ÇOK BOYUTLU ÖLÇEKLEME VE K-ORTALAMALAR KÜMELEME ANALİZİ İLE BİR GÖRSEL VERİ MADENCİLİĞİ UYGULAMASI

A VISUAL DATA MINING APPLICATION WITH MULTIDIMENSIONAL SCALING AND K-MEANS CLUSTERING

Muhammet ATALAY*
* Dr. Öğr. Üyesi, Kırklareli Üniversitesi İİBF İşletme Bölümü, muhammetatalay@gmail.com, ORCID:0000-0003-3960-500X

MAKALE BİLGİSİ

Gönderilme Tarihi
05.03.2022
Revizyon Tarihi
26.03.2022
Kabul Tarihi
29.03.2022
Makale Kategorisi
Araştırma Makalesi

JEL Kodları
C38
C39
H55

ÖZ

Görsel veri madenciliği, verileri veya analiz sonucunda elde edilen bulguları görselleştirerek örtük ve faydalı bilgileri keşfetmeye yarar. Bu çalışmada; 2020 yılı aktif sigortalı sayılarına ait veriler yardımıyla, Türkiye'deki illerin benzerlik ve farklılıklarının ortaya çıkarılması amaçlanmıştır. Yöntem olarak istatistiksel veri analizi ve veri madenciliği tekniklerinden çok boyutlu ölçekleme ve kümeleme analizleri bulguları görselleştirilerek kullanılmıştır. Tüm analizler R programlama dili kullanılarak yapılmıştır. Çok boyutlu ölçeklemede uyum iyiliği değerleri incelenmiştir. Kümeleme analizinde optimal küme sayısını tespit etmek için içsel kümeleme performansı indeksleri karşılaştırılmıştır. Elde edilen sonuçlara göre, iki boyutlu uzayda elde edilen harita, gerçek uzaklıklarla karşılaştırıldığında iyi derecede uyum göstermektedir. En iyi kümeleme için içsel indekslerin çoğu küme sayısının iki olması gerektiğini söylemektedir. Buna göre çok boyutlu ölçekleme ile elde edilen iki boyutlu dağılımda iller iki kümeye ayrışmaktadırlar. 4/a ve 4/b sigortalılarda dağılım dengeli iken 4/c sigortalılarda çok az sayıda ilin diğerlerinden ayrıştığı belirlenmiştir.

Anahtar Kelimeler: Çok Boyutlu Ölçekleme, K-Ortalamalar Kümeleme, İstatistiksel Analiz, Veri Görselleştirme, Görsel Veri Madenciliği

ARTICLE INFO

Received
05.03.2022
Revized
26.03.2022
Accepted
29.03.2022
Article Classification:
Research Article

JEL Codes
C38
C39
H55

ABSTRACT

Visual data mining serves to discover implicit and useful information by visualizing the data or the findings obtained as a result of the analysis. In this study; It is aimed to reveal the similarities and differences of the provinces in Turkey with the help of data on the number of active insured persons in 2020. As a method, multidimensional scaling and clustering analysis findings from statistical data analysis and data mining techniques were used by visualizing them. All analyzes were performed using the R programming language. Goodness-of-fit values were examined in multidimensional scaling. In order to determine the optimal number of clusters in cluster analysis, internal clustering performance indices were compared. According to the results obtained, the map obtained in two-dimensional space shows good agreement when compared to the actual distances. For the best clustering, most of the internal indices say that the number of clusters should be two. Accordingly, in the two-dimensional distribution obtained by multidimensional scaling, the provinces are divided into two clusters. While the distribution is balanced in 4/a and 4/b insureds, it has been determined that very few provinces differ from others in 4/c insureds.

Keywords: Multidimensional Scaling, K-Means Clustering, Statistical Analysis, Data Visualization, Visual Data Mining

Atıf (Citation): Atalay, M. (2022). "Çok Boyutlu Ölçekleme ve K-Ortalamalar Kümeleme Analizi ile Bir Görsel Veri Madenciliği Uygulaması", *Kapanaltı Muhasebe Finans Ekonomi Dergisi*, (1): 31-44

Giriş

Ülkelerin makro göstergelerinin analizinde istatistik ve veri madenciliği tekniklerinin tanımlayıcı ve çıkarımsal olarak kullanımı yaygındır. Bu analizlerde veri ve bulguların görselleştirilerek sunulması, yorumlanması ve anlaşılır hale gelmesi bakımından destekleyici olmaktadır. Görsel veri madenciliği, verileri veya analiz sonucunda elde edilen bulguları görselleştirme tekniklerini kullanarak büyük veri kümelerinden örtük ve faydalı bilgileri keşfetmeye yarar. Bu anlamda görsel veri madenciliği, veri görselleştirme ve veri madenciliğinin entegrasyonu olarak görülebilir. Veri görselleştirme ile veriler, farklı ayrıntı düzeyi ve soyutlama seviyelerinde, farklı nitelik veya boyutlarda görüntülenebilir. Veri madenciliği sonuçları görselleştirilerek elde edilen bulgu veya bilgiler daha anlaşılır hale getirilebilmektedir (Han vd., 2012, s. 602-603).

Veri madenciliği yöntemleri genel olarak dört görevi gerçekleştirmek üzere uygulanmaktadır. Bu görevler; betimleme ve özetleme, sınıflandırma, öngörü ve tahmin, kümelemedir (Oğuzlar, 2005a). Kümeleme analizi, nesneleri incelenen değişkenler bakımından benzerliklerine göre gruplandırarak betimleyici ve özetleyici bilgiler verirken, daha sonra yapılacak çalışmalar için bir ön işlem olma özelliği de taşır (Vatansever & Büyüklü, 2009). Kümeleme analizinin sonuçlarının görselleştirilmesi, yüksek boyutlu veri setlerinde kümeleme yapısının etkileşimli olarak keşfedilmesi için olanak sağlar. Bilgisayar programlarının artan grafik yetenekleriyle, görsel geri bildirimlere dayalı olarak kümeleme sürecine insanı entegre eden daha anlaşılır ve tatmin edici sonuçlar elde edilebilmektedir (Sips, 2009, s. 3355).

Çok boyutlu ölçekleme, nesneler arasındaki benzerlikler ile nesne kümeleri arasındaki uzaklık veya farklılıkların görsel bir temsilidir (Kruskal & Wish, 1978). Çok boyutlu ölçeklemede nesne çiftleri için uzaklık değerleri girdi olarak alınır. Sonrasında yöntemde kullanılan farklı uzaklık ölçütleri olmasına rağmen başlıca amacı görselleştirmedir. Yöntem çok değişkenli verilerle çalıştığından, bulgular çok değişkenli bir veri görselleştirme aracı ile anlaşılabilecektir (Buja vd., 2012). Kümeleme gibi çok boyutlu ölçekleme de veriyi özetleme ve anlaşılır kılma tekniği olmaktadır.

