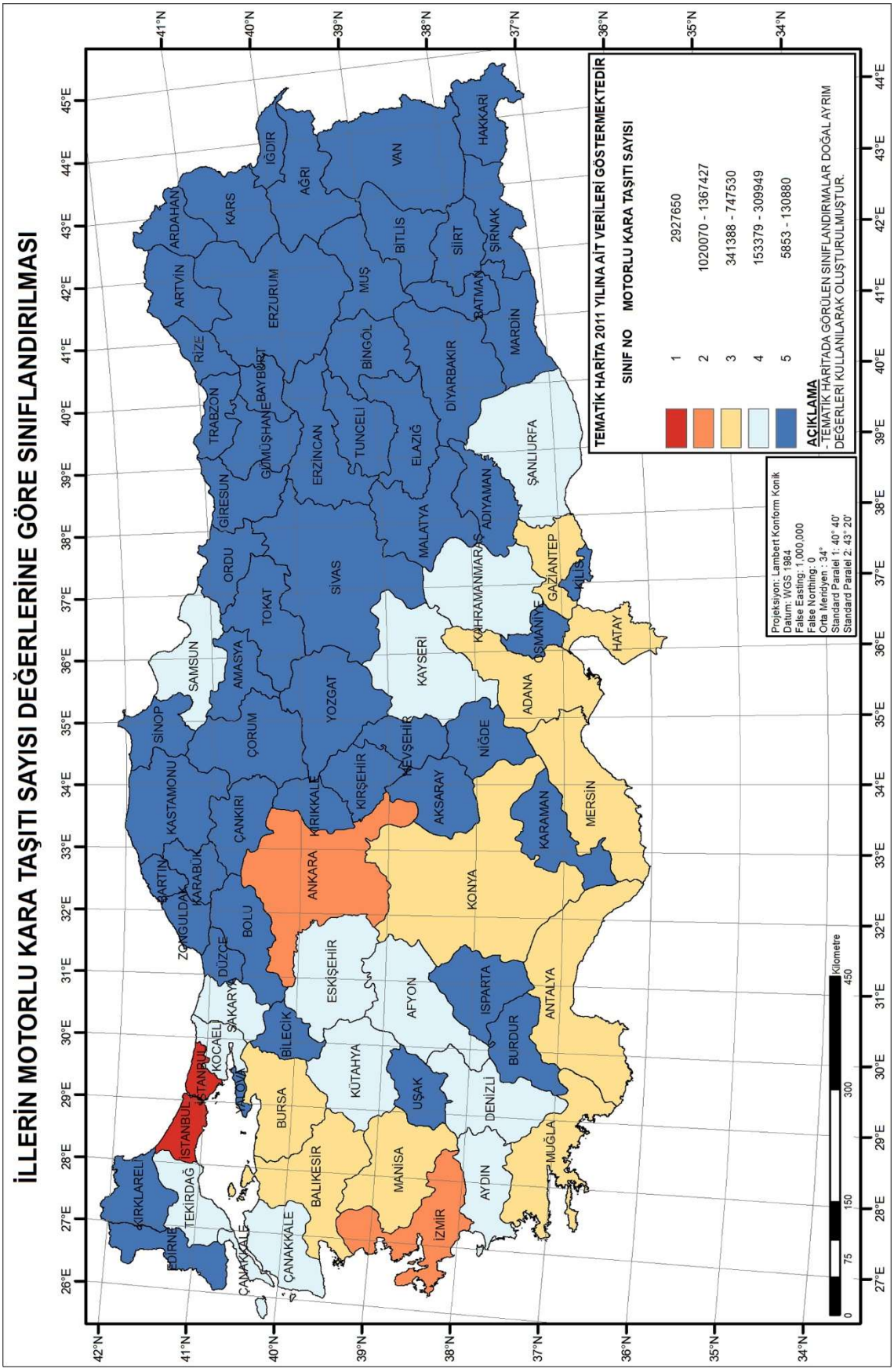
False Easting: 1000 000 m

False Northing: 0

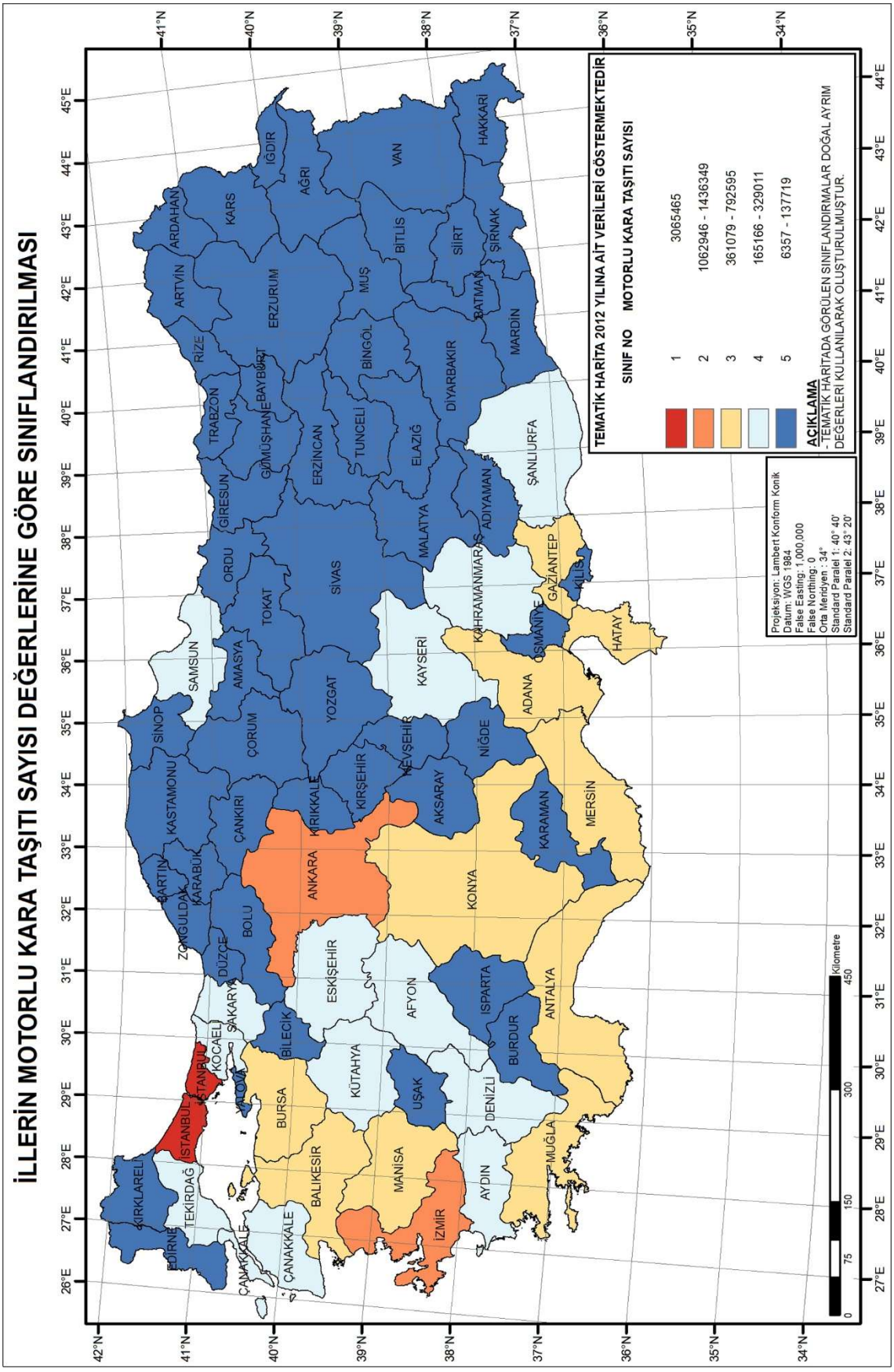
Haritalarda kullanılacak renk seçiminde ise; kümeleme analizi sonucunda oluşan kümelerin her birinin farklı bir renk ile gösterimi tercih edilmiştir. Renk seçiminde kartografik tasarımlar için renk tavsiyesinde bulunan colorbrewer 2.0 yazılımından yararlanılmıştır (URL7).

Kümeleme analizi sonuçlarıyla üretilen çok değişkenli haritaların yorumlamada yardımcı olması amacıyla öncelikle Türkiye illerinin motorlu kara taşıtı sayısı, ölümlü ve yaralanmalı trafik kaza sayısı, ölü sayısı ve yaralı sayısı değerlerine göre ArcGIS yazılımıyla 2011, 2012 ve 2013 yılları için sınıflandırma yöntemi ile tek değişkenli haritalar üretilmiştir (Şekil 4.4 - Şekil 4.15).

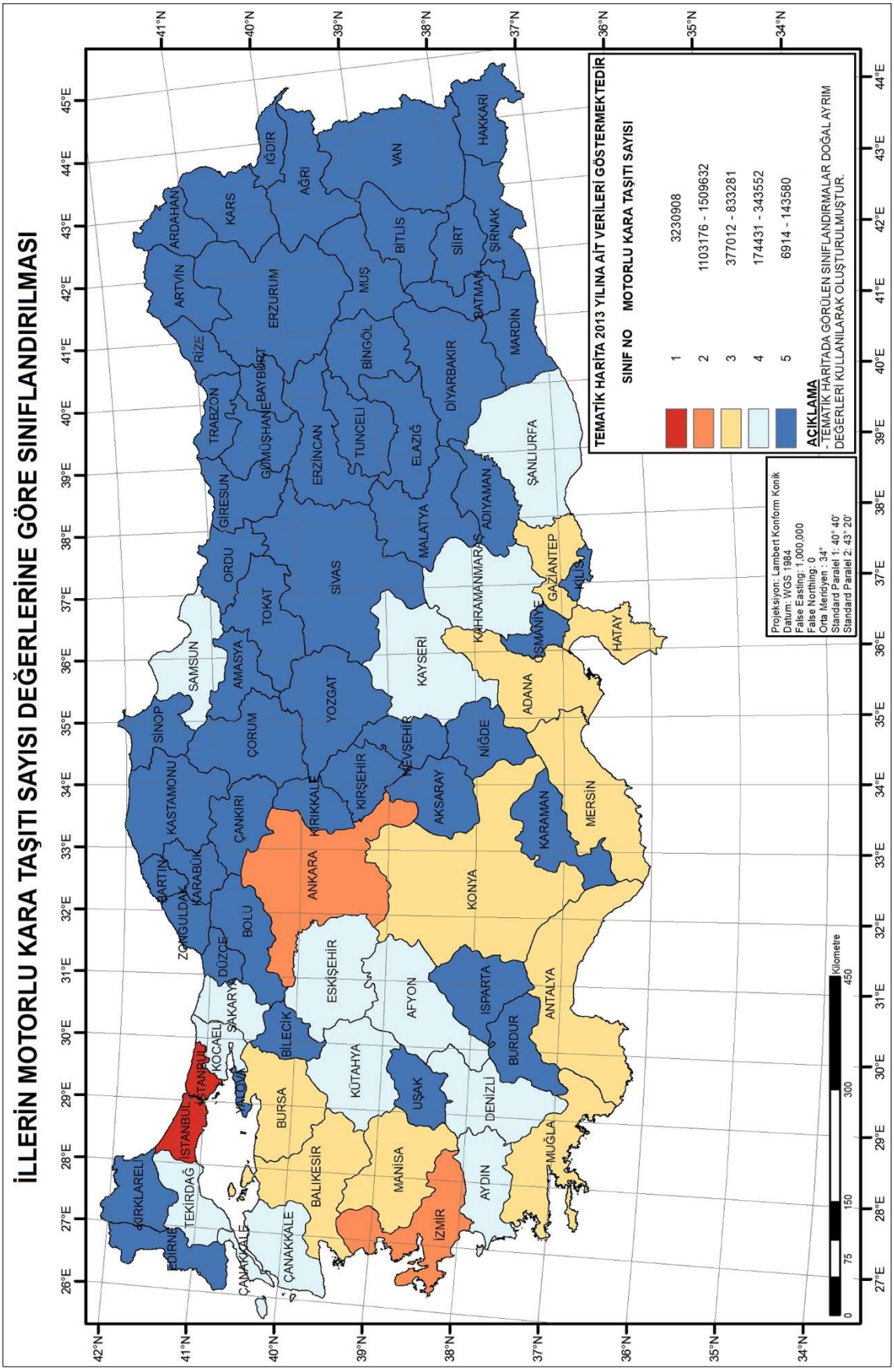
AGNES, K-Ortalama ve K-Medoids kümeleme analiz yöntemlerinin uygulanmasıyla veri setleri 4 farklı değer (motorlu kara taşıtı sayısı, ölümlü ve yaralanmalı trafik kaza sayısı, ölü sayısı ve yaralı sayısı) kullanarak kümelenmiş ve olduğu ArcGIS yazılımıyla 2011, 2012 ve 2013 yılları için çok değişkenli haritalar üretilmiştir (Şekil 4.16 - Şekil 4.24).



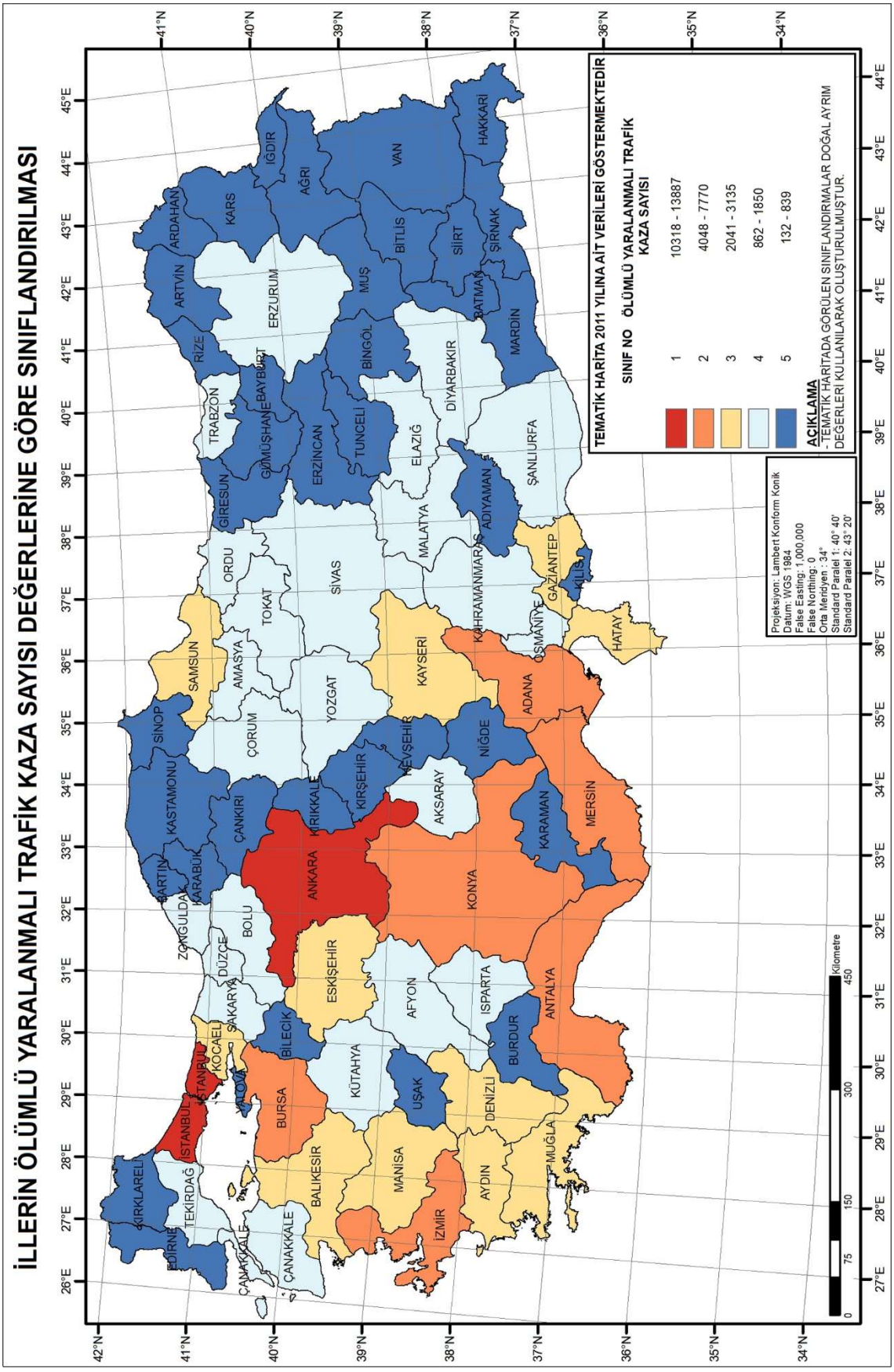
Şekil 4.4 2011 Yılı Motorlu Kara Taşıtı Sayısına Göre Üretilen Tek Değişkenli Harita