Kümeleme ve çok boyutlu ölçekleme, amaçları yanında kullanılan uzaklık ölçüleri bakımından da yakın yöntemlerdir. Çok boyutlu ölçeklemeyle, çok değişkenli verilerle uzaklıklar yardımıyla daha az boyutlu grafikler elde edilebilmektedir. Ancak bu grafiklerde birimlerin yakınlık ve uzaklıklarının yorumlanması güç olabilmektedir. Kümeleme birimleri uzaklıklarına göre küme merkezleri etrafında gruplandırmaktadır. Bu iki yöntemin birlikte kullanılması her ikisinin bulgularının daha anlaşılır hale gelmesini sağlamaktadır. Yapılan çalışmalarda iki yöntemin sentezlenerek; sporcular için optimal zihinsel sağlık ve performansa katkıda bulunan davranış, düşünce ve duyguların kümelenmesi (Ayala vd., 2022), hayvancılık ekolojik bölgelerinin dağılımlarının betimlenmesi (Velado-Alonso vd., 2022), makroekonomik göstergeler bakımından ülkelerin ekonomik durum ve refah düzeyine göre gruplandırılması (Aliukov & Buleca, 2022), psikiyatrik semptomların kümelenmesi (Fleming vd., t.y.), madde bağımlılığı ile mücadelede kullanılacak kavram haritalarının oluşturulması (Montgomery vd., 2022), illerin işsizlik oranlarına göre gruplandırılması (Almeira & Graciella Juanda, 2021), kamu harcamalarının illere göre dağılımı (Allahverdi vd., 2021) gibi çeşitli konularda son yıllarda kullanıldığı görülmektedir. Öte yandan bu çalışmanın odaklandığı sosyal güvenlik sistemi ve sigortalılar konularında da yöntemlerin birlikte kullanımına rastlanır. Büyük ölçekli afet ve kazalarda neden ve olayların düşük boyutlu temsillerinin çıkarılarak kümelerin ortaya çıkarılması (Lopes & Machado, 2022), sağlık sigortası kapsamında kronik hastalıklar arasındaki benzerlik veya farklılıkların belirlenmesi (Roux, 2008), refah ve sosyal uyum rejimleri tipolojisi oluşturmak için ekonomik özgürlük, eşitlik ve aile dayanışması algılarına göre ülkelerin gruplandırılması (Borsenberger vd., 2016), hayat dışı sigorta şirketlerinin performansının sigorta ve bireysel emeklilik verileri kullanılarak analizi (Karadağ Erdemir & Tatlıdil, 2018), sigorta pazar payının makro ekonomi ve sigortacılık göstergeleri yardımıyla analiz edilerek OECD ülkelerinin benzerliklerinin incelenmesi (Arı & Gülcemal, 2019), sağlık göstergeleri bakımından OECD ülkelerinin benzer ve farklı yönlerinin tespiti (Kırcı Çevik, 2021), Türkiye ve Avrupa ülkelerindeki ölümlü iş kazalarını karşılaştırmak için ülkelerin gruplandırılması (Demirel Top, Yapici & Cetinkaya, 2018) gibi çalışmalar bunlardan bazılarıdır.

Literatür incelemesine göre, Türkiye'deki illerin, bu illerde yaşayan sigortalılara ait veriler kullanılarak çok değişkenli istatistik ve veri madenciliği yöntemleriyle araştırıldığı çalışmaların az sayıda olduğu ve mevcut olanların güncel olmadığı görülmektedir. Bu çalışmada, sigortalı çalışanlar ve bunların bakmakla yükümlü oldukları ile aylık ve gelir alanlara ait en güncel veriler kullanılarak konu çok boyutlu ölçekleme ve kümeleme analizi yardımıyla araştırılacak ve literatüre katkı sağlanmış olacaktır. Ayrıca yöntemlerin görsel veri analiziyle sunulması bu çalışmanın bir diğer farklı yönüdür. Bu eksende çalışmanın amacı, 2020 yılı aktif sigortalı sayılarına ait veriler yardımıyla, Türkiye'deki illerin benzerlik ve farklılıklarının kümeleme ve çok boyutlu ölçekleme yöntemlerinin bulguları görselleştirilerek ortaya çıkarılmasıdır. Araştırmada istatistik ve veri madenciliği yöntemleri veri görselleştirme ile sentezlenerek sunulacaktır.

Sosyal güvenlik sisteminin amacına ulaşabilmesi ve sürdürülebilir olması, sigortalıların sektörel ve mekânsal bazda dağılımlarıyla da ilişkilidir. Nüfusun ve demografik yapının ülkenin yerleşim birimlerindeki dağılımı, sosyal güvence dağılımını da etkileyecektir. Kamu ve özel kesimde istihdamın yapısının çeşitlenmesiyle birlikte sosyal güvence sisteminin önemli bir göstergesi olan sigortalı yapısı da çeşitlenmektedir. Türkiye Cumhuriyeti devleti yasalarıyla bu dağılım belirli kapsamlarla belirginleştirilmiştir. Buna göre temel yapıyı üç kısımdan oluşan bir sigortalılık sistemi oluşturmaktadır.

4/1-a kapsamındaki aktif sigortalılar, 5510 sayılı Sosyal Sigortalar ve Genel Sağlık Sigortası Kanunu'nun 4. maddesinin 1. fıkrasının (a) bendi kapsamına göre, hizmet akdi ile bir veya birden fazla işveren tarafından çalıştırılan sigortalıları ifade etmektedir. Bu sigortalılar; zorunlu sigortalılar, stajyerler, kursiyerler, çıraklar, yurtdışı topluluklar ve diğer sigortalılardan oluşmaktadır. Zorunlu sigortalılar, uzun vade sigorta kolları kapsamında işyerlerince yapılan bildirimleri ifade etmektedir (SGK, 2021). Bu sebeple stajyer, kursiyer, çırak, yurtdışı topluluk ve diğer sigortalılar çalışmada kısa vade sigortalı olarak zikredilmiştir. Türkiye'de en fazla sigortalı çalışan sayısı 4/1-a kapsamındaki aktif sigortalılara aittir.

4/1-b kapsamındaki aktif sigortalılar, 5510 Sayılı Kanun'un 4. maddesine göre; köy ve mahalle muhtarları ile hizmet akdine bağlı olmaksızın kendi adına ve hesabına bağımsız çalışanları içermektedir. Bu çalışanlar; ticarî kazanç veya serbest meslek kazancı nedeniyle gerçek veya basit usulde gelir vergisi mükellefi olanlar, gelir vergisinden muaf olup, esnaf ve sanatkâr siciline kayıtlı olanlar, anonim şirketlerin yönetim kurulu üyesi olan ortakları, sermayesi paylara bölünmüş komandit şirketlerde komandite ortaklar, diğer şirket ve donatma iştiraklerinde ise tüm ortaklar, tarımsal faaliyette bulunanlar ve isteğe bağlı sigortalılardır (SGK, 2021). Bu kapsamdaki aktif sigortalılar öncelikle zorunlu ve isteğe bağlı sigortalılar olarak tasnif edilmektedir. Zorunlu sigortalılar ise tarım ve tarım dışı (muhtarlar ve diğer zorunlu) sigortalılardan oluşmaktadır.

4/1-c kapsamındaki aktif sigortalılar, kamu idarelerinde; 4/1-a veya 4/1-b kapsamında olmayan, kadro ve pozisyonlarda sürekli ya da sözleşmeli olarak çalışıp ilgili kanunlarında (a) bendi kapsamına girenler gibi sigortalı olması öngörülmemiş olanlar ile 657 sayılı Devlet Memurları Kanunu'nun 86. maddesi uyarınca açıktan vekil atananları içermektedir. Bu kapsamda da zorunlu ve isteğe bağlı sigortalılık söz konusu olup; zorunlu sigortalılar, diğer sigortalılar hariç uzun vadeli sigortalı kolları kapsamındaki bildirimleri ifade etmektedir (SGK, 2021). Bu kapsamdaki sigortalı çalışan sayısı, 4/1-a kapsamındaki sigortalı sayısından sonra Türkiye'de ikinci sıradadır.

Bu maddeler kapsamında çalışmayan ancak gelir ve aylık alan sigortalılar da bulunmaktadır. Gelir, iş kazası veya meslek hastalığı halinde sigortalıya veya sigortalının ölümü halinde hak sahiplerine yapılan sürekli ödemeyi ifade etmektedir. Aylık ise malûllük, yaşlılık ve ölüm sigortaları ile vazife malûllüğü halinde yapılan sürekli ödemeyi ifade etmektedir. Çalışan, gelir ya da aylık alan sigortalıların, sigortalı sayılmayan veya isteğe bağlı sigortalı olmayan, kendi sigortalılığı nedeniyle gelir veya aylık bağlanmamış olan bakmakla yükümlü oldukları da sigortalı kapsamında değerlendirilmektedir (SGK, 2021).

Çalışmanın devamında araştırmanın veri seti tanıtılacak ve kullanılan yöntemler açıklanacaktır. Ardından analiz bulguları verilecek ve son olarak bulgular tartışılarak çıkarılan sonuçlar paylaşılacaktır.

1. Gereç ve Yöntem

Veri madenciliği, verilerden anlamlı bilgileri üretme süreci olup, değişkenler arasındaki ilişkilerin, örüntülerin ve kuralların modellenmesi ve keşfedilmesi süreci olarak tanımlanabilir. İstatistik, bilgisayar bilimleri, makine öğrenmesi, veri tabanı yönetimi gibi alanların tekniklerinden bir ya da birkaçı kullanılarak veriden bilgiye erişilir (Albayrak & Koltan Yılmaz, 2009). Çok değişkenli görsel istatistik ve veri madenciliği yöntemlerinden olan çok boyutlu ölçekleme ve kümeleme bu çalışmada kullanılan yöntemlerdir. Tüm analizler R programlama dili kullanılarak RStudio'da gerçekleştirilmiştir (RStudio, 2022; The R Foundation, 2022). Çalışmada kullanılan R paketleri şunlardır: *readxl*, *magrittr*, *dplyr*, *ggpubr*, *stats*, *MASS, smacof, clusterCrit* (Wickham, François, Henry, & Müller, 2020; Wickham & Bryan, 2019; Bache & Wickham, 2020; Kassambara, 2020; Desgraupes, 2018; Leeuw & Mair, 2009; Mair, Groenen, & Leeuw, 2021; Team, 2020).