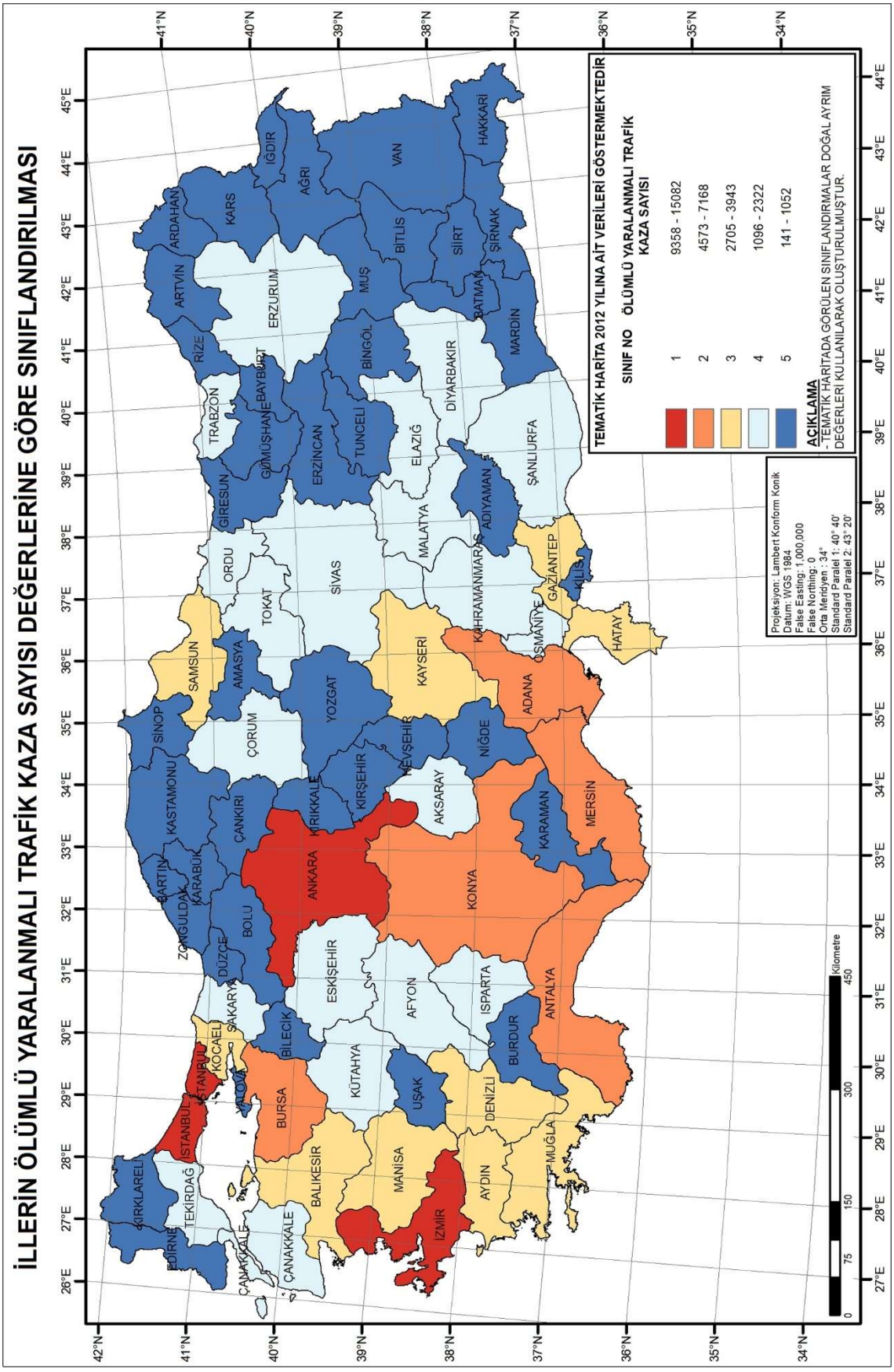
Şekil 4.5 2012 Yılı Motorlu Kara Taşıtı Sayısına Göre Üretilen Tek Değişkenli Harita



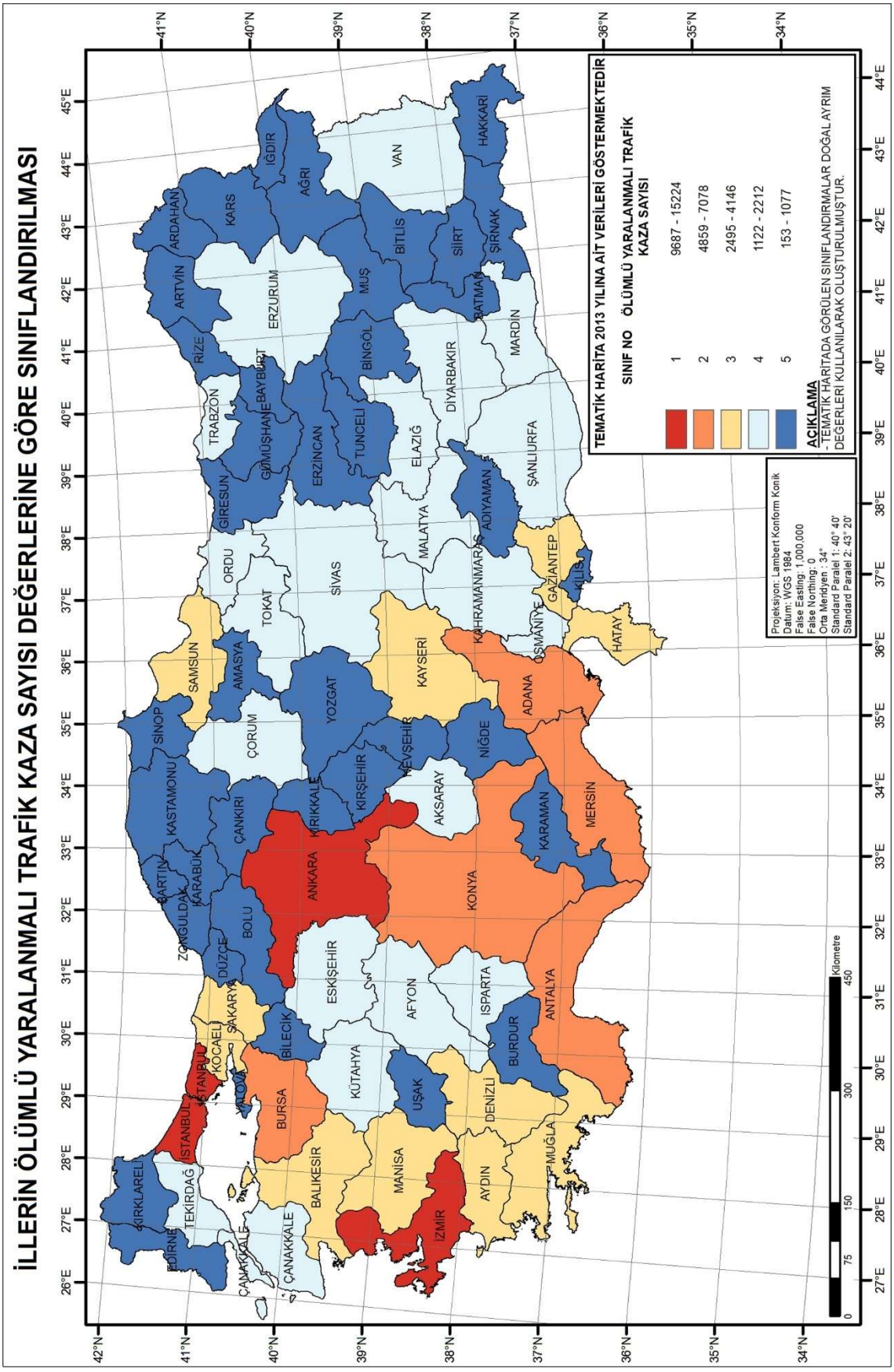
Şekil 4.6 2013 Yılı Motorlu Kara Taşıtı Sayısına Göre Üretilen Tek Değişkenli Harita



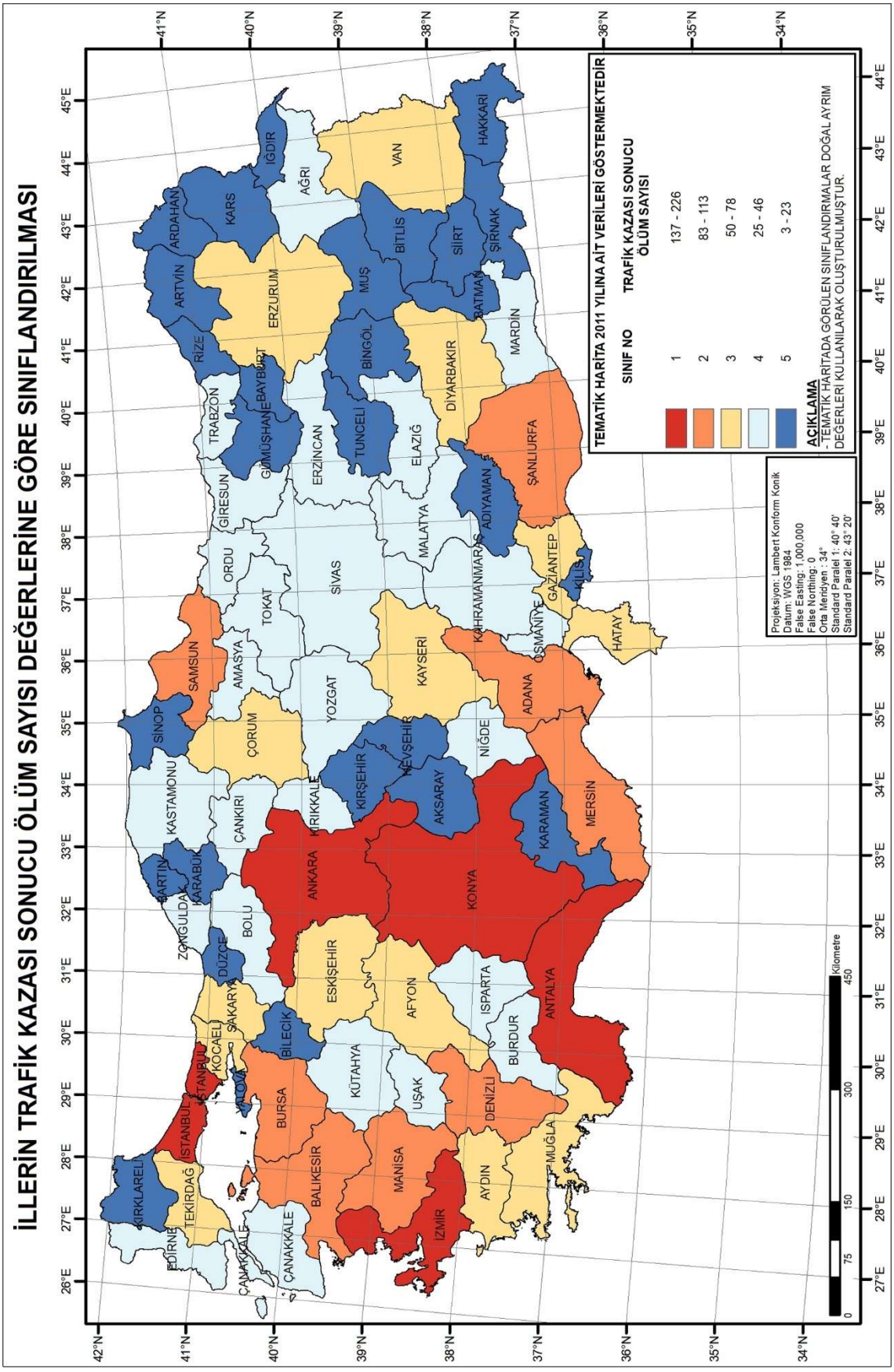
Şekil 4.7 2011 Yılı Ölümlü Yaralanmalı Trafik Kaza Sayısına Göre Üretilen Tek Değişkenli Harita



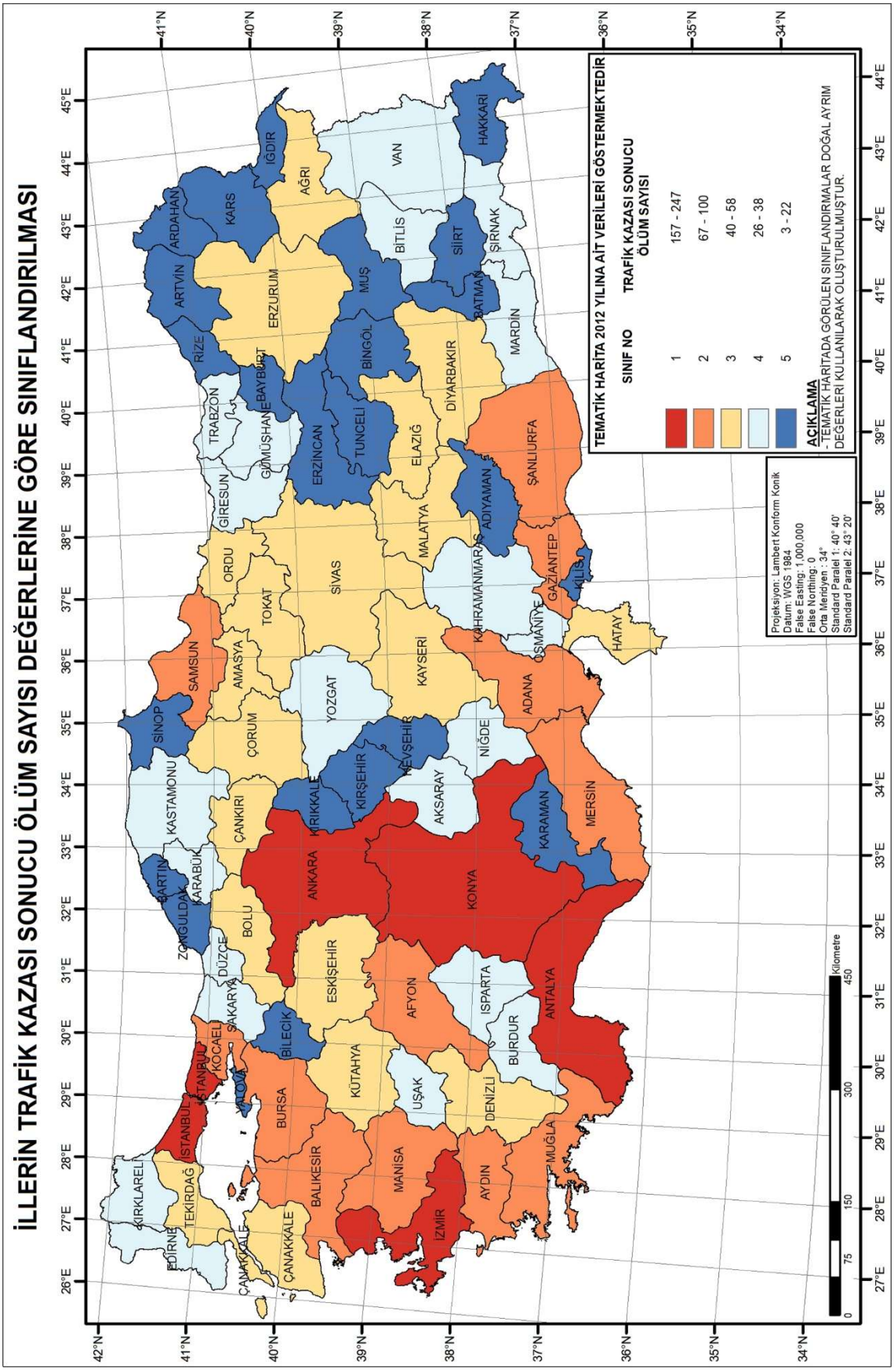
Şekil 4.8 2012 Yılı Ölümlü Yaralanmalı Trafik Kaza Sayısına Göre Üretilen Tek Değişkenli Harita



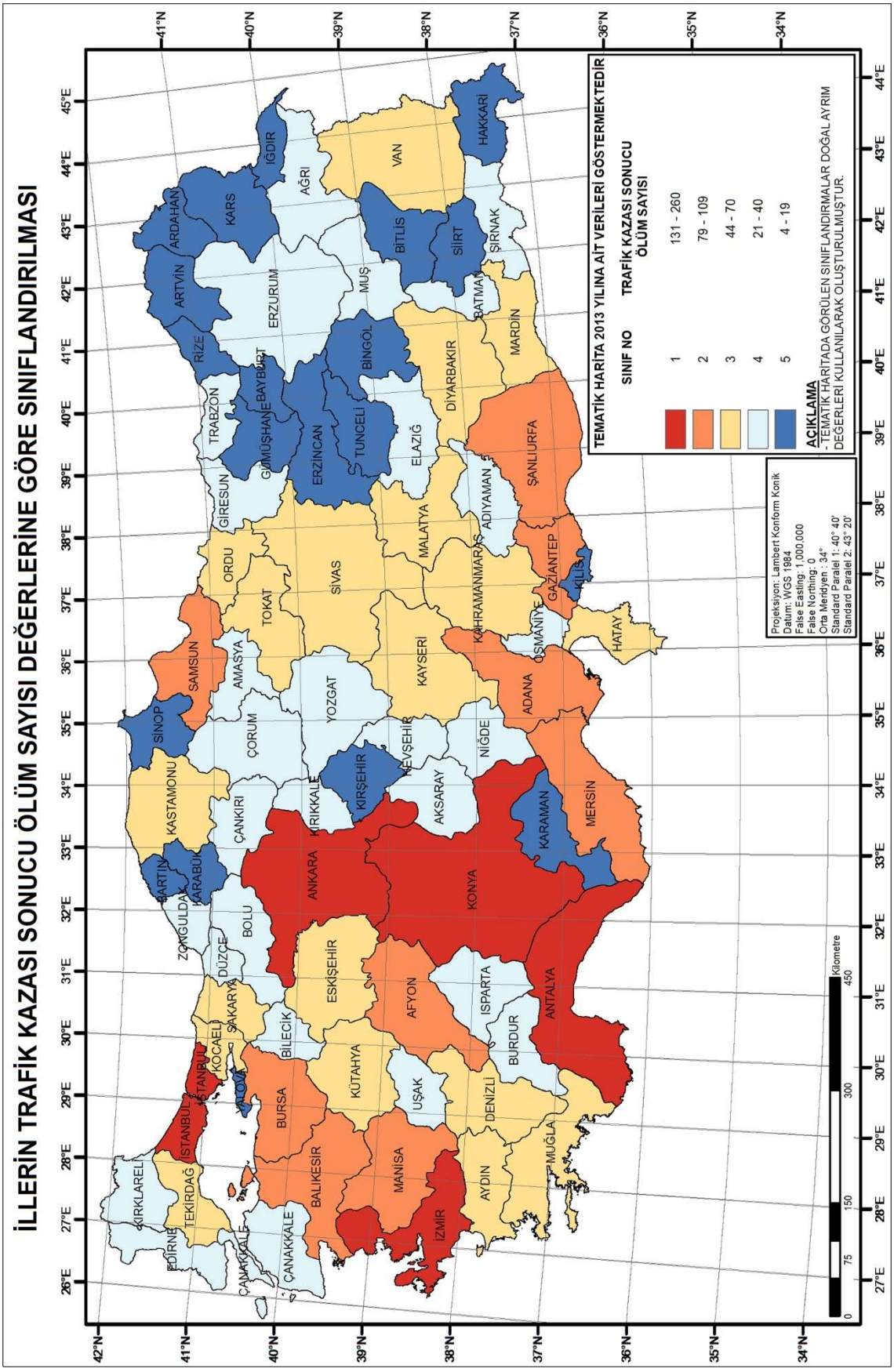
Şekil 4.9 2013 Yılı Ölümlü Yaralanmalı Trafik Kaza Sayısına Göre Üretilen Tek Değişkenli Harita