1.1. Veri Seti

Çalışmanın veri seti, Türkiye Cumhuriyeti Sosyal Güvenlik Kurumu (SGK) tarafından yayımlanan 2020 yılına ait sigortalı ve işyeri istatistiklerinden oluşturulmuştur. Bu verilere SGK İstatistik Yıllıklarından (SGK, 2021) ulaşılmış olup; 4-1/a, 4-1/b ve 4-1/c kapsamındaki aktif sigortalıların sayısı, ilgili ilin toplam aktif sigortalı sayısına oranlanarak, aylık ve gelir alanların sayısı ilgili ilin aylık ve gelir alanlar toplamına, sigortalıların bakmakla yükümlü olduklarının sayısı da yine ilgili ilin bakmakla yükümlü oldukları toplamına oranlanarak kullanılmıştır. Veri setinde kullanılan oranlar Türkiye'deki her bir il için hesaplanmış olup 81 ile ait 15 adet değişken (öznitelik) bulunmaktadır. Bu değişkenler ve hesaplama yöntemleri Tablo 1'de verilmiştir.

Tablo 1: *Veri Setinde Bulunan Değişkenler (Öznitelikler)* (SGK, 2021)

No	Kısaltması	Değişken (Öznitelik)	Hesaplanışı
1	A1	4/1-a Toplam Aktif Sigortalı Oranı	A1/(A1+B1+C1)
2	A2	4/1-a Toplam Kısa Vade Sigortalı Oranı	A2/(A1+B1+C1)
3	A3	4/1-a Toplam Zorunlu Sigortalı Oranı	A3/(A1+B1+C1)
4	A4	4/1-a Toplam Gelir ve Aylık Alan Oranı	A4/(A4+B4+C4)
5	A5	4/1-a Toplam Bakmakla Yükümlü Oranı	A5/(A5+B5+C5)
6	B1	4/1-b Toplam Aktif Sigortalı Oranı	B1/(A1+B1+C1)
7	B2	4/1-b Toplam Zorunlu Sigortalı Oranı	B2/(A1+B1+C1)
8	B3	4/1-b İsteğe Bağlı Sigortalı Oranı	B3/(A1+B1+C1)
9	B4	4/1-b Toplam Gelir ve Aylık Alan Oranı	B4/(A4+B4+C4)
10	B5	4/1-b Toplam Bakmakla Yükümlü Oranı	B5/(A5+B5+C5)
11	C1	4/1-c Toplam Aktif Sigortalı Oranı	C1/(A1+B1+C1)
12	C2	4/1-c Toplam Zorunlu Sigortalı Oranı	C2/(A1+B1+C1)
13	C3	4/1-c Toplam İsteğe Bağlı Sigortalı Oranı	C3/(A1+B1+C1)
14	C4	4/1-c Toplam Gelir ve Aylık Alan Oranı	C4/(A4+B4+C4)
15	C5	4/1-c Toplam Bakmakla Yükümlü Oranı	C5/(A5+B5+C5)

Değişkenlerden ilk 5 tanesi 4/1-a, sonraki 5 tanesi 4/1-b ve son 5 tanesi 4/1-c kapsamındaki aktif sigortalılarla ilgilidir. Bunlar belirlenirken öncelikle ilgili kanun maddelerinde yapılan tasnifler esas alınmıştır. Bu veri seti, 4/1-a, 4/1-b ve 4/1-c kapsamındaki sigortalı sayılarının iller bazında detaylı olarak incelenmesi için kullanılacaktır. Böylelikle veri seti üç ayrı grup olarak kullanılarak analizler yapılacaktır. Elde edilen değişkenler $[0,1)$ aralığında yer aldığından normalizasyona gerek bulunmamaktadır.

1.2. Çok Boyutlu Ölçekleme

Çok boyutlu ölçekleme, aralarındaki benzerlik veya benzersizlikleri haritalandırmak amacıyla birimleri birer nokta olarak grafiklerde göstermek için kullanılan çok değişkenli bir veri analizi tekniğidir. Amaç uzaklıklardan faydalanılarak nesneler arasındaki ilişkilerin ortaya çıkarılması ve görselleştirilmesidir (Wickelmaier, 2003, s. 4). Yöntem verilerin dağılımıyla ilgili bir varsayım gerektirmemektedir (Alpar, 2017, s. 375). Çok boyutlu ölçekleme algoritması girdi verisi olarak, nesne çiftleri arasındaki mesafeleri temsil eden uzaklıklar matrisini alır. Bu matrisin elde edilebildiği durumlarda metrik ölçekleme yapılırken, uzaklık değerlerinin yalnızca sıralama ifade ettiği durumlarda metrik olmayan ölçekleme kullanılır (Gürsakal, 2019, s. 181-183). Metrik uzaklıklar olarak öklid, minkowski, manhatan city-block gibi; metrik olamayan uzaklıklar olarak ise ki-kare ölçüsü ve phi-kare ölçüsü gibi uzaklıklar tercih edilmektedir (Shanti, 2019, s. 328). Bu çalışmada tüm değişkenler oransal olduğu için metrik ölçekleme, metrik uzaklık olarak kareli öklid ölçüsü kullanılmıştır. Metrik ölçeklemede nesneler arası orijinal uzaklıklar ve ölçeklemeyle hesaplanan haritadaki uzaklıklar oranlı ölçeklenmiş olacaktır (Oğuzlar, 2005b).

Çok boyutlu ölçekleme, en uygun çözümü, değişken sayısından daha düşük boyutta bir uzayda verileri temsil etmek için uzaklıklara dönüştürür. Boyut sayısı araştırmacı tarafından önceden belirlenir ve iki boyut seçilirse, iki boyutlu bir dağılım grafiği için birimlerin konumları, orijinal konumlarına çok yakın bir şekilde belirlenir (Kruskal & Wish, 1978, s. 15). Orijinal uzaklıklar ile gösterim uzaklıkları arasındaki uygunluk yani bir çok boyutlu ölçekleme çözümünün uyum iyiliğini değerlendirmek için çoğunlukla Kruskal $stress$ ölçüsü (Kruskal & Carmone, 1967) kullanılır. Küçük $stress$ değeri iyi bir uyum çözümünü gösterirken, yüksek bir değer, kötü bir uyumu gösterir. $Stress$ değerinin uyum iyiliğinde yorumlanması için Tablo 2'deki karşılaştırmalar kullanılabilir (Kruskal, 1964). Buna göre, $stress$ değeri 0,20'den küçük olan çözümlerin kabul edilebilir, 0,05'ten küçük olan çözümlerin görece daha iyi olduğu anlaşılmaktadır.

Tablo 2: *Stress ve uyum iyiliği*

Stress	Uyum İyiliği
0,20	Zayıf
0,10	Orta
0,05	İyi
0,025	Çok İyi
0,00	Mükemmel

Ancak verilerdeki yüksek bir hatanın $stress$ değerini de yükseltebildiği ve yalnızca bu değere bakılarak karar verilmesinin hatalı olabileceği söylenmektedir (Borg & Groenen, 2005, s. 48). İlaveten, elde edilen iki boyutlu haritanın orijinal harita ile uyumunu ölçen bir uyum iyiliği (GOF) değeri hesaplanmaktadır. Bu değer; iki boyutlu haritadaki nokta ile haritadaki orijin arasındaki kareli öklid mesafesinin, gerçek uzaydaki nokta ile orijin arasındaki kareli öklid mesafesine oranıdır. Belirli bir nokta çifti için hesaplanan uyum iyiliği ise bu noktalar arasındaki mesafelerin oranlanmasıyla bulunur ve noktaların uyumunu gösterir. Bu oran 1' e yaklaştıkça yapılan çözüm daha iyi bir uyum iyiliğine sahip olacaktır (Graffelman, 2020). Bu uyum, tahmin edilen mesafeler ile veri noktaları arasında gözlenen mesafeler arasındaki korelasyon katsayısı (R) hesaplanarak da incelenebilir. Elde edilecek R^2 değeri, çözümün açıklama oranını gösterecektir (Hair vd., 2014, s. 497).

1.3. K-Ortalamalar (K-Means) Kümeleme Yöntemi

Kümeleme, ham verileri uygun gruplara ayırma ve veri setinde var olabilecek gizli kalıpları arama yöntemidir. Bu anlamda istatistiki bir çıkarım yapmadan, aynı kümedeki veriler benzer, ancak farklı kümelere ait veriler benzemez olacak şekilde birimleri ayrık kümeler halinde gruplandırma işlemidir (Huang, 1998). Küme sayısı baştan bilinmeksizin aşamalı bir süreç izlenerek ve her bir aşamada bir önceki aşamada oluşan kümeler kullanılarak analiz sonrasında küme sayısına karar verilen yöntemler hiyerarşik kümeleme yöntemleri olarak bilinir. Bu yöntemler birleştirici ve ayırıcı olarak uygulanabilirler. Küme sayısının baştan belirlenerek tek bir optimal kümeleme sonucunun elde edildiği yöntemler ise hiyerarşik olmayan kümeleme yöntemleridir. Bu algoritmalar genellikle tüm noktalar merkezlerle ilişkili olana kadar küme merkezlerini değiştirir. (Inekwe, Maharaj, & Bhattacharya, 2020). K-Ortalamalar algoritması, kümeleme hatasını en aza indiren popüler bir hiyerarşik olmayan kümeleme tekniğidir. Birçok kümeleme uygulamasında kullanılan hızlı ve yinelemeli bir algoritmadır. Başlangıçta rastgele konumlara yerleştirilen küme merkezleriyle başlayan ve kümeleme hatasını en aza indirmek için küme merkezlerini her adımında değiştirerek ilerleyen bir yöntemdir (Likas, Vlassis, & Verbeek, 2003). Diğer kümeleme algoritmalarına kıyasla, basit ve sağlam, oldukça verimli ve çok çeşitli veri türleri için kullanılabilmesi gibi bazı belirgin avantajlara sahiptir. Küresel olmayan kümeler için kötü performans gösterme ve aykırı değerlere duyarlı olma gibi bazı dezavantajları bulunmakla birlikte, bazı uyarlamalarla bunlar giderilebilmektedir (Wu, 2012, s. 8).

K-Ortalamalar yöntemi, veri setinden elde edilen kümelerdeki gözlemlerinin küme merkezine olan küme içi uzaklıklarının kareler toplamını en küçüklemeye dayanır. Bir $X = \{x_1, x_2, x_3, \dots, x_N\}$ veri seti verildiğinde, M-kümeleme problemi bu veri setini, bir kümeleme kriterini optimize edecek şekilde $C_1, C_2, C_3, \dots, C_M$ ayrık alt kümelerine bölmeyi amaçlar. En yaygın kullanılan kümeleme kriteri, $\forall k \in [1, M]$ için C_k alt kümesinin küme merkezi m_k olmak üzere $\forall x_i \in C_k$ değeri ile m_k arasındaki 35klid uzaklıklarının toplamı ile belirlidir. Bu kriter kümeleme hatası olarak adlandırılır ve $m_1, m_2, m_3, \dots, m_M$ küme merkezlerine bağlı olarak hesaplanır:

$$E(m_1, m_2, m_3, \dots, m_M) = \sum_{i=1}^{N} \sum_{k=1}^{M} I(x_i \in C_k) \|x_i - m_k\|^2$$

Burada I üyelik fonksiyonu olup, X doğruysa $I(X) = 1$ dir, aksi takdirde $I(X) = 0$ dır. Eğer m_k küme merkezi, x_i değerine uzaklık olarak en yakınsa x_i değeri o kümeye aittir. Bu durumda k-ortalamalar algoritması x_i değerini k. Kümeye atar ve $I(x_i) = 1$ değerini alır. Aksi halde $I(x_i) = 0$ değerini almaktadır (Likas, Vlassis, & Verbeek, 2003; Žalik, 2008).

K-Ortalamalar ve diğer kümeleme algoritmalarında küme sayısının belirlenmesi kümeleme analizinin temel sorunlarındandır. Optimal küme sayısının belirlenmesi için değerlendirme ölçütü olarak indeksler kullanılmaktadır. Bunlar içsel ve dışsal indeksler olmak üzere iki gruba ayrılmaktadır. Dışsal indekslerde küme elemanlarının hangi kümede olması gerektiğini belirleyen bir etiket değişkenine veya uzman görüşüne ihtiyaç duyulmaktadır. İçsel indeksler ise noktaların kümelere atanması ile ilgili bir önsel bilgiye ihtiyaç duymazlar ve kümeleme analizi sonucunda elde edilen kümelerin küme içi ve/veya kümeler arası uzaklıkları üzerinden hesaplanırlar (Koçoğlu & Esnaf, 2019, s. 257-261). Bu indeksler değerlendirilirken; küme içi uzaklıkları minimum, yani küme içi benzerliklerin en çok, kümeler arası uzaklıkları maksimum, yani kümeler arası benzerlikleri en az yapacak küme sayısı aranır. Başlıca olarak bu indekslere; Dunn, Xie-Beni, Silhouette, Davies-Bouldin, BIC (Bayesian Information Criterion), Calinski-Harabasz, Wemmert-Gancarski, SD, S-Dbw indeksleri örnek verilebilir (Rendón, Abundez, Arizmendi, & Quiroz, 2011; Maulik & Bandyopadhyay, 2002; Agrawal, Garg, & Patel, 2015; Desgraupes, 2017; Liu, Li, Xiong, Gao, & Wu, 2010).

Kümeleme analizi de çok boyutlu ölçekleme gibi birimlerin benzerlik veya benzemezliklerini (uzaklıklar) esas almaktadır. Bu benzerliklerin bir 35klid uzayında görselleştirilmesi ve yapılandırılması, çok boyutlu ölçekleme ve kümeleme yöntemlerinin birlikte kullanımı ile mümkündür (Hofmann & Buhmann, 1995, s. 461). Bu

çalışmada her iki yöntemde uzaklık ölçüsü olarak kareli 36klid uzaklığı kullanılmış ve k-ortalamalar ile elde edilen kümelerin çok boyutlu ölçekleme yardımıyla 2-boyutlu uzayda birlikte haritalandırılması sağlanmıştır.

2. Bulgular

Bu bölümde 4/1-a, 4/1-b ve 4/1-c kapsamında sigortalılarla ilgili elde edilen değişkenler yardımıyla, Türkiye'deki illerin 2020 yılı için bu değişkenler bakımından benzerlik ve farklılıkları incelenecektir. Önce çok boyutlu ölçekleme ile iller bazında 2-boyutlu haritalar elde edilecektir. Sonra iller, oluşturulan uzaklıklar matrisi kullanılarak k-ortalamalar algoritmasıyla kümelenecek, elde edilen kümeler, bu iki boyutlu haritalar üzerinde görselleştirilecektir. Çok boyutlu ölçeklemede uyum iyiliği değerlendirmesi için $stress$, GOF ve R^2 değerleri incelenecektir. K-Ortalamalar kümeleme algoritmasında optimal küme sayısı için, Silhouette, Xie-Beni, Davies-Bouldin, Calinski-Harabasz, Wemmert-Gancarski, S-Dbw ve Dunn indeksleri hesaplanarak karşılaştırılacaktır. Bu indekslerden Silhouette, Calinski-Harabasz, Wemmert-Gancarski ve Dunn indeksleri büyük, Xie-Beni, S-Dbw ve Davies-Bouldin değerleri ise küçük olduğunda kümeleme daha başarılı demektir (Koçoğlu & Esnaf, 2019, s. 280; Agrawal, Garg, & Patel, 2015; Liu, Li, Xiong, Gao, & Wu, 2010).