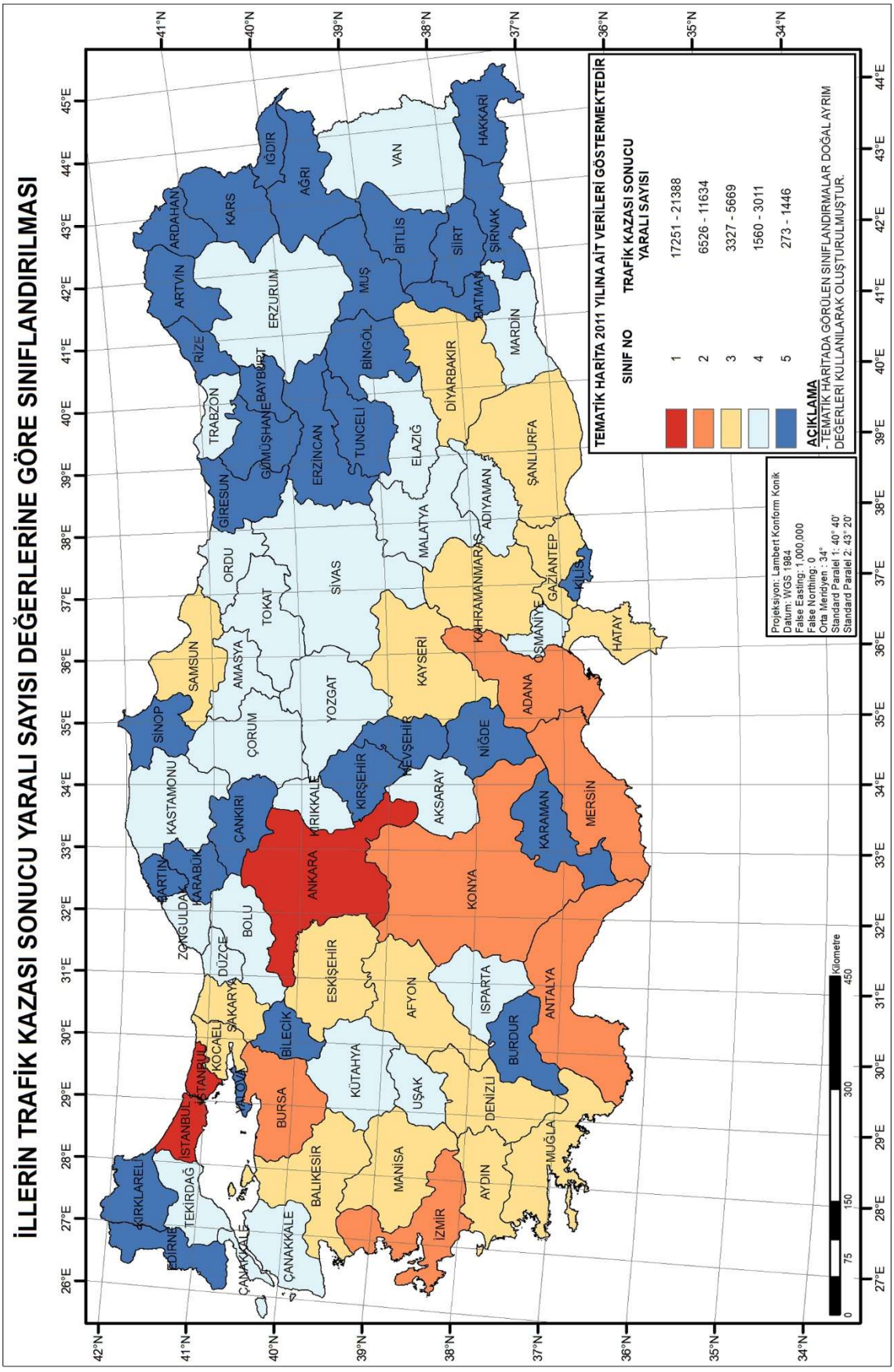
Şekil 4.10 2011 Yılı Trafik Kazası Sonucu Ölüm Sayısına Göre Üretilen Tek Değişkenli Harita



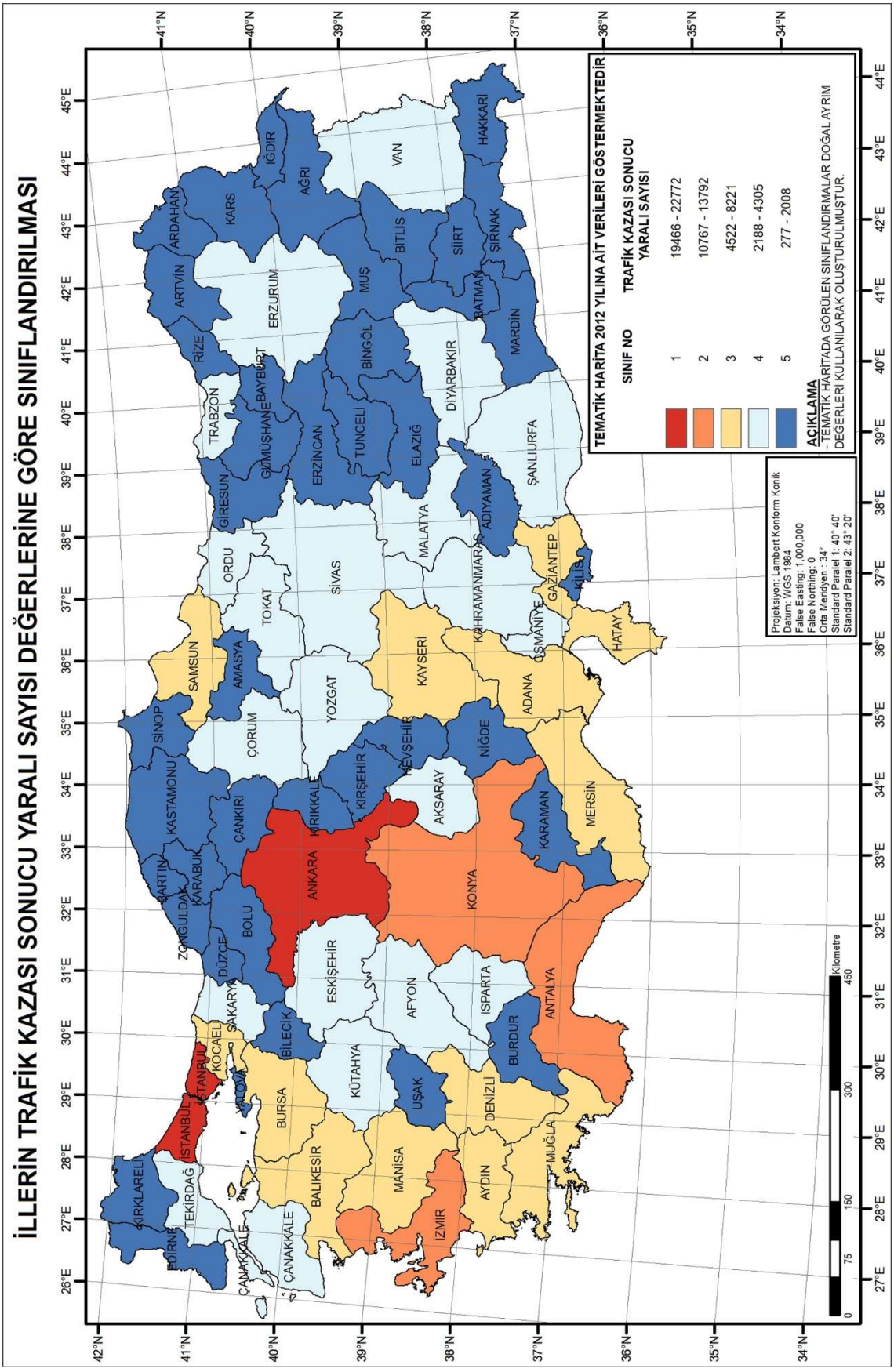
Şekil 4.11 2012 Yılı Trafik Kazası Sonucu Ölüm Sayısına Göre Üretilen Tek Değişkenli Harita



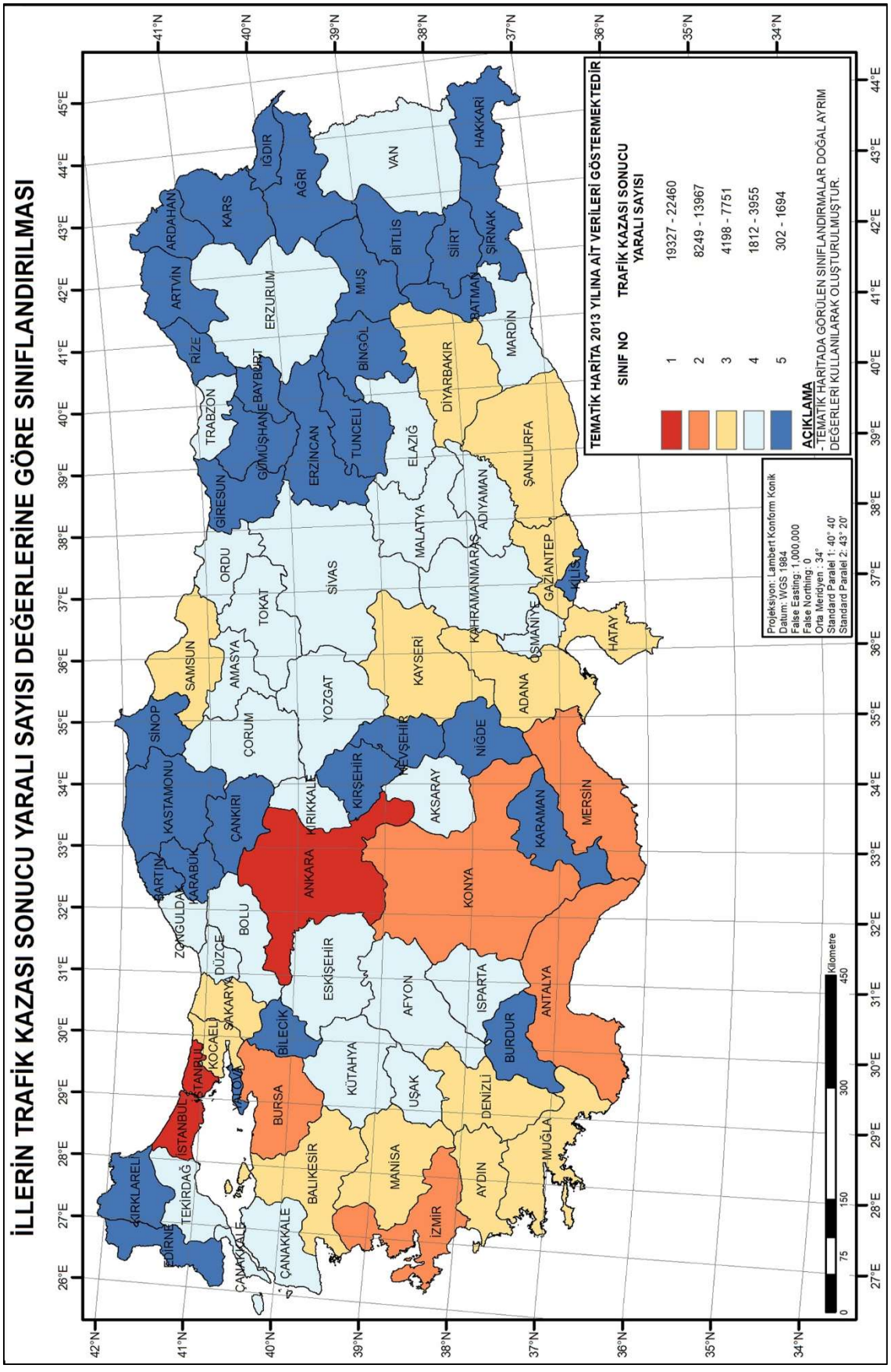
Şekil 4.12 2013 Yılı Trafik Kazası Sonucu Ölüm Sayısına Göre Üretilen Tek Değişkenli Harita



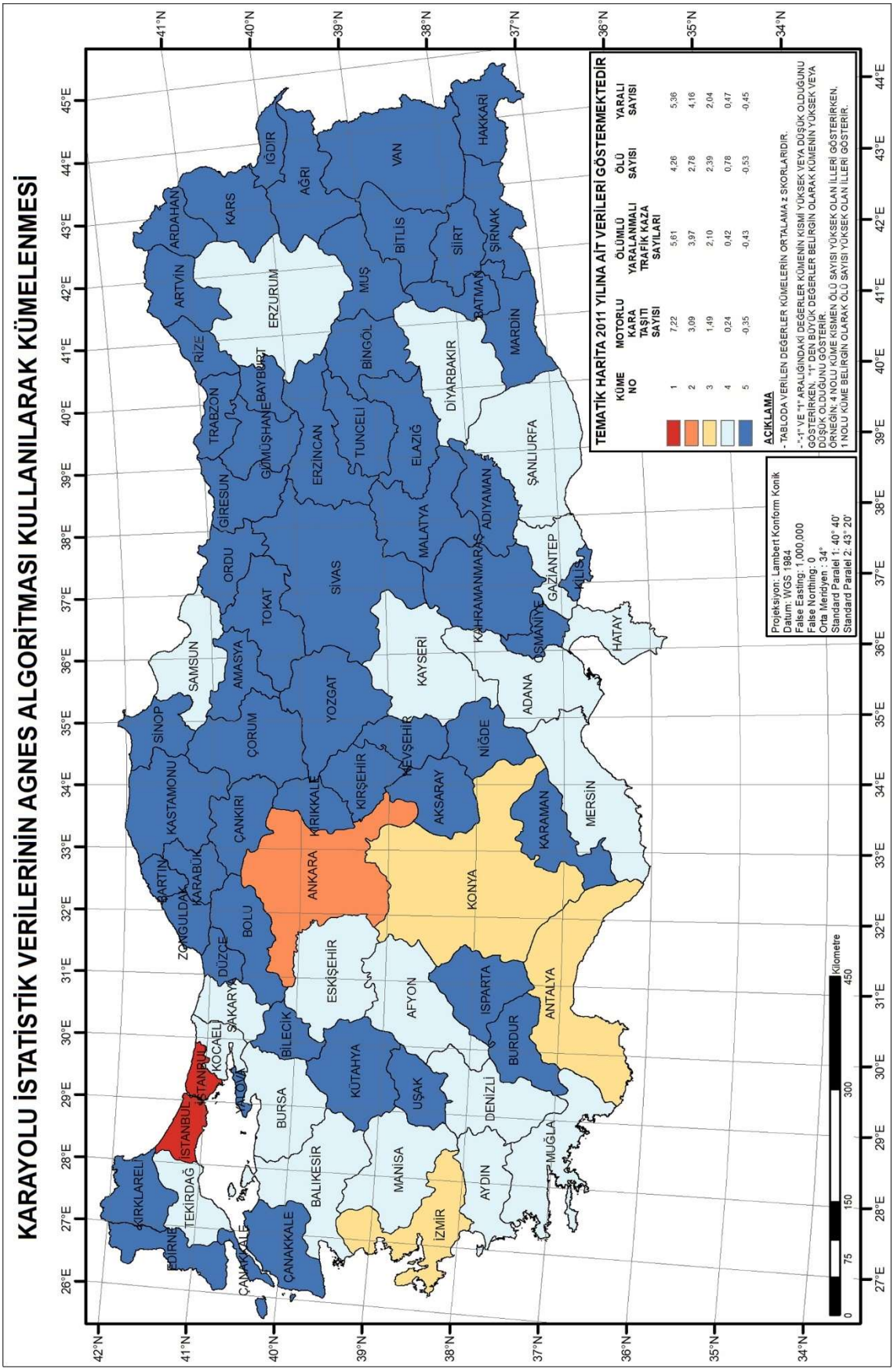
Şekil 4.13 2011 Yılı Trafik Kazası Sonucu Yaralı Sayısına Göre Üretilen Tek Değişkenli Harita