Türkiye'deki 81 ilde 2020 yılında 4/1-a kapsamındaki aktif sigortalıların sayısıyla elde edilen değişkenlerin aldığı değerlere göre hesaplanan uzaklıklarla illerin 2-boyutlu uzayda çok boyutlu ölçekleme ile elde edilen dağılımı Şekil 1'de görülmektedir. Bu dağılımın uyum iyiliğini gösteren değerler Tablo 3'te verilmiştir. Buna göre, beş değişken için 5-boyutlu uzaydaki gerçek uzaklıklarla 2-boyutlu uzaya indirgendiğinde elde edilen uzaklıkların uyumu, Kruskal $stress$ ölçüsüne göre (0,07421258) "iyi" düzeydedir. Ayrıca GOF ve R^2 değerleri 1'e yakın bulunmuştur. Uzaklıklar arasındaki ilişkiyi test eden F testi sonucunda göre de (F=162325,8; p=0,000) bu ilişki anlamlı düzeydedir.

Tablo 3: *Uyum İyiliği Değerleri*

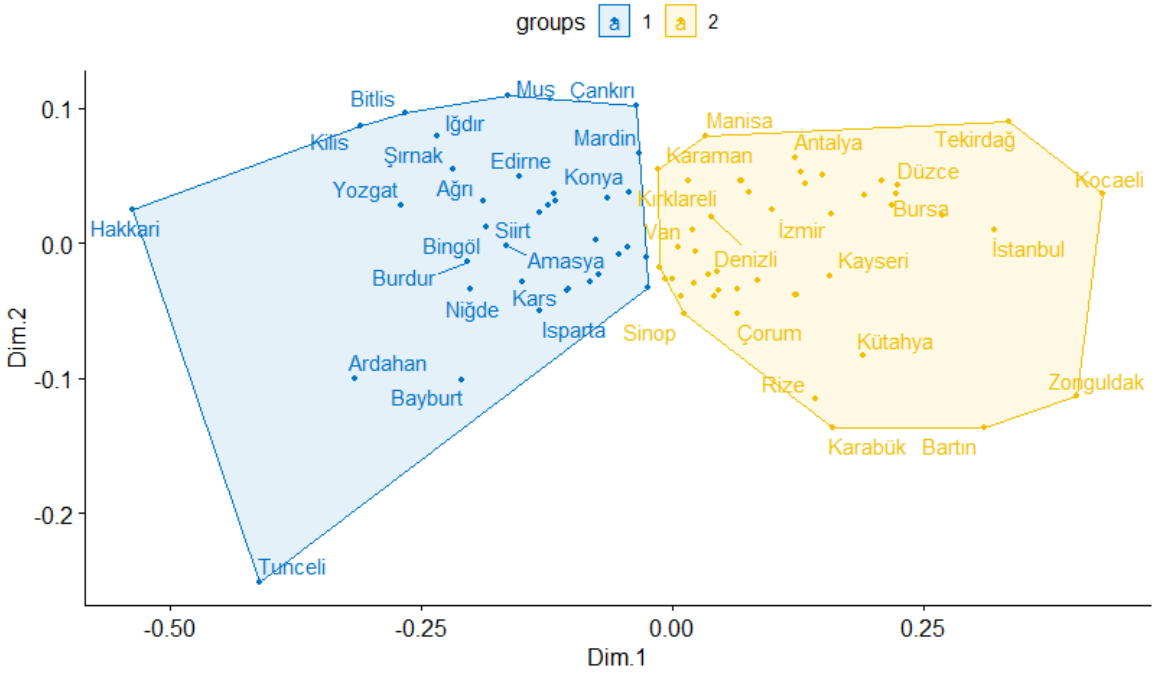
	GOF	R	R^2	F	sd	p	$stress$
4/1-a Oranları	0,9250329	0,990173	0,9804426	162325,8	3238	0,000	0,07421258
4/1-b Oranları	0,9511514	0,9938204	0,987679	259566,4	3238	0,000	0,05365879
4/1-c Oranları	0,9626831	0,996325	0,9926634	438113,6	3238	0,000	0,0526461

Aynı veri ile k-ortalamalar kümeleme analizi de yapılmıştır. Bunun için önce optimal küme sayısı belirlenmiştir. Algoritmayla küme sayısı 2 ile 10 arasında olacak şekilde kümelemeler elde edilmiş, sonra her biri için kümeleme performansı içsel indeksler yardımıyla karşılaştırılarak Tablo 4' de verilmiştir.

Tablo 4: *Kümeleme İndeksleri (4/1-a Değişkenleri İçin)*

k	**Wemmert-Gancarski**	**Silhouette**	**Dunn**	**Calinski-Harabasz**	**Xie-Beni**	**Davies-Bouldin**	**S-Dbw**
2	0,55884111	0,40421096	0,07868573	80,11230005	10,97969168	0,85938474	1,70435713
3	0,50296416	0,32456507	0,05697116	78,13229229	16,90317063	0,92374499	2,84658805
4	0,49608298	0,28914678	0,07979458	70,28883919	13,57937501	0,90736117	NaN
5	0,46257258	0,26952597	0,08514087	67,30748866	11,17597269	0,96923577	NaN
6	0,45045766	0,25359359	0,05690373	61,01362947	22,42718217	1,08291717	NaN
7	0,43878666	0,23484887	0,09731267	57,6861219	6,84509281	1,13392832	NaN
8	0,46759871	0,26253816	0,11767808	57,7506706	4,06478722	1,05155782	NaN
9	0,41539147	0,20966312	0,1342671	49,92203244	3,11804153	1,11477547	NaN
10	0,45210529	0,28090273	0,08514087	54,45791207	6,42365009	0,99313808	NaN
Karar Kriteri	***max***	***max***	***max***	***max***	***min***	***min***	***min***

Buna göre optimal küme sayısı Wemmert-Gancarski, Silhouette, Calinski-Harabasz, Davies-Bouldin ve S-Dbw indekslerine göre 2, Dunn ve Xie-Beni indekslerine göre 9 olmalıdır. İndekslerin çoğu 2 kümeyi işaret ettiği için k=2 olacak şekilde kümeler aynı düzlemde gösterilerek Şekil 1' de verilmiştir.



Şekil 1: *4/1-a Kapsamındaki Sigortalı Sayısı Oranlarına Göre İllerin Uzaklıkları ve Kümeleri*

Oluşan grafik incelendiğinde, illerin iki kümeye dengeli olarak dağıldığı görülmektedir. Birinci grupta Hakkâri ve Tunceli, ikinci grupta Bartın, Zonguldak ve Kocaeli kendi kümelerinde küme merkezinden uzakta kalmaktadır. İllerin kümelere dağılımı Tablo 5'te verilmiştir.

Tablo 5: *İllerin 4-1/a Oranlarına Göre Gruplara Dağılımı*

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

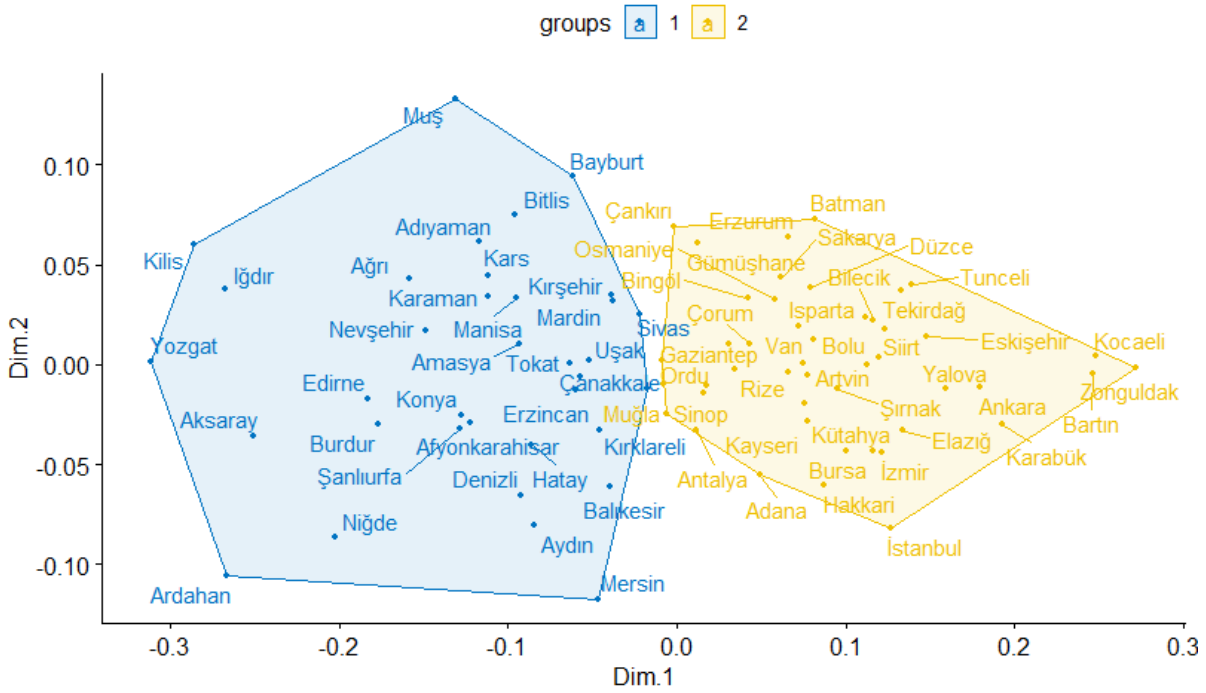
Türkiye'deki 81 ilde 4/1-b kapsamındaki aktif sigortalıların sayısıyla elde edilen değişkenlerin aldığı değerlere göre hesaplanan uzaklıklarla illerin 2-boyutlu uzayda çok boyutlu ölçekleme ile elde edilen dağılımı Şekil 2'de görülmektedir. Bu dağılımın uyum iyiliğini gösteren değerler (Tablo 3) incelendiğinde, beş değişken için 5-boyutlu uzaydaki gerçek uzaklıklarla 2-boyutlu uzaya indirgendiğinde elde edilen uzaklıkların uyumu, Kruskal $stress$ ölçüsüne göre (0,05365879) "iyi" düzeydedir. Ayrıca GOF ve R^2 değerleri 1'e yakın bulunmuştur. Uzaklıklar arasındaki ilişkiyi test eden F testi sonucunda göre de (F=259566,4; p=0,000) bu ilişki anlamlı düzeydedir.