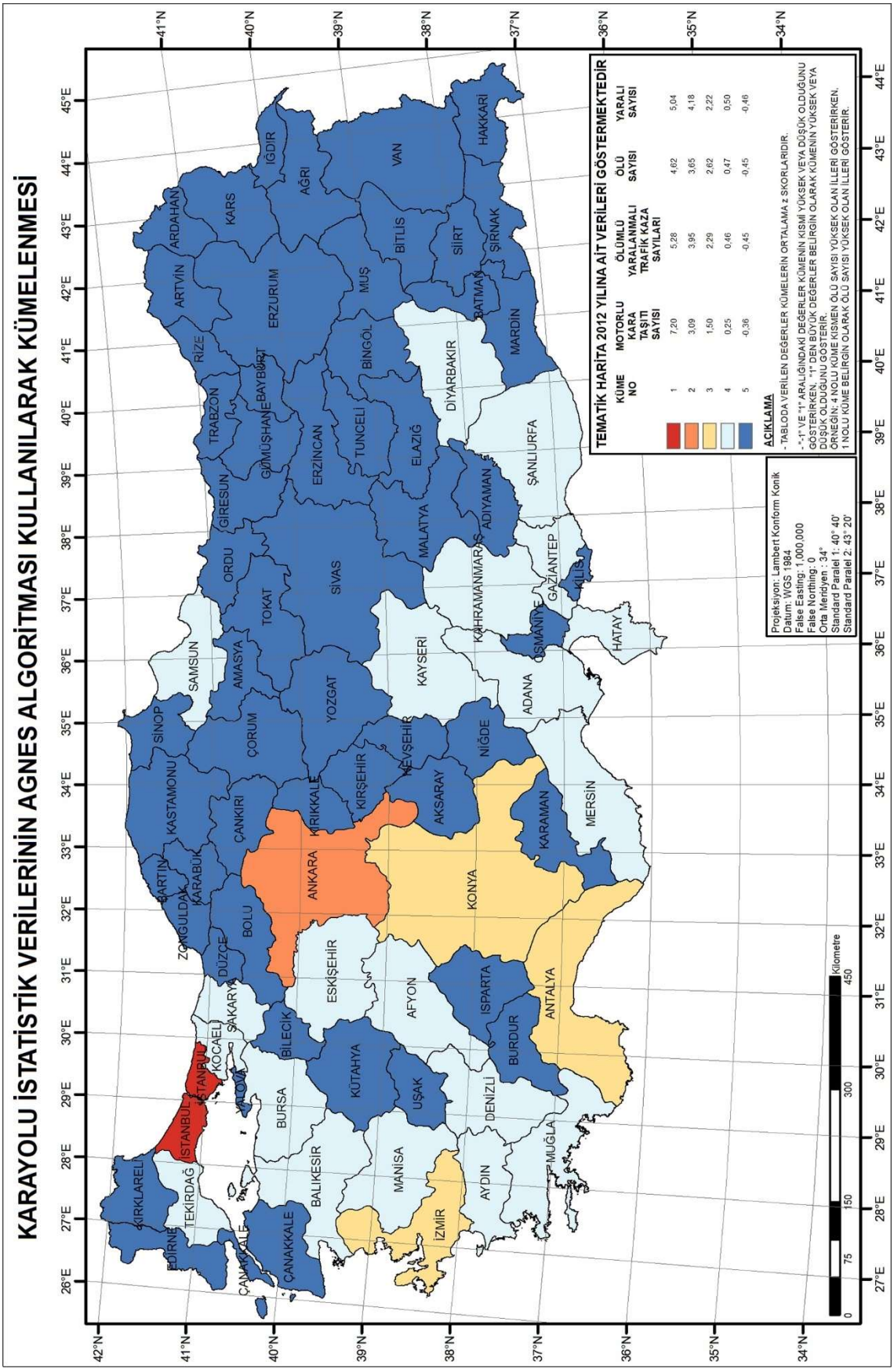
Şekil 4.14 2012 Yılı Trafik Kazası Sonucu Yaralı Sayısına Göre Üretilen Tek Değişkenli Harita



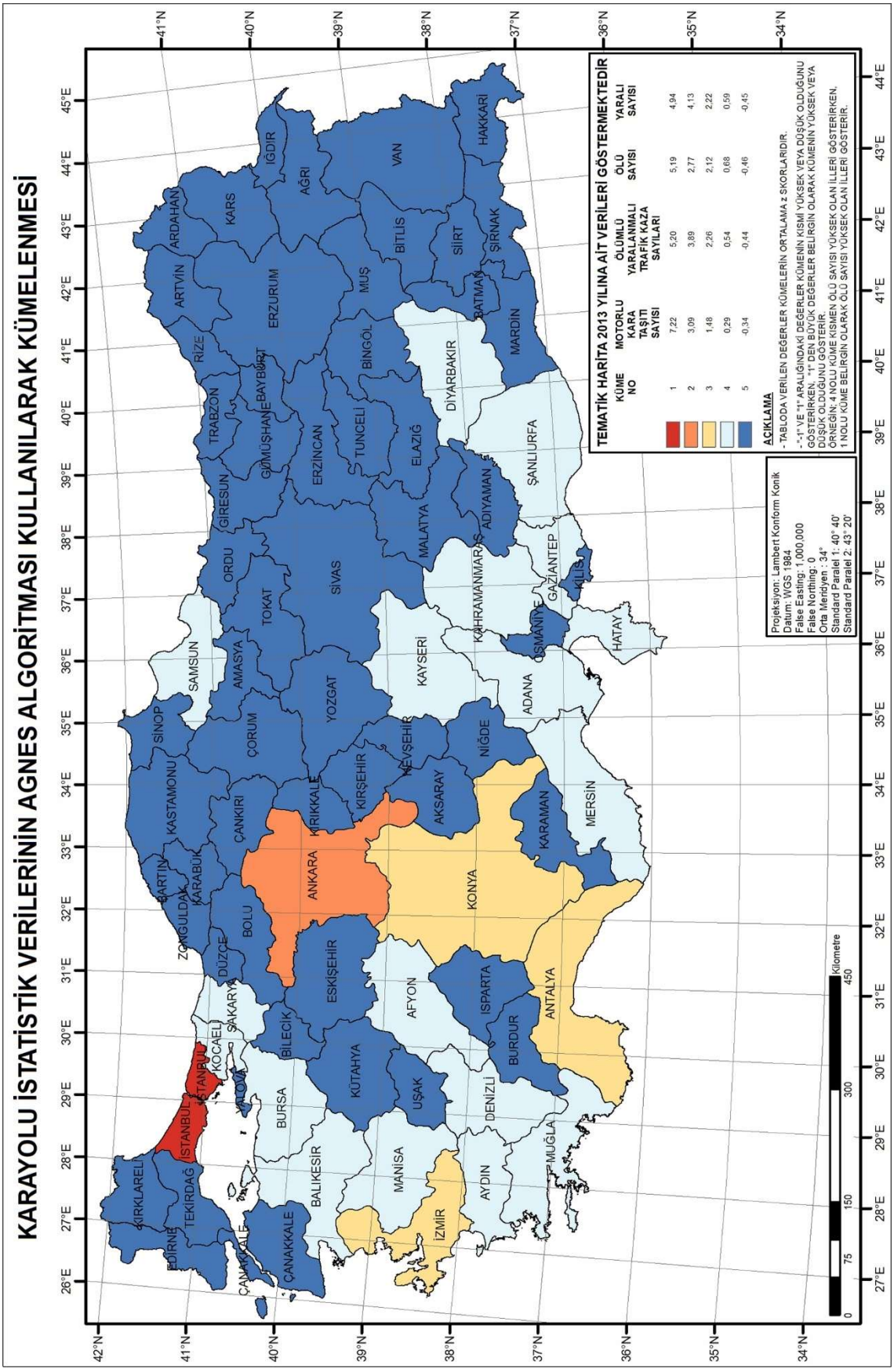
Şekil 4.15 2013 Yılı Trafik Kazası Sonucu Yaralı Sayısına Göre Üretilen Tek Değişkenli Harita



Şekil 4.16 2011 Yılı AGNES Metoduyla Üretilen Çok Değişkenli Harita



Şekil 4.17 2012 Yılı AGNES Metoduyla Üretilen Çok Değişkenli Harita

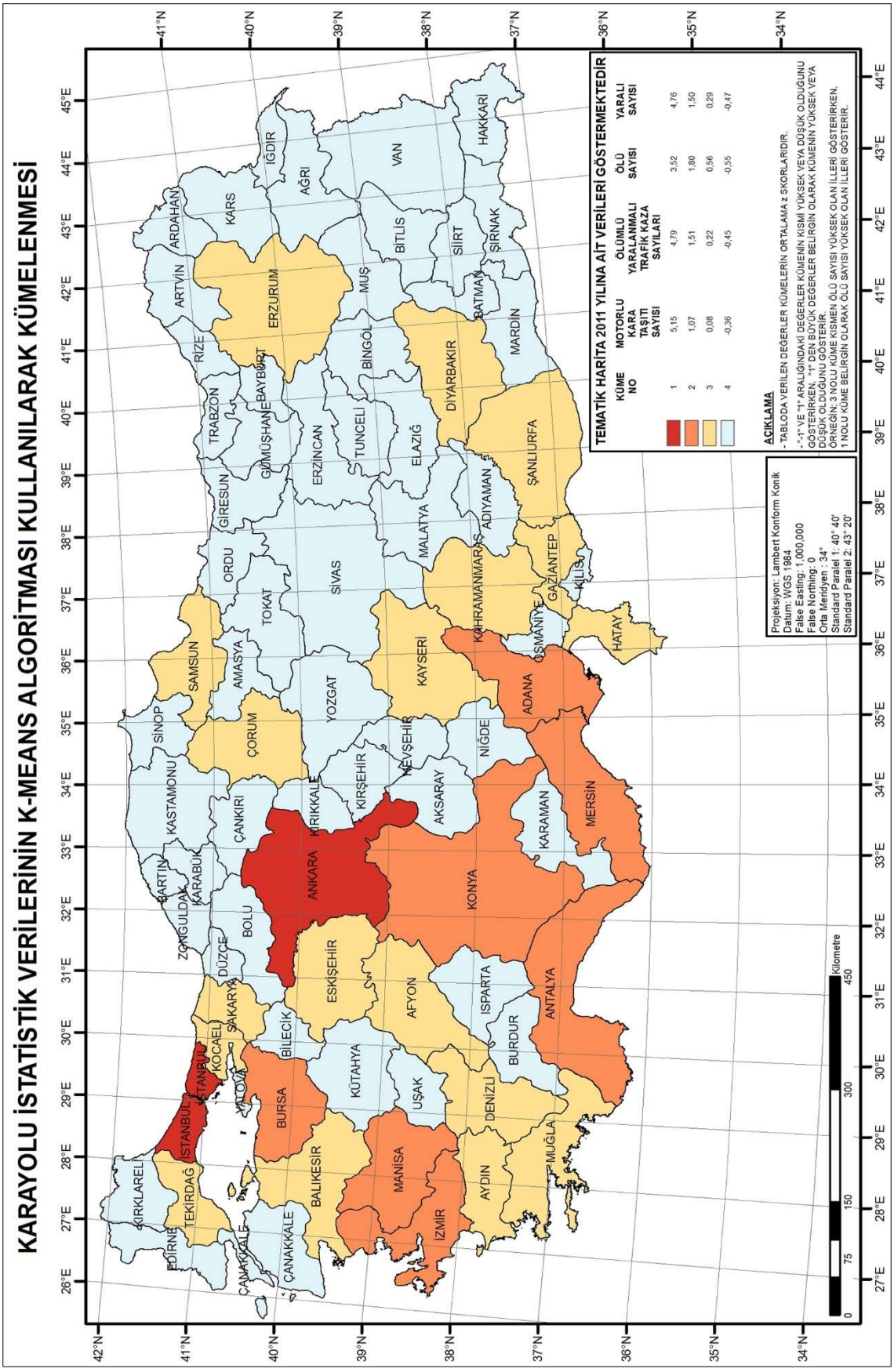


Şekil 4.18 2013 Yılı AGNES Metoduyla Üretilen Çok Değişkenli Harita

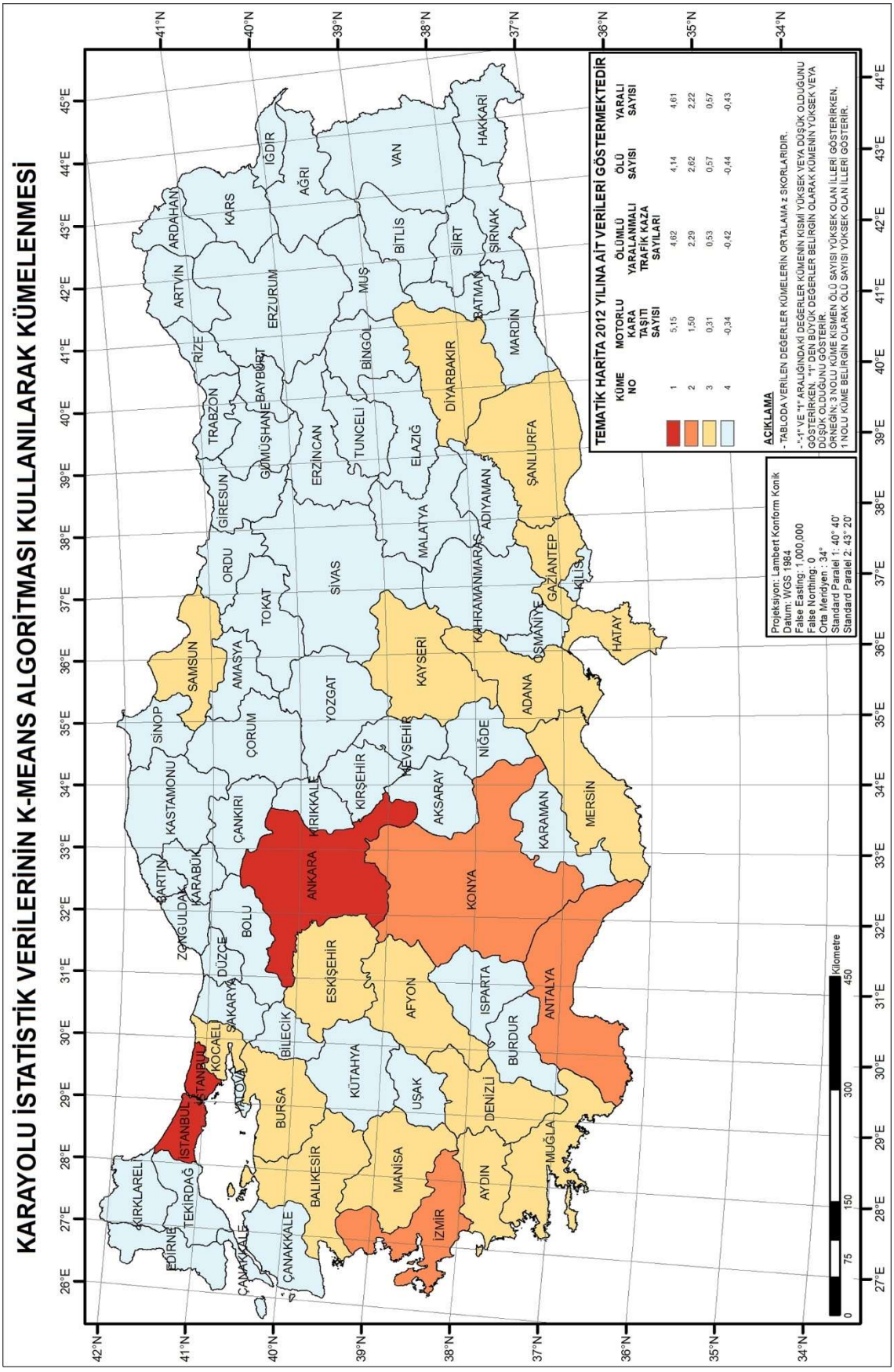
(Şekil 4.16 - Şekil 4.18)'de görülen AGNES kümeleme analizi sonuçlarıyla üretilen 2011, 2012 ve 2013 yıllarına ait çok değişkenli haritalar incelendiğinde; 1 ve 2 nolu kümeleri Türkiye'nin metropol illeri olan İstanbul ve Ankara illerinin; 3 nolu Antalya, İzmir ve Konya büyükşehirleri; 4 nolu kümeyi Adana, Aydın, Balıkesir, Bursa, Denizli, Diyarbakır, Erzurum, Eskişehir, Gaziantep, Hatay, Kayseri, Kocaeli, Manisa, Mersin, Muğla, Sakarya, Samsun, Şanlıurfa ve Tekirdağ büyükşehirlerinin yanı sıra büyükşehir olmayan Afyonkarahisar ili; 5 nolu kümeyi ise büyükşehir olmayan 50 il ile birlikte Kahramanmaraş, Malatya, Mardin, Ordu, Trabzon ve Van büyükşehirleri oluşturmaktadır.

Eğer veri setleri 4 farklı değer (motorlu kara taşıtı sayısı, ölümlü ve yaralanmalı trafik kaza sayısı, ölü sayısı ve yaralı sayısı) kullanmak yerine sadece motorlu kara taşıtı sayısı dikkate alınsaydı 1 nolu küme İstanbul; 2 nolu küme Ankara ve İzmir; 3 nolu küme Antalya, Konya, Adana, Balıkesir, Bursa, Gaziantep, Hatay, Manisa, Mersin, Muğla; 4 nolu küme Denizli, Eskişehir, Kayseri, Kocaeli, Sakarya, Samsun, Tekirdağ, Şanlıurfa, Kahramanmaraş, Afyonkarahisar, Kütahya, Aydın, Çanakkale ve 5 nolu küme diğer 55 il şeklinde oluşacaktı (Şekil 4.4 – Şekil 4.6).

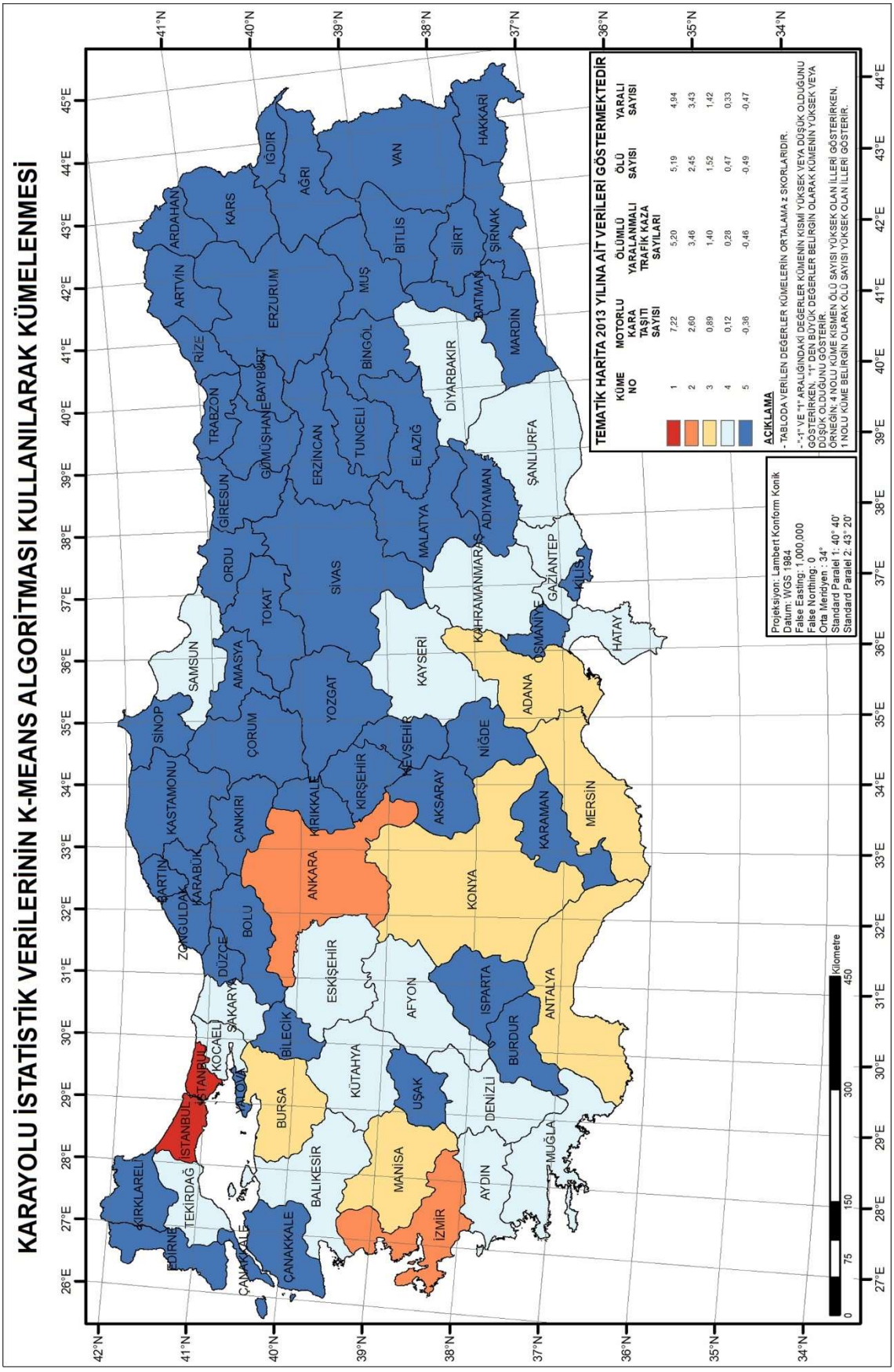
(Şekil 4.4 – Şekil 4.6) ve (Şekil 4.16 – Şekil 4.18)'de yer alan haritalar karşılaştırıldığında; İzmir ilinin tek değişkenli haritalarda Ankara ili ile birlikte bir sınıf oluştururken çok değişkenli haritalarda Konya ve Antalya illeri ile bir küme oluşturduğu görülebilmektedir. Yine (Şekil 4.4 – Şekil 4.6)'te yer alan haritalarda Konya ve Antalya illerinin Balıkesir, Bursa, Gaziantep, Hatay, Manisa, Mersin ve Muğla illeriyle bir sınıf oluştururken (Şekil 4.16 – Şekil 4.18)'deki çok değişkenli haritalarda İzmir iliyle ayrı bir küme oluşturdukları görülmektedir. Bu karşılaştırmadan; İzmir ilinin her ne kadar motorlu kara taşıtı sayısına göre Ankara iliyle aynı sınıfta yer alsa da ölümlü yaralanmalı trafik kaza sayısı, trafik kazası sonucu ölüm ve yaralanma sayısına göre kümelendirildiğinde Ankara ilinden daha güvenli olduğu; Konya ve Antalya illerininde aynı şekilde her ne kadar motorlu kara taşıtı sayısına göre Balıkesir, Bursa, Gaziantep, Hatay, Manisa, Mersin ve Muğla illeriyle aynı sınıfta yer alsa da ölümlü yaralanmalı trafik kaza sayısı, trafik kazası sonucu ölüm ve yaralanma sayısına göre kümelendirildiğinde Balıkesir, Bursa, Gaziantep, Hatay, Manisa, Mersin ve Muğla llerinde daha tehlikeli olduğu sonuçlarına ulaşılmaktadır.



Şekil 4.19 2011 Yılı k-Ortalama Metoduyla Üretilen Çok Değişkenli Harita



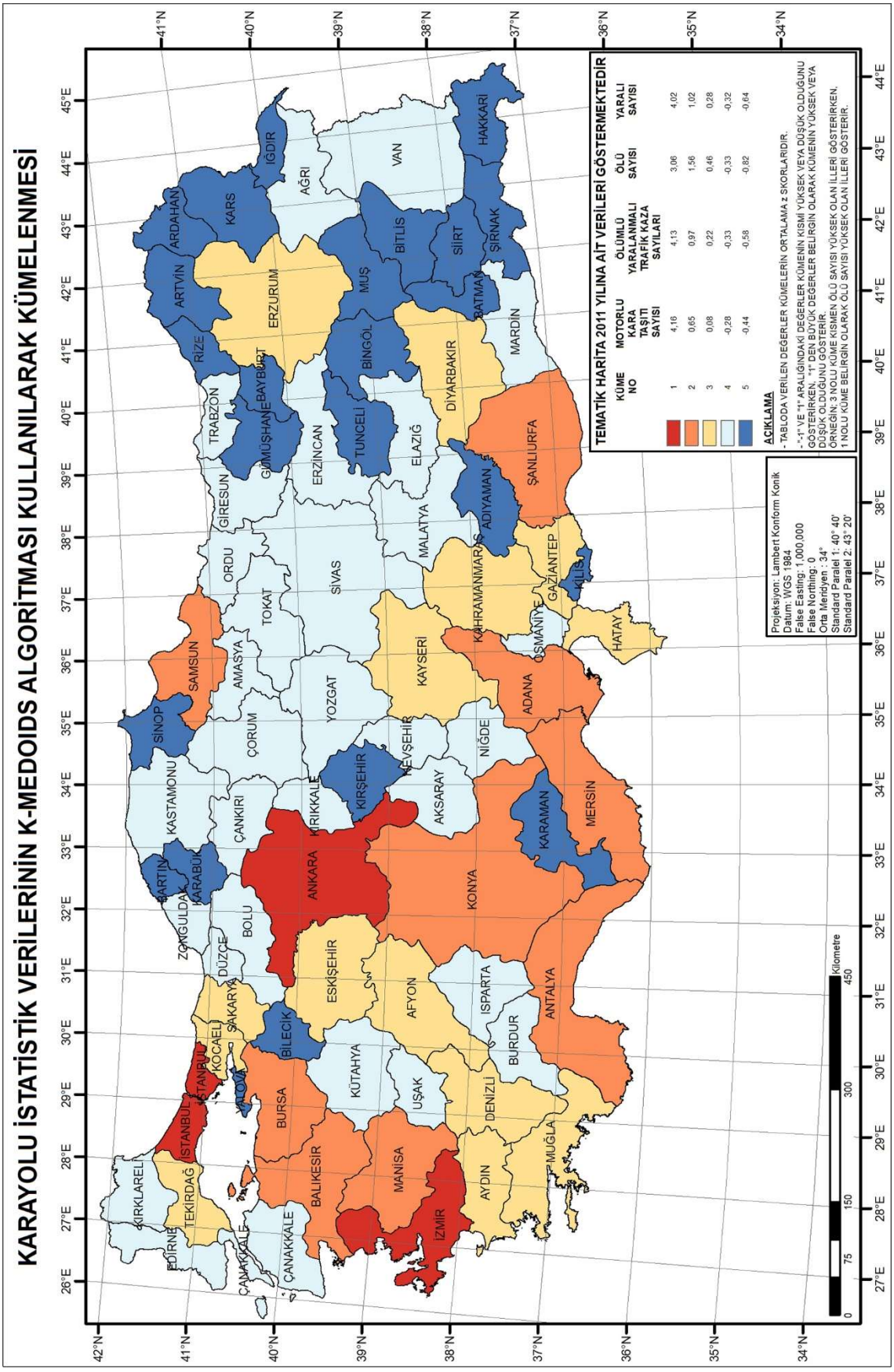
Şekil 4.20 2012 Yılı k-Ortalama Metoduyla Üretilen Çok Değişkenli Harita



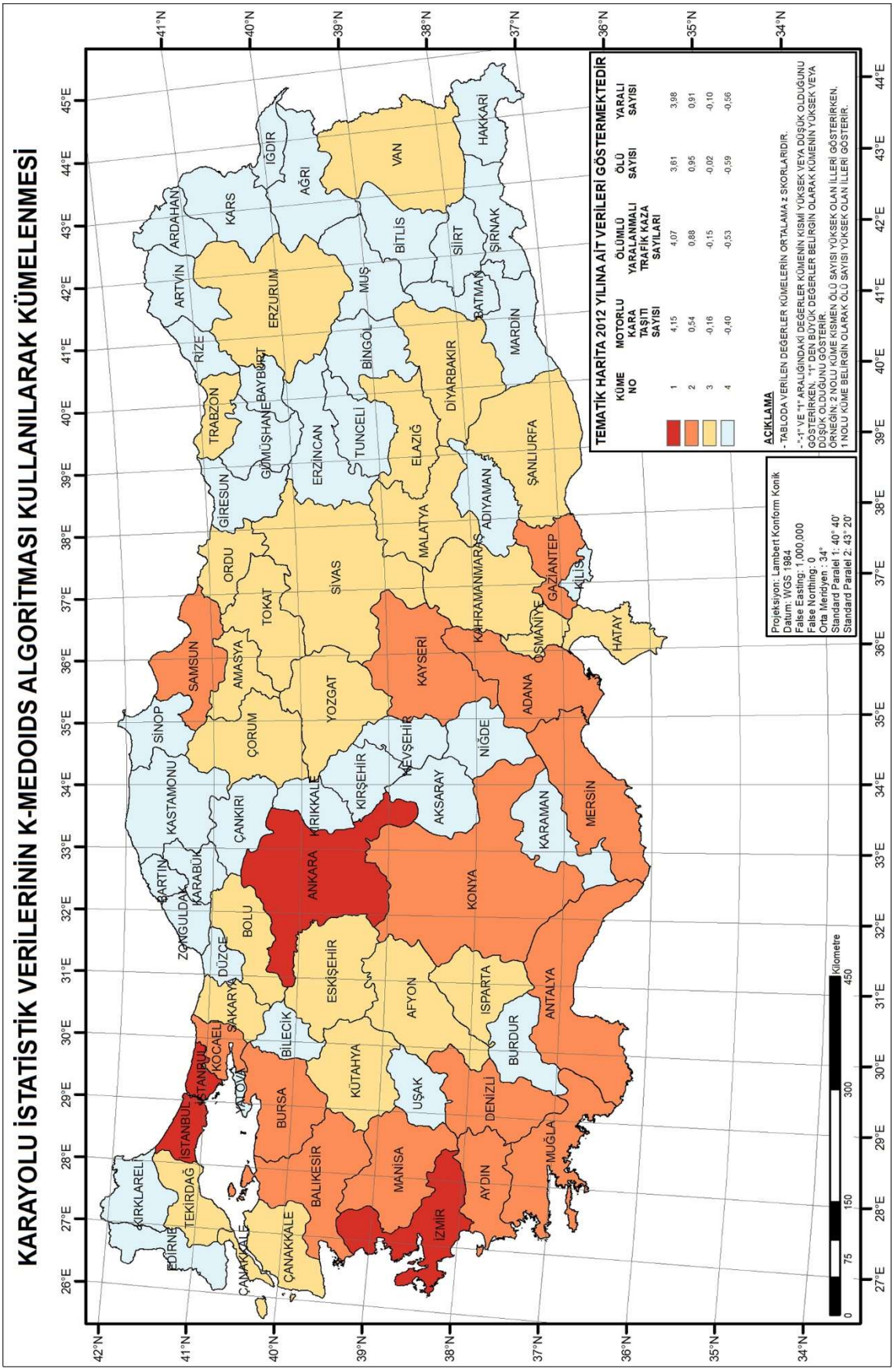
Şekil 4.21 2013 Yılı k-Ortalama Metoduyla Üretilen Çok Değişkenli Harita

(Şekil 4.19 – Şekil 4.21)'de görülen k-means kümeleme analizi sonuçlarıyla üretilen 2011, 2012 ve 2013 yıllarına ait çok değişkenli haritalar incelendiğinde; 2011 ve 2012 yıllarının 4 kümeye ayrılırken 2013 yılının 5 kümeye ayrıldığı görülmektedir. (Şekil 4.16 - Şekil 4.18)'de yer alan AGNES kümeleme analizi sonuçlarına göre üretilen çok değişkenli haritalar ile (Şekil 4.19 – Şekil 4.21)'de k-means kümeleme analizi sonuçlarına göre üretilen çok değişkenli haritalar incelendiğinde küme sayısında değişiklikler görülsede verilerin değerlendirilmesinde çok büyük farklılıkların olmadığı görülmektedir. Ankara ilinin (Şekil 4.19 – Şekil 4.20)'de 2011 ve 2012 yıllarına ait haritalarda oluşturulan küme sayısından dolayı İstanbul ili ile aynı küme olduğu, küme sayısı değişsede Konya ve Antalya illerinin aynı kümenin ve Balıkesir, Eskişehir,Afyon, Denizli, Aydın ve Muğla illerinin aynı kümenin elemanları olduğu görülmektedir.