Aynı veri ile k-ortalamalar kümeleme analizi de yapılmıştır. Bunun için önce optimal küme sayısı belirlenmiştir. Algoritmayla küme sayısı 2 ile 10 arasında olacak şekilde kümelemeler elde edilmiş, sonra her biri için kümeleme performansı içsel indeksler yardımıyla karşılaştırılarak Tablo 6' da verilmiştir.

Tablo 6: *Kümeleme İndeksleri (4/1-b Değişkenleri İçin)*

k	Wemmert-Gancarski	Silhouette	Dunn	Calinski-Harabasz	Xie-Beni	Davies-Bouldin	S-Dbw
2	0,60921349	0,45911573	0,04902342	107,6789107	37,45560365	0,76643945	1,06774224
3	0,5093533	0,33272861	0,0652704	89,57836656	17,35253448	0,95089646	1,6543283
4	0,48093747	0,3212105	0,09020401	85,3658126	7,68704777	0,9581336	2,7559853
5	0,47387533	0,3015395	0,1082811	77,6959234	6,53409708	0,99874618	4,53809778
6	0,48635398	0,34300736	0,09499568	78,01809236	8,96182102	0,92222423	NaN
7	0,48982877	0,35157536	0,11409662	73,43755726	7,99121813	0,92465099	NaN
8	0,44789782	0,25102522	0,09499371	59,00495469	6,48950247	1,05284135	NaN
9	0,48416937	0,30026383	0,08253806	66,65000167	10,93289108	0,93877284	NaN
10	0,44889475	0,26686357	0,04957362	61,96496878	28,76961585	1,01209622	NaN
Karar Kriteri	***max***	***max***	***max***	***max***	***min***	***min***	***min***

Buna göre optimal küme sayısı Wemmert-Gancarski, Silhouette, Calinski-Harabasz, Davies-Bouldin ve S-Dbw indekslerine göre 2, Dunn indeksine göre 7 ve Xie-Beni indeksine göre 8 olmalıdır. Bu nedenle k=2 olacak şekilde kümeler aynı düzlemde gösterilmiş ve Şekil 2'de verilmiştir.



Şekil 2: *4/1-b Kapsamındaki Sigortalı Sayısı Oranlarına Göre İllerin Uzaklıkları ve Kümeleri*

Oluşan grafik incelendiğinde, illerin iki kümeye dengeli olarak dağıldığı görülmektedir. Birinci grupta Ardahan, Muş, Kilis ve Yozgat, ikinci grupta Bartın, Zonguldak ve Kocaeli kendi kümelerinde küme merkezinden nispeten uzakta kalmaktadır. İllerin kümelere dağılımı Tablo 7'de verilmiştir.

Tablo 7: *İllerin 4-1/b Oranlarına Göre Gruplara Dağılımı*

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

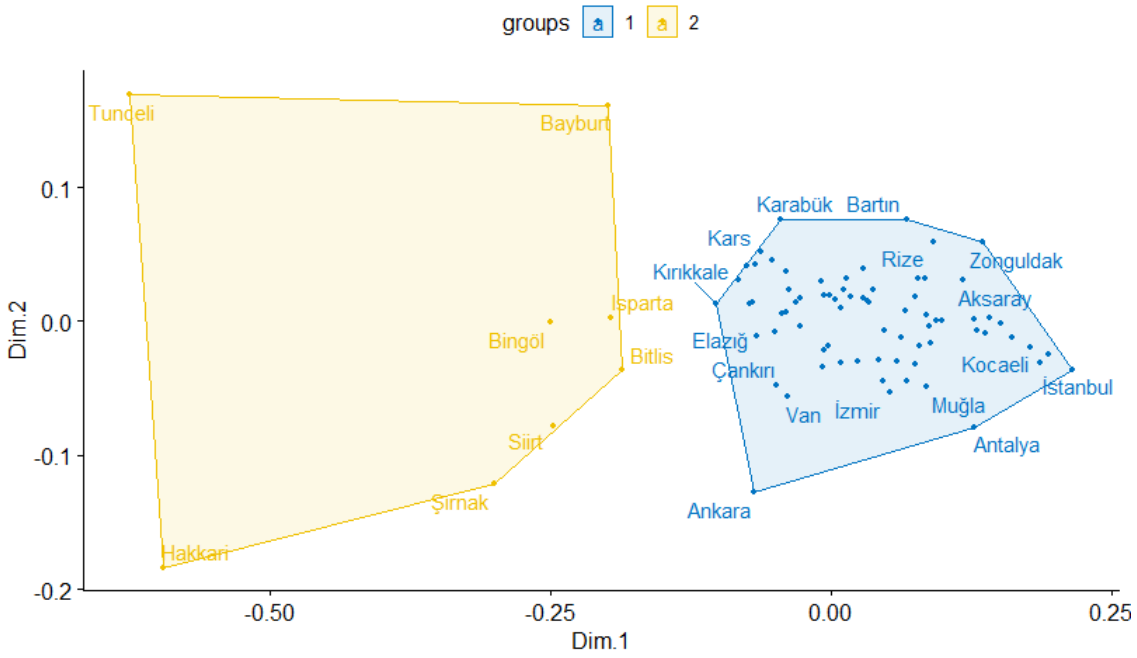
81 ildeki 4/1-c kapsamındaki aktif sigortalıların sayısıyla elde edilen değişkenlerin aldığı değerlere göre hesaplanan uzaklıklarla illerin 2-boyutlu uzayda çok boyutlu ölçekleme ile elde edilen dağılımı Şekil 3'te görülmektedir. Bu dağılımın uyum iyiliğini gösteren değerler (Tablo 3) incelendiğinde, beş değişken için 5-boyutlu uzaydaki gerçek uzaklıklarla 2-boyutlu uzaya indirgendiğinde elde edilen uzaklıkların uyumu, Kruskal $stress$ ölçüsüne göre (0,0526461) "iyi" düzeydedir. Ayrıca GOF ve R^2 değerleri 1'e yakın bulunmuştur. Uzaklıklar arasındaki ilişkiyi test eden F testi sonucunda göre de (F=438113,6; p=0,000) bu ilişki anlamlı düzeydedir.

Aynı veri ile k-ortalamalar kümeleme analizi de yapılmıştır. Bunun için önce optimal küme sayısı belirlenmiştir. Algoritmayla küme sayısı 2 ile 10 arasında olacak şekilde kümelemeler elde edilmiş, sonra her biri için kümeleme performansı içsel indeksler yardımıyla karşılaştırılmış ve Tablo 8' de verilmiştir.