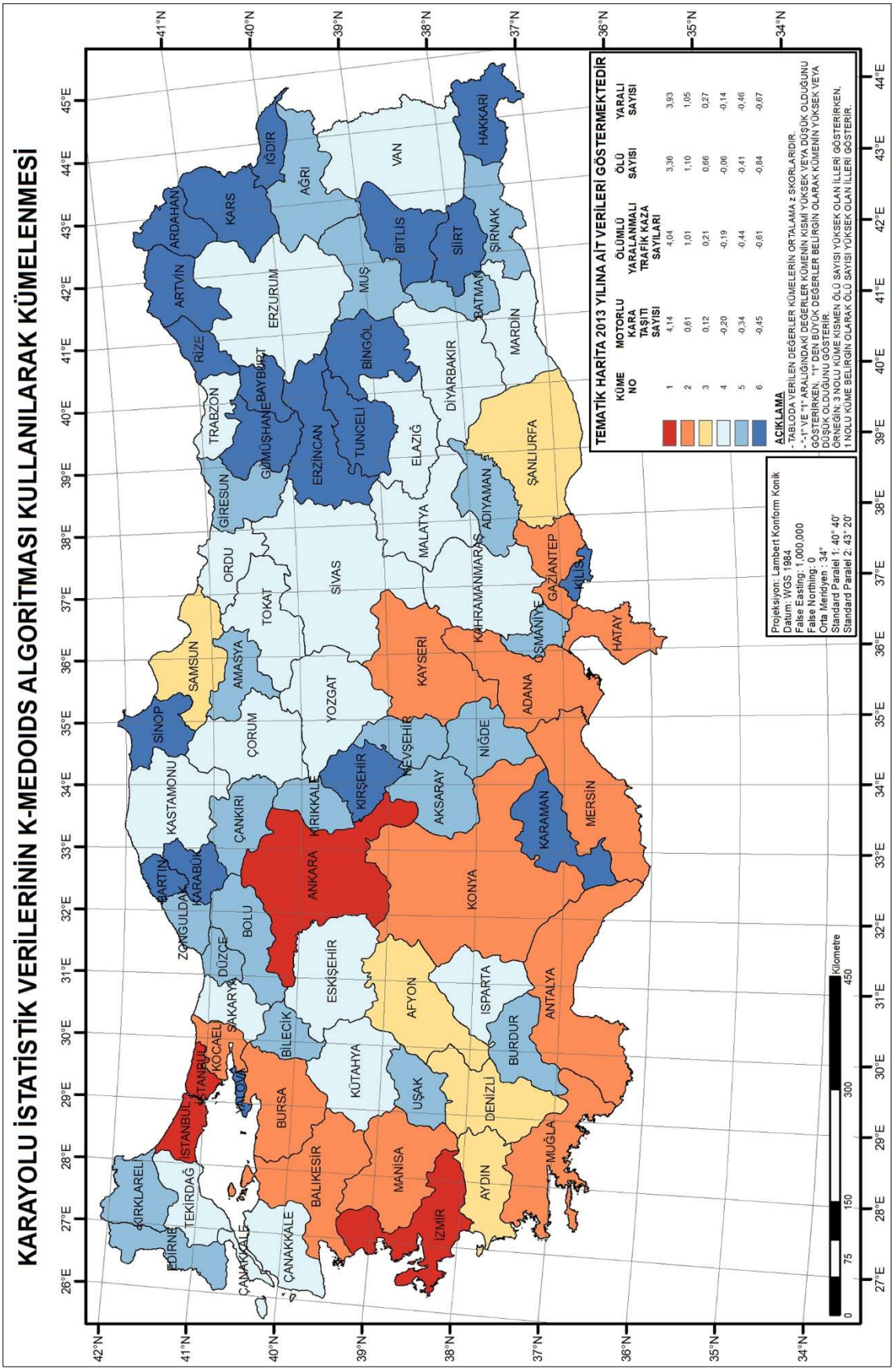
(Şekil 4.7 – Şekil 4.9) ve (Şekil 4.19 – Şekil 4.21)'de yer alan 2013 yılına ait haritalar karşılaştırıldığında; İstanbul, Ankara ve İzmir illerinin tek değişkenli (ölümlü yaralanmalı trafik kaza sayısı) haritalarda aynı sınıfın elemanı olduğu görülürken çok değişkenli haritada farklı bir küme oluşturduğu; Konya, Antalya, Adana, Bursa ve Mersin illerinin tek değişkenli harita görülen sınıflanma yapısını çok değişkenli haritada da aynı kümenin elemanları olarak birbirleriyle olan ilişkilerini korudukları; Manisa ilinin ise tek değişkenli harita da Kocaeli, Sakarya, Balıkesir, Aydın, Denizli, Muğla, Samsun, Kayseri, Gaziantep ve Hatay illeriyla olan sınıflandırma yapısını çok değişkenli haritada koruyamadığı, ölümlü yaralanmalı trafik kaza sayısı sonucu meydana gelen ölüm ve yaralanma sayılarındaki yükseklik sebebiyle Konya, Antalya, Adana, Bursa ve Mersin illeriyle aynı kümeye dahil olduğu; Çorum ve Trabzon illerinin tek değişkenli haritada Eskişehir, Diyarbakır, Şanlıurfa, Tekirdağ ve Afyon illeriyle aynı sınıfı paylaşırken çok değişkenli haritada bu ilişkiyi koruyamadıkları ve trafik kazaları sonucunda gerçekleşen ölüm ve yaralanma sayısındaki düşüklük nedeniyle Sinop, Rize ve Çankırı illerinin bulunduğu kümeye dahil oldukları görülmektedir. k-means algoritması ile yaptığımız kümeleme değerlendirmesinin her zaman kesin doğru sonuç verdiğini söylemek doğru değildir. Özellikle Şekil 4.9'da yer alan 2013 yılı haritasında Erzurum ve Van illerine baktığımızda ölümlü yaralanmalı trafik kaza sayısı sınıflandırmasında Eskişehir, Afyon ve Sakarya illeri ile aynı sınıfı paylaşırken Şekil 4.21'de yer alan 2013 yılına ait çok değişkenli haritada Rize, Artvin ve Gümüşhane gibi daha küçük illerde aynı kümede bulunduğu gözlemlenmiştir. Bu sonucun Erzurum ve Van illerindeki motorlu kara taşıtı sayısındaki düşüklük nedeniyle olduğuda gözden kaçmamaktadır.



Şekil 4.22 2011 Yılı k-Medoids Metoduyla Üretilen Çok Değişkenli Harita



Şekil 4.23 2012 Yılı k-Medoids Metoduyla Üretilen Çok Değişkenli Harita



Şekil 4.24 2013 Yılı k-Medoids Metoduyla Üretilen Çok Değişkenli Harita

(Şekil 4.22 – Şekil 4.24)'de görülen k-medoids kümeleme analizi sonuçlarıyla üretilen 2011, 2012 ve 2013 yıllarına ait çok değişkenli haritalar incelendiğinde; 2011 yılındaki haritanın 5 kümeye ayrıldığı, 2012 yılındaki haritanın 4 kümeye ayrıldığı ve 2013 yılının 5 kümeye ayrıldığı görülmektedir. 2011, 2012 ve 2013 yıllarına ait verilerdeki değişikliklerden dolayı her ne kadar farklı küme sayıları kullanılsada k-medoids kümeleme analizi sonuçlarına göre üretilen çok değişkenli haritalar incelendiğinde 2011 yılında bir kümeyi oluşturan illerin diğer yıllarda da biribirleriyle olan ilişkilerinin sürdüğü görülmektedir.

(Şekil 4.22 – Şekil 4.24)'de yer alan 2011, 2012 ve 2013 yıllara ait çok değişkenli haritalar incelendiğinde her üç yılda da İstanbul, Ankara ve İzmir metropol illerinin aynı kümenin özelliklerini taşıdığı; Bursa, Balıkesir, Manisa, Konya, Antalya, Mersin ve Adana yine üç yılda da aynı kümenin özelliklerini ve Bartın, Karabük, Sinop, Kırşehir, Karaman, Gümüşhane, Bayburt, Rize, Artvin, Ardahan, Kars, Iğdır, Bingöl, Bitlis, Siirt ve Hakkari illerinin de yine üç yılda da aynı kümenin özelliklerini taşıdığı görülmektedir.

Şekil 4.9 ve Şekil 4.24'de yer alan 2013 yılına ait haritalar karşılaştırıldığında İstanbul, Ankara ve İzmir illerinin tek değişkenli (ölümlü yaralanmalı trafik kaza sayısı) haritalarda aynı sınıfın elemanı olduğu ve birbirleriyle olan bu ilişkilerini çok değişkenli haritada aynı kümenin elemanları olarak sürdürdüğü; Balıkesir, Manisa, Muğla, Kayseri, Gaziantep ve Hatay illerinin tek değişkenli haritada birbirleriyle aynı sınıfın özelliklerini taşıyıp Konya, Antalya, Bursa, Adana, Mersin ve Bursa illerinden ayrı bir sınıfın elemanı olduğu görülsede çok değişkenli haritada motorlu kara taşıtı sayısı değerinin yüksekliği sebebiyle aynı kümenin elemanları oldukları görülmektedir.

5. DEĞERLENDİRME VE SONUÇ

İnsanoğlu'nun yaşamındaki en temel kavramlardan biri şüphesiz "mekân"dır. Günümüzde elde edilen verinin yaklaşık %80'i mekânsal bileşenlere sahiptir. Veri toplama, veri işleme ve veri saklama teknolojilerindeki yaşanan gelişmelerle birlikte mekânsal veri setlerinin miktarı ve kapsamı da hızla büyümektedir. Veri Madenciliği bu büyük veri setlerinde gizlenmiş olan mekânsal bilgiyi ortaya çıkarmada kullanılan önemli bir disiplindir. Günümüzde birçok insan veri madenciliği disiplinini farkına varmaksızın mekânsal özellikleri keşfetmekte kullanmaktadır. Örneğin; bir emlak-konut işlemi (alım-satım-kiralama) yapmak isteyen kullanıcı web ortamında servis veren siteler (sahibinden, zingat vb.) üzerinden kendine özgü belirlediği değişkenleri (metrekare, oda sayısı, kat sayısı, ısınma türü vb.) kullanarak aradığı binayı hangi mekânsal konumda (il, ilçe, köy, mahalle) bulabileceğinin sorgulamasını kolaylıkla yapmaktadır. Veri madenciliği disiplini ile mekânsal olan veriler ile mekânsal olmayan öznitelik bilgeleri arasındaki örüntüler ortaya çıkarılmaktadır. Arazi verimliliklerinin değerlendirilmesi, su havzası kalitesinin analizi, lojistik rota planlaması vb. uygulamalar mekânsal veri madenciliğinin örnek uygulamaları arasında yer almaktadır.

AGNES kümeleme analizi sonucu kullanılarak üretilen harita ve Dendrogramlar üzerinden belirlenen kümeler incelendiğinde AGNES yöntemiyle üretilen 2011, 2012 ve 2013 haritalarının birbirleriyle oldukça uyumlu olduğu görülmektedir. Bu sonuç AGNES yöntemiyle üretilen çok değişkenli haritaların risk yönetimi açısından da oldukça önemli olduğunu göstermektedir. Çünkü 2011 verileriyle öngörülen risk bölgeleri 2012 ve 2013 verilerinde de doğrulanmıştır.

Hiyerarşik olmayan k-ortalama ve k-Medoids kümeleme algoritmaları k giriş parametresine göre n tane nesneyi k tane kümeye bölme mantığıyla çalışmaktadır. Nesneler birbirlerine benziyor ve diğer kümelerdeki nesnelere benzemiyorlarsa aynı kümeyi oluşturmaktadırlar. Bu algoritmaların uygulanmasında ki en büyük sorun oluşturulacak k küme sayısının belirlenmesidir. k sayısının belirlenmesinde Dunn indeksi ve Davies-Bouldin indeksi testlerinden yararlanılabilir. Bu da yapılan birkaç uygulama deneyimi ile belirlenebilmektedir. Yapılan çalışmada kullanılan veri setleri için farklı k değerleri için kümeleme sonuçları gözlemlenmiştir. Her iki algoritmanın da kümeleme başarımları benzerlik gösterse de her iki yöntemle oluşturulan kümelerin ortalama z-skoru tabloları incelendiğinde k-Medoids algoritmasında kümelerin birbirinden daha iyi

ayrıldığı gözlemlenmiştir. Kümeleme çalışmalarındaki amaç küme içi benzerliklerin maksimum, kümeler arası benzerliklerin ise minimum olması olduğundan k-Medoids yönteminin bu veriler için daha iyi sonuç verdiği söylenebilir.

Oluşturulan haritalar incelendiğinde; İstanbul, Ankara, İzmir gibi metropol şehirlerin bu çalışmada ele alınan 4 değişken (motorlu kara taşıtı sayısı, ölümlü ve yaralanmalı trafik kaza sayıları, ölü sayıları ve yaralı sayıları) dikkate alındığında aynı risk seviyesinde ve benzer kümelerde olduğu görülmektedir. Yine Antalya, Konya, Eskişehir, Afyon gibi turistik ve ana ulaşım koridorundaki şehirlerde de oluşan trafik kazalarında benzerlikler görülmektedir. Ayrıca nüfus yoğunluğu düşük olan ve ana yol güzergâhlarında bulunmayan Bartın, Karabük, Kırşehir, Tunceli, Bitlis vb. şehirlerin de bu 4 parametre dikkate alındığında trafik kazalarının niteliği açısından aynı risk grubunda oldukları gözlenmektedir.

Bu çalışmayla kümeleme yöntemlerinin kullanılmasıyla birden fazla özellik dikkate alınarak farklı mekânsal objelerin benzer yönlerinin ortaya çıkarılabileceği gösterilmiştir.

Mekânsal ve mekânsal olmayan veri setlerinin veri madenciliği disiplini kullanılarak analizlerinin yapılması ve analiz sonuçlarının kartografik görselleme teknikleriyle hazırlanan ve mekânsal - mekânsal olmayan verilerin birbirleriyle etkileşimli olarak sunumunun yapıldığı çok değişkenli haritaların ekolojik, çevresel yönetim, ulaşım, hak güvenliği, ticari, tarım, endüstri, turizm, seyahat, risk analizi, ulusal savunma vb. konularda uzman kararlarının alınmasında önemli katkılar sağlayacağı düşünülmektedir.

KAYNAKLAR

Akat, Y., 2007, Ülkelerin Askeri Benzerliklerine Göre Kümeleme Analizi Yardımıyla Sınıflandırılması, Yüksek Lisans Tezi, *İstanbul Teknik Üniversitesi Fen Bilimleri Üniversitesi*, İstanbul.

Akın, Y., K., 2008, Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi, Doktora Tezi, *Marmara Üniversitesi Sosyal Bilimler Enstitüsü*, İstanbul.

Akpınar, H., 2000, Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, *İ.Ü. İşletme Fakültesi Dergisi*, C:29, S: 1/Nisan 2000.

Alkan, H., 2012, Kümeleme Analizi İle Elektrik Tüketiminin Sınıflandırılması, Yüksek Lisans Tezi, *Fırat Üniversitesi Fen Bilimleri Enstitüsü*, Elazığ.

Apte, C., Liu, B., Pednault, E.P.D., Symth., P., 2002, Business Application of Data Mining, Communications of The Acm August 2002/Vol. 45, No. 8.

Argüden, Y. ve Erşahin, B., 2008, Veri Madenciliği, ARGE Yayınları, İstanbul.

Atalay, A. ve Tortum, A., 2010, Türkiye'deki İllerin 1997-2006 Yılları Arası Trafik Kazalarına Göre Kümeleme Analizi, *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, Cilt 16, Sayı 3, Sayfa 335-343.

Berry Michael J., Linoff, Gordon S., 2004, Data Mining Techniques For Marketing Sales And Customer Support, New York: Wiley

Buckley, A., 2008, Multivariate Mapping. In K. KEMP (Eds.) Encyclopedia of Geographic Information Science (pp. 300-303).

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi, A., 1998, Discovering Data Mining: From Concept to Implementation, Prentice Hall, Upper Saddle River, NJ.

Çakmak, Z., Uzgören, N., Keçek, G., 2005, "Kümeleme analizi teknikleri ile illerin kültürel yapılarına göre sınıflandırılması ve değişimlerinin incelenmesi", *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, cilt12, s. 15-36.

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

Çetinkaya, S., 2008, İstanbul'daki Binaların Veri Madenciliği Yaklaşımıyla Kümelenmesi, Yüksek Lisans Tezi, *Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü*, İstanbul.

Çiçekdağı, H.İ., 2013, Deprem Verilerinin Analizi İçin Veri Madenciliği Yaklaşımı, Yüksek Lisans Tezi, *Dumlupınar Üniversitesi Fen Bilimleri Enstitüsü*, Kütahya.

DiBiase, D., 1994, "Designing Animated Maps for A Multimedia Encyclopedia", Cartographic Perspectives,19, 3-7.

DiBiase, D., Sloan, J. L., and Paradis, T. (1994). Weighted isolines: an alternative method for depicting statistical surfaces. *Professional Geographer*, *46*(2), 218-228.

Dinçer E.Ş., 2006, Veri Madenciliğinde K-means Algoritması ve Tıp Alanında Uygulanması, *Kocaeli Üniversitesi Fen Bilimleri Enstitüsü*, Kocaeli.

Gorunescu, F., 2011, Data Mining Concepts, Models and Techniques, Springer-Verlag Berlin Heidelberg.

Han, J. ve Kamber, M., 2006, Data Mining: Concepts and Techniques, San Francisco.

Han, J., Kamber, M ve Pei, J.,2011, Data Mining: Concepts and Techniques, Waltham.

Jenks, G. F., 1953, "Pointillism as a Cartographic Technique", *The Professional Geographer*, 5, 4-6.

Karpat, G. ve Yılmaz, V., 2002, "Türkiye'deki Trafik Kazaları Oluş Şekillerinin, Kazanın Olduğu Yerdeki Trafik, Aydınlatma ve Yol Durumuna Göre İller Bazında İncelenmesi", Uluslararası Trafik ve Yol Güvenliği Kongresi, Gazi Üniv., Ankara.

Kocabaş, F.M., 2010, Veri Madenciliği Süreci ve Gerçek Bir Veri Seri Üzerinde Uygulanması, Yüksek Lisans Tezi, *Hacettepe Üniversitesi Fen Bilimleri Enstitüsü*, Ankara.

Larose, D.T., 2005, Discovering Knowledge in Data: An Introduction to Data Mining, Wiley Publishing.

Luan, J., 2002, Data Mining and Its Applications in Higher Education.

Nelson, E.S., 2000, "Designing Effective Bivariate Symbols: The Influence of Perceptual Grouping Processes", *Cartography and Geographic Information Science*, 27, 4, 261-278.

Özkan, Y., 2008, Veri Madenciliği Yöntemleri, Papatya Yayıncılık, İstanbul.

Reyes, N. J. J., 2009, Ideas For the Use of Chernoff Faces in School, *24. International Cartographic Conference*, Chile.

Romesburg H.C., 1984, Cluster Analysis for Researchers, Belmont, CA: Life time Learning Publications.

Sarıman, G., 2011, Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması, *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 15-3, 192-202.

Silahtaroğlu, G., 2013, Veri Madenciliği (Kavram ve Algoritmaları), Papatya Yayıncılık, İstanbul.

Slocum, T.A., McMaster, R.B., Kessler, F.C., Howard, H.H., 2009, Thematic Cartography and Geovisualization, Pearson Education Inc. Third Edition, USA.

Şekerler, A. ve Murat Y.Ş., 2009, Trafik Kaza Verilerinin Kümeleme Analizi Yöntemi ile Modellenmesi, *İMO Teknik Dergi*, 4759-4777, Yazı 311.

SPSS, 1999, Field-Tested Data Mining 10 Essential Strategies and Tips.