Tablo 8: *Kümeleme İndeksleri (4/1-c Değişkenleri İçin)*

k	Wemmert-Gancarski	Silhouette	Dunn	Calinski-Harabasz	Xie-Beni	Davies-Bouldin	S-Dbw
2	0,74491233	0,47393264	0,18309561	81,06295249	1,21594724	0,72754397	1,25104541
3	0,55726238	0,38092817	0,02228885	82,18791825	118,1416988	0,76575455	1,92313753
4	0,47911951	0,32917881	0,01770695	73,25117706	150,9318546	0,84566537	2,18554932
5	0,5072216	0,31984853	0,05783552	89,35159434	9,56097901	0,86143023	2,39858475
6	0,45726892	0,28672116	0,05220483	80,33070889	10,5295644	0,98726775	NaN
7	0,46100425	0,27966327	0,04808565	72,42015738	11,4779756	1,00540066	4,62455449
8	0,46745473	0,31023695	0,04808565	63,53617669	11,12096463	0,97798297	NaN
9	0,47821413	0,27794793	0,04799627	64,18733396	9,73565218	0,93028634	NaN
10	0,50006012	NaN	0,04808565	54,73919245	9,93548689	0,87895826	NaN
Karar Kriteri	***max***	***max***	***max***	***max***	***min***	***min***	***min***

Buna göre optimal küme sayısı Calinski-Harabasz indeksine göre 5, diğer tüm indekslere göre 2 olmalıdır. Bu sebeple k=2 olacak şekilde kümeler aynı düzlemde gösterilerek Şekil 3' te verilmiştir.



Şekil 3: *4/1-c Kapsamındaki Sigortalı Sayısı Oranlarına Göre İllerin Uzaklıkları ve Kümeleri*

Oluşan grafik incelendiğinde, illerin iki kümeye dağılımında dengesizlik göze çarpmaktadır. Birinci grupta Bayburt, Hakkâri ve Tunceli, ikinci grupta nispeten Ankara, İstanbul, Antalya, Bartın, Zonguldak ve Karabük kendi kümelerinde küme merkezinden uzakta kalmaktadır. İllerin kümelere dağılımı Tablo 9'da verilmiştir.

Tablo 9: *İllerin 4-1/c Oranlarına Göre Gruplara Dağılımı*

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

3. Sonuç

Veri görselleştirme, istatistik ve veri madenciliği yöntemleriyle çalışmalarda sıklıkla başvurulan bir konu olmakla birlikte, her yöntemin kendi içinde bazı kısıtlı yönleri bulunmaktadır. Çok boyutlu ölçeklemeyle, çok değişkenli verilerle uzaklıklar yardımıyla iki veya üç boyutlu haritalar elde edilebilmektedir. Ancak bu grafiklerde boyutlara göre noktaların yakınlık ve uzaklıklarının yorumlanması güç olabilmektedir. Kümeleme ise noktaları küme merkezleri etrafında gruplandırabilmektedir. Bu iki yöntemin birlikte kullanılması hem dayandıkları uzaklıkların aynı seçilebilmesi hem de her iki yöntemin de bir çıkarım yapma amacı olmadan tanımlayıcı yöntemler olması dolayısıyla mümkündür. Görselleştirme tekniği olarak, birimlerin yakınlık ve uzaklıklarının çok boyutlu ölçeklemenin boyutları bakımından yorumlanması kümeleme sayesinde kolay ve anlaşılır hale gelmektedir. Açık kaynak kodlu R programlama dili bu analizlerin yapılmasına olanak sağlamaktadır.

Bu çalışmada, Türkiye'de her bir il için 2020 yılında 4/1-a, 4/1-b ve 4/1-c kapsamındaki; aktif sigortalıların toplam aktif sigortalılar içindeki oranları, aylık ve gelir alanların toplam aylık ve gelir alanlar içindeki oranları ve sigortalıların bakmakla yükümlü olduklarının toplam bakmakla yükümlü olunanlar içindeki oranları kullanılarak bir analiz gerçekleştirilmiştir. Bu sigorta grupları için ayrı ayrı yapılan analizlerde beşer adet değişken (öznitelik) bulunmaktadır. Bu değişkenler için çok boyutlu ölçekleme ile öklid uzaklığı kullanılarak her bir il 2-boyutlu uzayda nokta olarak belirtilmiştir. Noktaların bu düzlemdeki dağılımlarının gerçek

dağılımlarıyla uyumu iyi düzeyde bulunmuştur. Ardından bu noktalar kümeleme analizi ile gruplandırılmıştır. Küme sayısının belirlenmesinde içsel kümeleme performans indeksleri kullanılmıştır. Buna göre noktalar iki grup olarak kümelenmektedirler.

Analiz bulgularıyla elde edilen görseller ile çalışanlar, aylık ve gelir alanlar veya bunların bakmakla yükümlü oldukları sigortalıların oranlarına göre illerin benzerlik ve farklılıkları yorumlanabilecektir. Bazı illerin üç grup sigortalı türünde de diğer illerden farklılaştığı görülmektedir. Bu o ilin sektörel dağılımı ve demografik yapısı ile ilgili olabileceği gibi istihdam yapısı ile ilgili de olabileceği düşünülebilir. Bulgularda 4/1-c kapsamındaki dağılımlara göre neredeyse tüm illerin bir kümede gruplandığı göze çarpmaktadır. Bu da bu kapsamdaki sigortalıların oransal olarak ülke genelinde aynı düzeylerde dağıldığını göstermektedir.

Bu çalışma, istatistik ve veri madenciliği yöntemlerinin, veri görselleştirme destekli olarak sosyal güvenlik alanında kullanılması yönüyle literatüre katkılar sağlamaktadır. Çalışma zaman boyutuyla da uygulanarak illerin sosyal güvenlik yapısındaki değişimler incelenebilir. Elde edilen görsellerin web uygulaması olarak, seçilen değişkenlerle ve iller için verideki değişime uyum sağlayabilecek ve güncellenebilecek şekilde R Shiny uygulaması üzerinden infografikler olarak uyarlanması mümkündür.

Kaynakça

Agrawal, K., Garg, S., & Patel, P. (2015). Performance Measures for Densed and Arbitrary Shaped Clusters. International Journal of Computer Science & Communication, 6(2), 338-350.

Albayrak, A. S., & Koltan Yılmaz, Ş. (2009). Veri Madenciliği: Karar Ağaçları ve İMKB Verileri Üzerine Bir Uygulama. Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 14(1), 31-52.

Aliukov, S., & Buleca, J. (2022). Comparative Multidimensional Analysis of the Current State of European Economies Based on the Complex of Macroeconomic Indicators. Mathematics 2022, Vol. 10, Page 847, 10(5), 847. https://doi.org/10.3390/MATH10050847

Allahverdi, F., Allahverdi, M., & Çevik, S. (2021). Türkiye'de Kamu Harcamalarının İl Düzeyinde Dağılımının Çok Boyutlu Ölçekleme ve Kümeleme Analizi ile İncelenmesi. Maliye Dergisi, 0(180), 31-60.

Almeira, D., & Graciella Juanda, G. (2021). Analisis Multidimensional Scaling dan k-Means Clustering untuk Pengelompokan Provinsi Berdasarkan Tingkat Pengangguran. E-Prosiding Nasional | Departemen Statistika FMIPA Universitas Padjadjaran, 10, 08. http://prosiding.statistics.unpad.ac.id/index.php/prosidingnasional/article/view/75

Alpar, R. (2017). Uygulamalı Çok Değişkenli İstatistik Yöntemler (5. b.). Ankara: Detay Yayıncılık.

Arı, E., & Gülcemal, M. E. (2019). OECD Ülkelerinin Sigorta Pazar Paylarının Çok Değişkenli İstatistiksel Yöntemlerle İncelenmesi. Batman Üniversitesi Yaşam Bilimleri Dergisi, 9(2), 136–157.

Ayala, E., Nelson, L., Bartholomew, M., & Plummer, D. (2022). A conceptual model for mental health and performance of North American athletes: A mixed methods study. Psychology of Sport and Exercise, 102176. https://doi.org/10.1016/J.PSYCHSPORT.2022.102176

Bache, S. M., & Wickham, H. (2020). magrittr: A Forward-Pipe Operator for R. R package version 2.0.1. https://CRAN.R-project.org/package=magrittr adresinden alındı

Borg, I., & Groenen, P. J. (2005). Modern Multidimensional Scaling: Theory and Applications. Springer.