Yılmaz, Ö., Temurlenk, Ö.Y., 2005, Türkiye'deki İstatistik Bölgelerin Kişi Başına Düşen Gelir Açısından Hiyerarşik ve Hiyerarşik Olmayan Kümeleme Analizi İle Değerlendirilmesi:1965–2001, Atatürk Üniversitesi İİBF, 19(2), 75-92.

Yomralıoğlu, T., 2000, Coğrafi Bilgi Sistemleri: Temel kavramlar ve uygulamalar, İstanbul.

URL1: https://tr.wikipedia.org/wiki/SPSS [Ziyaret tarihi: 3 Aralık 2017]

URL2: https://www.ibm.com/analytics/data-science/predictive-analytics/spss-statistical-software [Ziyaret tarihi: 3 Aralık 2017]

URL3: https://tr.wikipedia.org/wiki/RapidMiner [Ziyaret tarihi: 3 Aralık 2017]

URL4: https://rapidminer.com [Ziyaret tarihi: 3 Aralık 2017]

URL5: http://deim.urv.cat/~sergio.gomez/multidendrograms.php [Ziyaret tarihi: 3 Aralık 2017]

URL6: http://desktop.arcgis.com/en/ [Ziyaret tarihi: 3 Aralık 2017]

URL7: http://colorbrewer2.org [Ziyaret tarihi: 3 Aralık 2017]

EKLER

Ek1-a 2011 Yılı TUİK Trafik Kaza İstatistikleri Karayolu Verileri

SEHIR KODU	SEHIR ADI	MOTORLU KARA TAŞITI SAYISI	ÖLÜMLÜ YARALANMALI TRAFİK KAZA SAYILARI	ÖLÜ SAYILARI	YARALI SAYILARI
TR100	İstanbul	2927650	13887	226	21388
TR211	Tekirdağ	171505	1423	70	2607
TR212	Edirne	115352	715	25	1303
TR213	Kırklareli	89304	704	22	1283
TR221	Balıkesir	342478	2819	107	4958
TR222	Çanakkale	157312	1150	45	2103
TR310	İzmir	1020070	7770	137	11634
TR321	Aydın	309949	2119	59	3724
TR322	Denizli	285895	2245	83	3969
TR323	Muğla	341388	2905	76	4747
TR331	Manisa	431497	3135	113	5669
TR332	Afyonkarahisar	161638	1648	78	3339
TR333	Kütahya	158346	1131	34	2157
TR334	Uşak	96956	839	26	1758
TR411	Bursa	574926	4541	101	7563
TR412	Eskişehir	198841	2041	70	3482
TR413	Bilecik	46512	476	18	954
TR421	Kocaeli	258555	2865	69	5249
TR422	Sakarya	187110	1850	65	3327
TR423	Düzce	72118	898	20	1725
TR424	Bolu	81422	864	30	1871
TR425	Yalova	37309	383	15	645
TR510	Ankara	1367427	10318	164	17251
TR521	Konya	522578	4825	152	8787
TR522	Karaman	69810	589	18	1101
TR611	Antalya	747530	6037	154	9452
TR612	Isparta	130880	1062	37	2030
TR613	Burdur	101994	776	25	1446
TR621	Adana	480321	4222	95	6935
TR622	Mersin	433176	4048	108	6526
TR631	Hatay	348594	2211	50	3861
TR632	Kahramanmaraş	153379	1699	46	3454
TR633	Osmaniye	112545	1034	26	1765
TR711	Kırıkkale	49503	781	37	1677
TR712	Aksaray	80583	865	21	1809
TR713	Niğde	72178	659	43	1231
TR714	Nevşehir	81996	698	22	1406
TR715	Kırşehir	45236	443	8	996
TR721	Kayseri	262112	2906	69	5473
TR722	Sivas	109078	1328	30	3011
TR723	Yozgat	78866	928	29	2216

TR811	Zonguldak	115457	893	27	1797
TR812	Karabük	48098	476	23	938
TR813	Bartın	36405	393	5	806
TR821	Kastamonu	94387	803	46	1685
TR822	Çankırı	34517	562	35	1361
TR823	Sinop	42043	400	22	732
TR831	Samsun	244242	2293	85	4459
TR832	Tokat	127355	1080	41	2171
TR833	Çorum	129691	1335	50	2835
TR834	Amasya	80270	862	38	1833
TR901	Trabzon	117654	1183	44	2267
TR902	Ordu	88692	1146	34	2269
TR903	Giresun	57847	715	38	1355
TR904	Rize	50290	530	19	1003
TR905	Artvin	25637	309	13	617
TR906	Gümüşhane	15946	343	18	780
TRA11	Erzurum	87194	1292	65	2823
TRA12	Erzincan	42382	642	28	1424
TRA13	Bayburt	9893	165	5	356
TRA21	Ağrı	27801	479	31	1125
TRA22	Kars	33382	373	11	947
TRA23	Iğdır	20235	296	16	554
TRA24	Ardahan	11896	160	7	341
TRB11	Malatya	113874	1367	44	2871
TRB12	Elazığ	80993	938	33	2002
TRB13	Bingöl	11932	343	6	871
TRB14	Tunceli	5853	132	6	273
TRB21	Van	66605	709	50	1560
TRB22	Muş	23507	334	16	798
TRB23	Bitlis	16421	394	23	935
TRB24	Hakkari	9406	191	15	546
TRC11	Gaziantep	349139	2694	60	4883
TRC12	Adıyaman	70143	767	17	1591
TRC13	Kilis	30881	217	3	339
TRC21	Şanlıurfa	213782	1792	101	3795
TRC22	Diyarbakır	105149	1658	74	3681
TRC31	Mardin	56442	747	34	1578
TRC32	Batman	38176	415	6	802
TRC33	Şırnak	27302	326	9	635
TRC34	Siirt	14690	254	14	554
TOPLAM		**16089528**	**131845**	**3835**	**238074**

Ek1-b 2012 Yılı TUİK Trafik Kaza İstatistikleri Karayolu Verileri

SEHIR KODU	SEHIR ADI	MOTORLU KARA TAŞITI SAYISI	ÖLÜMLÜ YARALANMALI TRAFİK KAZA SAYILARI	ÖLÜ SAYILARI	YARALI SAYILARI
TR100	İstanbul	3065465	15082	247	22772
TR211	Tekirdağ	187665	1651	55	3020
TR212	Edirne	122491	847	30	1531
TR213	Kırklareli	94993	767	33	1389
TR221	Balıkesir	361079	3177	70	5396
TR222	Çanakkale	167198	1279	41	2188
TR310	İzmir	1062946	9358	158	13792
TR321	Aydın	329011	2705	67	4522
TR322	Denizli	303011	2935	53	4998
TR323	Muğla	363762	3535	70	5618
TR331	Manisa	458273	3811	100	6517
TR332	Afyonkarahisar	170684	1733	71	3458
TR333	Kütahya	167741	1285	41	2378
TR334	Uşak	102937	940	30	1818
TR411	Bursa	607585	5162	79	8221
TR412	Eskişehir	209910	2019	58	3432
TR413	Bilecik	49450	543	10	977
TR421	Kocaeli	276210	3227	74	5730
TR422	Sakarya	198851	2322	33	3963
TR423	Düzce	76994	1002	31	1927
TR424	Bolu	86284	835	43	1854
TR425	Yalova	40954	562	9	839
TR510	Ankara	1436349	11772	205	19466
TR521	Konya	556391	6297	166	10767
TR522	Karaman	73676	758	17	1342
TR611	Antalya	792595	7168	157	11138
TR612	Isparta	137719	1267	32	2240
TR613	Burdur	107084	837	38	1557
TR621	Adana	508751	4573	78	7609
TR622	Mersin	460568	4775	91	7650
TR631	Hatay	373274	2800	45	4804
TR632	Kahramanmaraş	165166	2146	31	3909
TR633	Osmaniye	120295	1432	31	2503
TR711	Kırıkkale	53894	835	19	1782
TR712	Aksaray	86424	1096	26	2268
TR713	Niğde	76886	731	29	1387
TR714	Nevşehir	87385	886	19	1703
TR715	Kırşehir	48717	590	22	1224
TR721	Kayseri	278029	3943	54	6657
TR722	Sivas	116696	1494	40	3188
TR723	Yozgat	82793	929	37	2267
TR811	Zonguldak	120911	853	20	1691
TR812	Karabük	50765	493	26	1106

TR813	Bartın	38471	435	8	829
TR821	Kastamonu	99626	794	35	1620
TR822	Çankırı	37413	627	41	1539
TR823	Sinop	45098	470	21	925
TR831	Samsun	257765	2805	91	5148
TR832	Tokat	134763	1300	46	2475
TR833	Çorum	137183	1406	42	2859
TR834	Amasya	85351	994	44	2008
TR901	Trabzon	127663	1437	37	2772
TR902	Ordu	96193	1327	51	2537
TR903	Giresun	62008	819	27	1552
TR904	Rize	54942	623	17	1203
TR905	Artvin	27359	289	16	561
TR906	Gümüşhane	17137	342	30	853
TRA11	Erzurum	93109	1337	56	2947
TRA12	Erzincan	45128	699	12	1430
TRA13	Bayburt	10649	193	3	334
TRA21	Ağrı	29514	603	41	1295
TRA22	Kars	35634	402	13	931
TRA23	Iğdır	21729	256	9	416
TRA24	Ardahan	13116	145	8	331
TRB11	Malatya	121542	1388	46	2567
TRB12	Elazığ	86800	1123	40	1984
TRB13	Bingöl	12778	367	14	871
TRB14	Tunceli	6357	141	6	277
TRB21	Van	71081	1052	37	2363
TRB22	Muş	25360	336	13	751
TRB23	Bitlis	17674	358	30	848
TRB24	Hakkari	9560	197	13	412
TRC11	Gaziantep	378144	3355	74	5830
TRC12	Adıyaman	75849	875	20	1827
TRC13	Kilis	33782	374	4	587
TRC21	Şanlıurfa	228449	2072	81	4305
TRC22	Diyarbakır	111074	1966	55	4000
TRC31	Mardin	62043	860	32	1756
TRC32	Batman	40661	585	14	1138
TRC33	Şırnak	28772	409	29	712
TRC34	Siirt	15774	339	8	688
TOPLAM		**17033413**	**153552**	**3750**	**268079**

Ek1-c 2013 Yılı TUİK Trafik Kaza İstatistikleri Karayolu Verileri

SEHIR KODU	SEHIR ADI	MOTORLU KARA TAŞITI SAYISI	ÖLÜMLÜ YARALANMALI TRAFİK KAZA SAYILARI	ÖLÜ SAYILARI	YARALI SAYILARI
TR100	İstanbul	3230908	15224	260	22460
TR211	Tekirdağ	202487	1720	48	3118
TR212	Edirne	128568	756	34	1240
TR213	Kırklareli	101016	763	23	1353
TR221	Balıkesir	377012	3458	103	5882
TR222	Çanakkale	174991	1441	28	2509
TR310	İzmir	1103176	9687	134	13967
TR321	Aydın	343552	2831	56	4575
TR322	Denizli	318461	2945	53	4734
TR323	Muğla	384148	3806	70	5895
TR331	Manisa	479962	3961	109	6720
TR332	Afyonkarahisar	178251	1777	92	3610
TR333	Kütahya	174431	1295	60	2412
TR334	Uşak	108680	1077	27	2023
TR411	Bursa	642836	5524	90	8725
TR412	Eskişehir	221240	1949	52	3361
TR413	Bilecik	52669	547	22	1061
TR421	Kocaeli	294640	3276	64	5640
TR422	Sakarya	211628	2495	45	4198
TR423	Düzce	82962	1042	32	1925
TR424	Bolu	91298	879	35	1831
TR425	Yalova	44793	606	9	957
TR510	Ankara	1509632	11883	160	19327
TR521	Konya	581064	6450	135	10947
TR522	Karaman	76294	790	8	1348
TR611	Antalya	833281	7078	131	10956
TR612	Isparta	143157	1472	27	2482
TR613	Burdur	111026	905	22	1602
TR621	Adana	535149	4859	87	7751
TR622	Mersin	484893	5394	99	8249
TR631	Hatay	393217	3298	67	5368
TR632	Kahramanmaraş	176022	2187	50	3955
TR633	Osmaniye	128188	1653	21	2540
TR711	Kırıkkale	58117	898	23	1812
TR712	Aksaray	92336	1208	30	2211
TR713	Niğde	80069	860	37	1532
TR714	Nevşehir	92505	937	28	1670
TR715	Kırşehir	52449	612	13	1312
TR721	Kayseri	293922	4146	58	7084
TR722	Sivas	124213	1350	48	2788
TR723	Yozgat	87845	1014	35	2360
TR811	Zonguldak	126330	970	23	1865
TR812	Karabük	53833	476	11	915

TR813	Bartın	40909	479	10	808
TR821	Kastamonu	104872	850	58	1694
TR822	Çankırı	40210	607	26	1487
TR823	Sinop	48072	464	18	857
TR831	Samsun	271041	2808	79	5007
TR832	Tokat	141929	1447	46	2615
TR833	Çorum	143580	1556	40	3211
TR834	Amasya	90383	950	26	1912
TR901	Trabzon	137560	1461	30	2700
TR902	Ordu	102915	1368	45	2538
TR903	Giresun	66907	817	37	1593
TR904	Rize	58854	754	16	1282
TR905	Artvin	29199	335	10	616
TR906	Gümüşhane	18378	363	16	852
TRA11	Erzurum	98295	1337	31	2808
TRA12	Erzincan	47107	652	19	1333
TRA13	Bayburt	11449	175	6	362
TRA21	Ağrı	30606	685	32	1429
TRA22	Kars	37858	364	6	791
TRA23	Iğdır	23608	329	4	551
TRA24	Ardahan	14498	153	6	375
TRB11	Malatya	128950	1415	46	2745
TRB12	Elazığ	92456	1279	30	2250
TRB13	Bingöl	13437	423	18	905
TRB14	Tunceli	6914	168	5	302
TRB21	Van	73425	1377	50	2687
TRB22	Muş	27221	387	28	908
TRB23	Bitlis	18641	412	10	952
TRB24	Hakkari	9467	221	12	440
TRC11	Gaziantep	405168	3607	79	6241
TRC12	Adıyaman	81729	1026	30	2014
TRC13	Kilis	36439	441	6	659
TRC21	Şanlıurfa	237559	2212	83	4198
TRC22	Diyarbakır	114720	2059	44	4345
TRC31	Mardin	66292	1122	49	2116
TRC32	Batman	41548	655	31	1158
TRC33	Şırnak	28991	576	34	1016
TRC34	Siirt	17009	473	10	802
TOPLAM		**17939447**	**161306**	**3685**	**274829**

Ek2-a Standartlaştırılmış 2011 Yılı TUİK Trafik Kaza İstatistikleri Karayolu Verileri

SEHIR KODU	SEHIR ADI	MOTORLU KARA TAŞITI SAYISI	ÖLÜMLÜ YARALANMALI TRAFİK KAZA SAYILARI	ÖLÜ SAYILARI	YARALI SAYILARI
TR100	İstanbul	1,000	1,000	1,000	1,000
TR211	Tekirdağ	-0,887	-0,812	-0,399	-0,779
TR212	Edirne	-0,925	-0,915	-0,803	-0,902
TR213	Kırklareli	-0,943	-0,917	-0,830	-0,904
TR221	Balıkesir	-0,770	-0,609	-0,067	-0,556
TR222	Çanakkale	-0,896	-0,852	-0,623	-0,827
TR310	İzmir	-0,306	0,111	0,202	0,076
TR321	Aydın	-0,792	-0,711	-0,498	-0,673
TR322	Denizli	-0,808	-0,693	-0,283	-0,650
TR323	Muğla	-0,770	-0,597	-0,345	-0,576
TR331	Manisa	-0,709	-0,563	-0,013	-0,489
TR332	Afyonkarahisar	-0,893	-0,780	-0,327	-0,710
TR333	Kütahya	-0,896	-0,855	-0,722	-0,822
TR334	Uşak	-0,938	-0,897	-0,794	-0,859
TR411	Bursa	-0,610	-0,359	-0,121	-0,309
TR412	Eskişehir	-0,868	-0,722	-0,399	-0,696
TR413	Bilecik	-0,972	-0,950	-0,865	-0,935
TR421	Kocaeli	-0,827	-0,603	-0,408	-0,529
TR422	Sakarya	-0,876	-0,750	-0,444	-0,711
TR423	Düzce	-0,955	-0,889	-0,848	-0,862
TR424	Bolu	-0,948	-0,894	-0,758	-0,849
TR425	Yalova	-0,978	-0,964	-0,892	-0,965
TR510	Ankara	-0,068	0,481	0,444	0,608
TR521	Konya	-0,646	-0,318	0,336	-0,194
TR522	Karaman	-0,956	-0,934	-0,865	-0,922
TR611	Antalya	-0,492	-0,141	0,354	-0,131
TR612	Isparta	-0,914	-0,865	-0,695	-0,834
TR613	Burdur	-0,934	-0,906	-0,803	-0,889
TR621	Adana	-0,675	-0,405	-0,175	-0,369
TR622	Mersin	-0,707	-0,431	-0,058	-0,408
TR631	Hatay	-0,765	-0,698	-0,578	-0,660
TR632	Kahramanmaraş	-0,899	-0,772	-0,614	-0,699
TR633	Osmaniye	-0,927	-0,869	-0,794	-0,859
TR711	Kırıkkale	-0,970	-0,906	-0,695	-0,867
TR712	Aksaray	-0,949	-0,893	-0,839	-0,855
TR713	Niğde	-0,955	-0,923	-0,641	-0,909
TR714	Nevşehir	-0,948	-0,918	-0,830	-0,893
TR715	Kırşehir	-0,973	-0,955	-0,955	-0,932
TR721	Kayseri	-0,825	-0,597	-0,408	-0,507
TR722	Sivas	-0,929	-0,826	-0,758	-0,741
TR723	Yozgat	-0,950	-0,884	-0,767	-0,816
TR811	Zonguldak	-0,925	-0,889	-0,785	-0,856
TR812	Karabük	-0,971	-0,950	-0,821	-0,937