Borsenberger, M., Fleury, C., & Dickes, P. (2016). Welfare regimes and social cohesion regimes: do they express the same values?, European Societies, 18(3), 221-244. https://doi.org/10.1080/14616696.2016.1172717

Buja, A., Swayne, D. F., Littman, M. L., Dean, N., Hofmann, H., & Chen, L. (2012). Data Visualization With Multidimensional Scaling, Journal of Computational and Graphical Statistics, 17(2), 444–472. https://doi.org/10.1198/106186008X318440

Demirel Top, E., Yapici, N., & Cetinkaya, C. (2018). Comparison of Fatal Occupational Accidents Statistics inTurkey with Some European Countries. International Journal of Scientific and Technological Research, Vol 4, No.6, 107-119.

Desgraupes, B. (2017). Package clusterCrit for R: Clustering Indices. Kasım 5, 2021 tarihinde https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf adresinden alındı

Desgraupes, B. (2018). clusterCrit: Clustering Indices. R package version 1.2.8. https://CRAN.R-project.org/package=clusterCrit adresinden alındı

Fleming, L., Lemonde, A.-C., Gold, J., Taylor., J., Malla, A., Joober, R., Iyer, V., Lepage, M., Shah, J., & Corlett, P. R. (t.y.). Reducing the dimensions of psychotic illness. PsyArXiv https://doi.org/10.31234/OSF.IO/WJ89F

Graffelman, J. (2020). Goodness-of-fit filtering in classical metric multidimensional scaling with large datasets. Journal of Applied Statistics, 47(11), 2011-2024.

Gürsakal, S. (2019). Sosyal Bilimlerde SPSS Uygulamalı Çok Değişkenli İstatistiksel Analiz Teknikleri. Bursa: Dora Yayıncılık.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining Trends and Research Frontiers, Editor(s): Jiawei Han, Micheline Kamber, Jian Pei, In The Morgan Kaufmann Series in Data Management Systems, Data Mining: Concepts and Techniques (Third Edition), Morgan Kaufmann, Pages 585-631. https://doi.org/10.1016/B978-0-12-381479-1.00013-7

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). Multivariate Data Analysis (7. b.). Edinburgh Gate: Pearson.

Hofmann, T., & Buhmann, J. (1995). Multidimensional scaling and data clustering. T. K. Leen, G. Tesauro, & D. S. Touretzky içinde, Advances in Neural Information Processing Systems 7 (s. 459-466). The MIT Press.

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery (2), 283-304.

Inekwe, J., Maharaj, E., & Bhattacharya, M. (2020). Drivers of carbon dioxide emissions: an empirical investigation using hierarchical and non-hierarchical clustering methods. Environ Ecol Stat (27), 1-40.

Karadağ Erdemir, Ö., & Tatlıdil, H. (2018). The Use of the Multivariate Statistical Methods in the Performance Analysis of Non-Life Insurance Companies. Finansal Araştırmalar ve Çalışmalar Dergisi, 56–69. https://doi.org/10.14784/marufacd.460659

Kassambara, A. (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr adresinden alındı

Kırcı Çevik, N. (2021). OECD Ülkeleri Sağlık Sistemi Göstergelerine Çok Boyutlu Bir Yaklaşım. Iğdır Üniversitesi Sosyal Bilimler Dergisi. https://doi.org/10.54600/igdirsosbilder.991828

Koçoğlu, F. Ö., & Esnaf, Ş. (2019). Veri Madenciliği Kümeleme Algoritmalarının Başarı Göstergesi Olarak Kümeleme İndeks Değerlerinin İncelenmesi. M. E. Balaban, & E. Kartal içinde, Veri Madenciliği ve Makine Öğrenmesi Temel Kavramlar, Algoritmalar, Uygulamalar (s. 243-288). İstanbul: Çağlayan Kitabevi.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29(1), 1-27.

Kruskal, J. B., & Carmone, F. J. (1967). How to Use MDSCAL, Version 5-M, and Other Useful Information. Murray Hill, NJ.: Bell Laboratories.

Kruskal, J. B., & Wish, M. (1978). Multidimensional Scaling, Sage University Papers Series, Quantitative Applications in the Social Sciences, No. 07-011 . Sage Publications, Inc. https://doi.org/10.4135/9781412985130

Leeuw, J. d., & Mair, P. (2009). Multidimensional Scaling Using Majorization: SMACOF in R. Journal of Statistical Software, 31(3), 1-30. https://www.jstatsoft.org/v31/i03/ adresinden alındı

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. Pattern Recognition(36), 451-461.

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of Internal Clustering Validation Measures. 2010 IEEE International Conference on Data Mining (s. 911-916). Sydney, Australia: IEEE.

Lopes, A. M., & Machado, J. A. T. (2022). Multidimensional scaling and visualization of patterns in global large-scale accidents. Chaos, Solitons & Fractals, 157, 111951. https://doi.org/10.1016/J.CHAOS.2022.111951

Mair, P., Groenen, P. J., & Leeuw, J. d. (2021). More on Multidimensional Scaling in R: smacof Version 2. Journal of Statistical Software.

Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(12), 1650-1654. https://doi.org/10.1109/TPAMI.2002.1114856

Montgomery, L. T., Vaughn, L. M., & Jacquez, F. (2022). Engaging Adolescents in the Fight Against Drug Abuse and Addiction: A Concept Mapping Approach. Health Education & Behavior. January 2022. https://doi.org/10.1177/10901981211068416

Oğuzlar, A. (2005a). Kümeleme Analizinde Yeni Bir Yaklaşım: Kendini Düzenleyen Haritalar (Kohonen Ağları). Journal of Economics and Administrative Sciences, 19(2), 93–107. https://dergipark.org.tr/en/pub/atauniiibd/issue/2688/35322

Oğuzlar, A. (2005b). Çok Boyutlu Ölçekleme Analizi Yardımıyla Avrupa Birliği Üyeliğini Etkileyen Faktörlerin Konumlandırılması. Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, XXIV(1), 33-43.

Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E. M. (2011). Internal versus external cluster validation indexes. International Journal of Computers and Communications, 5(1), 27-34.

Roux, I. (2008). Application of cluster analysis and multidimensional scaling on medical schemes data. Master Thesis, Stellenbosch University, http://scholar.sun.ac.za/handle/10019.1/2040

RStudio. (2022, Şubat 1). RStudio | Open source & professional software for data science teams. Kasım 5, 2021 tarihinde RStudio: https://www.rstudio.com/ adresinden alındı

SGK. (2021, Ekim 14). Sosyal Güvenlik Kurumu İstatistik Yıllığı Sigortalı ve İş Yeri İstatistikleri 2020. Ankara, Balgat, Türkiye. Kasım 2, 2021 tarihinde http://www.sgk.gov.tr/wps/portal/sgk/tr/kurumsal/istatistik/sgk_istatistik_yilliklari adresinden alındı

Shanti, R. (2019). Multivariate Data Analysis: Using SPSS and AMOS. Chennai: MJP Publisher.

Sips, M. (2009). Visual Clustering. In L. Liu & M. T. Özsu (Eds.), Encyclopedia of Database Systems (pp. 3355–3360). Springer. https://doi.org/10.1007/978-0-387-39940-9_1124

Team, R. C. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/ adresinden alındı

The R Foundation. (2022, Mart 10). The R Project for Statistical Computing. Kasım 5, 2021 tarihinde R 4.1.3: https://www.r-project.org/ adresinden alındı

Vatansever, M., & Büyüklü, A. H. (2009). Using Visual Data Mining Techniques in Clustering Analysis and An Application. Sigma Journal of Engineering and Natural Sciences, 27, 83–104.

Velado-Alonso, E., Morales-Castilla, I., & Gómez-Sal, A. (2022). The landscapes of livestock diversity: grazing local breeds as a proxy for domesticated species adaptation to the environment. Landscape Ecology 2022, 1–14. https://doi.org/10.1007/S10980-022-01429-5

Wickelmaier, F. (2003). An Introduction to MDS. Aalborg University. Denmark: Sound Quality Research Unit.

Wickham, H., & Bryan, J. (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl adresinden alındı

Wickham, H., François, R., Henry, L., & Müller, K. (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr adresinden alındı

Wu, J. (2012). Advances in K-means Clustering: A Data Mining Thinking. Berlin Heidelberg: Springer.,

Žalik, K. R. (2008). An efficient k′-means clustering algorithm. Pattern Recognition Letters, 29(9), 1385-1391. https://doi.org/10.1016/j.patrec.2008.02.014.