TR813	Bartın	-0,979	-0,962	-0,982	-0,950
TR821	Kastamonu	-0,939	-0,902	-0,614	-0,866
TR822	Çankırı	-0,980	-0,937	-0,713	-0,897
TR823	Sinop	-0,975	-0,961	-0,830	-0,957
TR831	Samsun	-0,837	-0,686	-0,265	-0,604
TR832	Tokat	-0,917	-0,862	-0,659	-0,820
TR833	Çorum	-0,915	-0,825	-0,578	-0,757
TR834	Amasya	-0,949	-0,894	-0,686	-0,852
TR901	Trabzon	-0,923	-0,847	-0,632	-0,811
TR902	Ordu	-0,943	-0,853	-0,722	-0,811
TR903	Giresun	-0,964	-0,915	-0,686	-0,898
TR904	Rize	-0,970	-0,942	-0,857	-0,931
TR905	Artvin	-0,986	-0,974	-0,910	-0,967
TR906	Gümüşhane	-0,993	-0,969	-0,865	-0,952
TRA11	Erzurum	-0,944	-0,831	-0,444	-0,758
TRA12	Erzincan	-0,975	-0,926	-0,776	-0,891
TRA13	Bayburt	-0,997	-0,995	-0,982	-0,992
TRA21	Ağrı	-0,985	-0,950	-0,749	-0,919
TRA22	Kars	-0,981	-0,965	-0,928	-0,936
TRA23	Iğdır	-0,990	-0,976	-0,883	-0,973
TRA24	Ardahan	-0,996	-0,996	-0,964	-0,994
TRB11	Malatya	-0,926	-0,820	-0,632	-0,754
TRB12	Elazığ	-0,949	-0,883	-0,731	-0,836
TRB13	Bingöl	-0,996	-0,969	-0,973	-0,943
TRB14	Tunceli	-1,000	-1,000	-0,973	-1,000
TRB21	Van	-0,958	-0,916	-0,578	-0,878
TRB22	Muş	-0,988	-0,971	-0,883	-0,950
TRB23	Bitlis	-0,993	-0,962	-0,821	-0,937
TRB24	Hakkari	-0,998	-0,991	-0,892	-0,974
TRC11	Gaziantep	-0,765	-0,627	-0,489	-0,563
TRC12	Adıyaman	-0,956	-0,908	-0,874	-0,875
TRC13	Kilis	-0,983	-0,988	-1,000	-0,994
TRC21	Şanlıurfa	-0,858	-0,759	-0,121	-0,666
TRC22	Diyarbakır	-0,932	-0,778	-0,363	-0,677
TRC31	Mardin	-0,965	-0,911	-0,722	-0,876
TRC32	Batman	-0,978	-0,959	-0,973	-0,950
TRC33	Şırnak	-0,985	-0,972	-0,946	-0,966
TRC34	Siirt	-0,994	-0,982	-0,901	-0,973

Ek2-b Standartlaştırılmış 2012 Yılı TUİK Trafik Kaza İstatistikleri Karayolu Verileri

SEHIR KODU	SEHIR ADI	MOTORLU KARA TAŞITI SAYISI	ÖLÜMLÜ YARALANMALI TRAFİK KAZA SAYILARI	ÖLÜ SAYILARI	YARALI SAYILARI
TR100	İstanbul	1,000	1,000	1,000	1,000
TR211	Tekirdağ	-0,881	-0,798	-0,574	-0,756
TR212	Edirne	-0,924	-0,905	-0,779	-0,889
TR213	Kırklareli	-0,942	-0,916	-0,754	-0,901
TR221	Balıkesir	-0,768	-0,594	-0,451	-0,545
TR222	Çanakkale	-0,895	-0,848	-0,689	-0,830
TR310	İzmir	-0,309	0,234	0,270	0,202
TR321	Aydın	-0,789	-0,657	-0,475	-0,623
TR322	Denizli	-0,806	-0,626	-0,590	-0,580
TR323	Muğla	-0,766	-0,546	-0,451	-0,525
TR331	Manisa	-0,705	-0,509	-0,205	-0,445
TR332	Afyonkarahisar	-0,893	-0,787	-0,443	-0,717
TR333	Kütahya	-0,894	-0,847	-0,689	-0,813
TR334	Uşak	-0,937	-0,893	-0,779	-0,863
TR411	Bursa	-0,607	-0,328	-0,377	-0,294
TR412	Eskişehir	-0,867	-0,749	-0,549	-0,719
TR413	Bilecik	-0,972	-0,946	-0,943	-0,938
TR421	Kocaeli	-0,824	-0,587	-0,418	-0,515
TR422	Sakarya	-0,874	-0,708	-0,754	-0,672
TR423	Düzce	-0,954	-0,885	-0,770	-0,853
TR424	Bolu	-0,948	-0,907	-0,672	-0,860
TR425	Yalova	-0,977	-0,944	-0,951	-0,950
TR510	Ankara	-0,065	0,557	0,656	0,706
TR521	Konya	-0,640	-0,176	0,336	-0,067
TR522	Karaman	-0,956	-0,917	-0,885	-0,905
TR611	Antalya	-0,486	-0,059	0,262	-0,034
TR612	Isparta	-0,914	-0,849	-0,762	-0,825
TR613	Burdur	-0,934	-0,907	-0,713	-0,886
TR621	Adana	-0,672	-0,407	-0,385	-0,348
TR622	Mersin	-0,703	-0,380	-0,279	-0,344
TR631	Hatay	-0,760	-0,644	-0,656	-0,598
TR632	Kahramanmaraş	-0,896	-0,732	-0,770	-0,677
TR633	Osmaniye	-0,926	-0,827	-0,770	-0,802
TR711	Kırıkkale	-0,969	-0,907	-0,869	-0,866
TR712	Aksaray	-0,948	-0,872	-0,811	-0,823
TR713	Niğde	-0,954	-0,921	-0,787	-0,901
TR714	Nevşehir	-0,947	-0,900	-0,869	-0,873
TR715	Kırşehir	-0,972	-0,940	-0,844	-0,916
TR721	Kayseri	-0,822	-0,491	-0,582	-0,433
TR722	Sivas	-0,928	-0,819	-0,697	-0,741
TR723	Yozgat	-0,950	-0,895	-0,721	-0,823
TR811	Zonguldak	-0,925	-0,905	-0,861	-0,874
TR812	Karabük	-0,971	-0,953	-0,811	-0,926

TR813	Bartın	-0,979	-0,961	-0,959	-0,951
TR821	Kastamonu	-0,939	-0,913	-0,738	-0,881
TR822	Çankırı	-0,980	-0,935	-0,689	-0,888
TR823	Sinop	-0,975	-0,956	-0,852	-0,942
TR831	Samsun	-0,836	-0,643	-0,279	-0,567
TR832	Tokat	-0,916	-0,845	-0,648	-0,805
TR833	Çorum	-0,914	-0,831	-0,680	-0,770
TR834	Amasya	-0,948	-0,886	-0,664	-0,846
TR901	Trabzon	-0,921	-0,827	-0,721	-0,778
TR902	Ordu	-0,941	-0,841	-0,607	-0,799
TR903	Giresun	-0,964	-0,909	-0,803	-0,887
TR904	Rize	-0,968	-0,935	-0,885	-0,918
TR905	Artvin	-0,986	-0,980	-0,893	-0,975
TR906	Gümüşhane	-0,993	-0,973	-0,779	-0,949
TRA11	Erzurum	-0,943	-0,840	-0,566	-0,763
TRA12	Erzincan	-0,975	-0,925	-0,926	-0,897
TRA13	Bayburt	-0,997	-0,993	-1,000	-0,995
TRA21	Ağrı	-0,985	-0,938	-0,689	-0,909
TRA22	Kars	-0,981	-0,965	-0,918	-0,942
TRA23	Iğdır	-0,990	-0,985	-0,951	-0,988
TRA24	Ardahan	-0,996	-0,999	-0,959	-0,995
TRB11	Malatya	-0,925	-0,833	-0,648	-0,796
TRB12	Elazığ	-0,947	-0,869	-0,697	-0,848
TRB13	Bingöl	-0,996	-0,970	-0,910	-0,947
TRB14	Tunceli	-1,000	-1,000	-0,975	-1,000
TRB21	Van	-0,958	-0,878	-0,721	-0,815
TRB22	Muş	-0,988	-0,974	-0,918	-0,958
TRB23	Bitlis	-0,993	-0,971	-0,779	-0,949
TRB24	Hakkari	-0,998	-0,993	-0,918	-0,988
TRC11	Gaziantep	-0,757	-0,570	-0,418	-0,506
TRC12	Adıyaman	-0,955	-0,902	-0,861	-0,862
TRC13	Kilis	-0,982	-0,969	-0,992	-0,972
TRC21	Şanlıurfa	-0,855	-0,742	-0,361	-0,642
TRC22	Diyarbakır	-0,932	-0,756	-0,574	-0,669
TRC31	Mardin	-0,964	-0,904	-0,762	-0,869
TRC32	Batman	-0,978	-0,941	-0,910	-0,923
TRC33	Şırnak	-0,985	-0,964	-0,787	-0,961
TRC34	Siirt	-0,994	-0,973	-0,959	-0,963

Ek2-c Standartlaştırılmış 2013 Yılı TUİK Trafik Kaza İstatistikleri Karayolu Verileri

SEHIR KODU	SEHIR ADI	MOTORLU KARA TAŞITI SAYISI	ÖLÜMLÜ YARALANMALI TRAFİK KAZA SAYILARI	ÖLÜ SAYILARI	YARALI SAYILARI
TR100	İstanbul	1,000	1,000	1,000	1,000
TR211	Tekirdağ	-0,879	-0,792	-0,656	-0,746
TR212	Edirne	-0,925	-0,920	-0,766	-0,915
TR213	Kırklareli	-0,942	-0,919	-0,852	-0,905
TR221	Balıkesir	-0,770	-0,561	-0,227	-0,496
TR222	Çanakkale	-0,896	-0,829	-0,813	-0,801
TR310	İzmir	-0,320	0,265	0,016	0,233
TR321	Aydın	-0,791	-0,645	-0,594	-0,614
TR322	Denizli	-0,807	-0,629	-0,617	-0,600
TR323	Muğla	-0,766	-0,515	-0,484	-0,495
TR331	Manisa	-0,707	-0,495	-0,180	-0,421
TR332	Afyonkarahisar	-0,894	-0,784	-0,313	-0,701
TR333	Kütahya	-0,896	-0,848	-0,563	-0,810
TR334	Uşak	-0,937	-0,877	-0,820	-0,845
TR411	Bursa	-0,606	-0,287	-0,328	-0,240
TR412	Eskişehir	-0,867	-0,762	-0,625	-0,724
TR413	Bilecik	-0,972	-0,948	-0,859	-0,931
TR421	Kocaeli	-0,822	-0,586	-0,531	-0,518
TR422	Sakarya	-0,873	-0,689	-0,680	-0,648
TR423	Düzce	-0,953	-0,882	-0,781	-0,854
TR424	Bolu	-0,948	-0,904	-0,758	-0,862
TR425	Yalova	-0,977	-0,940	-0,961	-0,941
TR510	Ankara	-0,068	0,557	0,219	0,717
TR521	Konya	-0,644	-0,164	0,023	-0,039
TR522	Karaman	-0,957	-0,915	-0,969	-0,906
TR611	Antalya	-0,487	-0,081	-0,008	-0,038
TR612	Isparta	-0,915	-0,825	-0,820	-0,803
TR613	Burdur	-0,935	-0,900	-0,859	-0,883
TR621	Adana	-0,672	-0,375	-0,352	-0,328
TR622	Mersin	-0,703	-0,304	-0,258	-0,283
TR631	Hatay	-0,760	-0,583	-0,508	-0,543
TR632	K.Maraş	-0,895	-0,730	-0,641	-0,670
TR633	Osmaniye	-0,925	-0,801	-0,867	-0,798
TR711	Kırıkkale	-0,968	-0,901	-0,852	-0,864
TR712	Aksaray	-0,947	-0,860	-0,797	-0,828
TR713	Niğde	-0,955	-0,906	-0,742	-0,889
TR714	Nevşehir	-0,947	-0,896	-0,813	-0,877
TR715	Kırşehir	-0,972	-0,939	-0,930	-0,909
TR721	Kayseri	-0,822	-0,470	-0,578	-0,388
TR722	Sivas	-0,927	-0,841	-0,656	-0,776
TR723	Yozgat	-0,950	-0,886	-0,758	-0,814
TR811	Zonguldak	-0,926	-0,892	-0,852	-0,859
TR812	Karabük	-0,971	-0,957	-0,945	-0,945

TR813	Bartın	-0,979	-0,957	-0,953	-0,954
TR821	Kastamonu	-0,939	-0,908	-0,578	-0,874
TR822	Çankırı	-0,979	-0,940	-0,828	-0,893
TR823	Sinop	-0,974	-0,959	-0,891	-0,950
TR831	Samsun	-0,836	-0,648	-0,414	-0,575
TR832	Tokat	-0,916	-0,828	-0,672	-0,791
TR833	Çorum	-0,915	-0,814	-0,719	-0,737
TR834	Amasya	-0,948	-0,894	-0,828	-0,855
TR901	Trabzon	-0,919	-0,826	-0,797	-0,784
TR902	Ordu	-0,940	-0,839	-0,680	-0,798
TR903	Giresun	-0,963	-0,912	-0,742	-0,883
TR904	Rize	-0,968	-0,920	-0,906	-0,912
TR905	Artvin	-0,986	-0,976	-0,953	-0,972
TR906	Gümüşhane	-0,993	-0,972	-0,906	-0,950
TRA11	Erzurum	-0,943	-0,843	-0,789	-0,774
TRA12	Erzincan	-0,975	-0,934	-0,883	-0,907
TRA13	Bayburt	-0,997	-0,997	-0,984	-0,995
TRA21	Ağrı	-0,985	-0,929	-0,781	-0,898
TRA22	Kars	-0,981	-0,972	-0,984	-0,956
TRA23	Iğdır	-0,990	-0,977	-1,000	-0,978
TRA24	Ardahan	-0,995	-1,000	-0,984	-0,993
TRB11	Malatya	-0,924	-0,833	-0,672	-0,779
TRB12	Elazığ	-0,947	-0,851	-0,797	-0,824
TRB13	Bingöl	-0,996	-0,964	-0,891	-0,946
TRB14	Tunceli	-1,000	-0,998	-0,992	-1,000
TRB21	Van	-0,959	-0,838	-0,641	-0,785
TRB22	Muş	-0,987	-0,969	-0,813	-0,945
TRB23	Bitlis	-0,993	-0,966	-0,953	-0,941
TRB24	Hakkari	-0,998	-0,991	-0,938	-0,988
TRC11	Gaziantep	-0,753	-0,542	-0,414	-0,464
TRC12	Adıyaman	-0,954	-0,884	-0,797	-0,845
TRC13	Kilis	-0,982	-0,962	-0,984	-0,968
TRC21	Ş.Urfa	-0,857	-0,727	-0,383	-0,648
TRC22	Diyarbakır	-0,933	-0,747	-0,688	-0,635
TRC31	Mardin	-0,963	-0,871	-0,648	-0,836
TRC32	Batman	-0,979	-0,933	-0,789	-0,923
TRC33	Şırnak	-0,986	-0,944	-0,766	-0,936
TRC34	Siirt	-0,994	-0,958	-0,953	-0,955

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı	**:**	Burak ÇAĞLAR
Uyruğu	**:**	T.C.
Doğum Yeri ve Tarihi	**:**	Çorum - 10.09.1985
Telefon	**:**	530 274 13 45
Faks	**:**	
e-mail	**:**	bcaglar19@hotmail.com

EĞİTİM

Derece		**Adı, İlçe, İl**	**BitirmeYılı**
Lise	:	Çorum Atatük Lisesi, Merkez, Çorum	2003
Üniversite	:	Selçuk Üniversitesi, Selçuklu, Konya	2008

İŞ DENEYİMLERİ

Yıl	**Kurum**	**Görevi**
2012-	ÇORUM İL ÖZEL İDARESİ	Harita Mühendisi
2011-2012	1902 ÇORUM LİHKAB	Harita Mühendisi
2010-2011	MAPA-LIMAK-MNG J.V. / S. Arabistan	Harita Mühendisi
2008-2009	AZIZ - MAPA CONSORTIUM / S. Arabistan	Harita Mühendisi

YAYINLAR

1. Selvi, H.Z., Çağlar, B. (2016) Using K-Means and K-Medoids Methods for Multivariate Mapping, International Journal of Applied Mathematics, Electronics and Computers, 4, 342-345.
2. Selvi, H.Z., Çağlar, B. (2016) Using Cluster Analysis Method for Multivariate Mapping, 2nd International Conference on Engineering and Natural Sciences (ICENS 2016), 69-75.
3. Selvi, H.Z., Çağlar, B. (2017) Çok Değişkenli Haritalama İçin Kümeleme Yöntemlerinin Kullanılması, Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi, Cilt 6, Sayı 2, 415-429.