İstatistik Araştırma Dergisi
Journal of Statistical Research
ISSN: 2791-7614

(2022), 12 (1)
Başvuru/Received: 08/04/2022
Kabul/Accepted: 27/06/2022

https://journal.tuik.gov.tr/

ARAŞTIRMA MAKALESİ RESEARCH ARTICLE

Tabaka Sınırlarının Belirlenmesinde, Kümeleme Analiz Yöntemleri ve Tabaka Sınırı Belirleme Yöntemlerinin Karşılaştırılması[1]

Osman SERT
Türkiye İstatistik Kurumu / TÜİK Uzmanı
osman.sert@tuik.gov.tr
Orcid No: 0000-0002-0928-851X

Alpaslan AKÇORAOĞLU
Hacı Bayram Veli Üniversitesi / Prof. Dr.
alpaslan.akcoraoglu@hbv.edu.tr
Orcid No: 0000-0003-4368-9998

Özet

Tabakalı örneklemede örnekleme çerçevesi birbiriyle örtüşmeyen tabakalara bölünür. Bu bölünme çoğu pratik durumda coğrafi bölgeler, cinsiyet, yaş gibi doğal durumu yansıtacak şekilde kendiliğinden oluşur. Bu şekilde kendiliğinden oluşan tabakalar araştırma değişkenine göre içsel olarak homojen olmayabilir. Fakat araştırma değişkeninin tabaka sınırları katı bir şekilde önceden belirlenmemişse tabakaların içsel olarak homojenliği sağlanabilir. Bu yolla araştırma değişkenine göre içsel olarak homojen tabakalar oluşturulur ve tahmin hassasiyeti en üst düzeye çıkarılmış olur. Bu amaçla bu çalışmada tabaka-içi homojenliği sağlamak için Kümeleme Analiz Yöntemleri ve Tabaka Sınırı Belirleme Yöntemleri karşılaştırılmıştır. Bu karşılaştırma için beş farklı çarpıklık değerine sahip veri setleri türetilmiş ve her veri seti için ayrıştırma yöntemlerine göre tabaka sınırları belirlenmiştir. Tahminlerin güvenilirliğini arttırmak için her yöntemden 1000 kez bağımsız örnek seçilmiştir. Böylece her yöntemden elde edilen ortalama tahmin edicisine ilişkin Kök Hata Kareler Ortalamaları (KHKO) hesaplanmış ve en küçük KHKO değerini veren yöntemin tabaka sınırlarının optimum sınırlar olduğu sonucuna ulaşılmıştır. Analizler R programında yer alan "NbClust" ve "Stratification" paketleri ile yapılmıştır. Çalışmada elde edilen sonuçlara göre en küçük üç çarpıklık değerine sahip simülasyonlar için "Lavallee-Hidiroglou", dördüncü simülasyon için "Ortalama Kümeleme" ve en büyük çarpıklık değerine sahip simülasyon için ise "K-Ortalamalar Kümeleme" yöntemleri ile elde edilen tabaka sınırları, optimum tabaka sınırları olarak belirlenmiştir.

Anahtar Sözcükler: Kümeleme Analizi, Optimum Tabaka Sınırları, Tabakalı Tesadüfi Örnekleme

[1] Bu çalışma ilk yazarın doktora tezinden türetilmiştir. Çalışmada ifade edilen görüşler tamamen yazarlara aittir ve Türkiye İstatistik Kurumu'nu bağlamaz.

Sorumlu Yazar / Corresponding Author: 1-Osman SERT, Türkiye İstatistik Kurumu, Yöntem Araştırmaları Daire Başkanlığı, Veri Analiz Teknikleri Grup Başkanlığı.

2-Alpaslan AKÇORAOĞLU, Hacı Bayram Veli Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Ekonometri Bölümü.

Atıf / Citation: SERT O., AKÇORAOĞLU A. (2022). Tabaka Sınırlarının Belirlenmesinde, Kümeleme Analiz Yöntemleri ve Tabaka Sınırı Belirleme Yöntemlerinin Karşılaştırılması. İstatistik Araştırma Dergisi, 12 (1), 68-81.

Atıf / Citation: SERT O., AKÇORAOĞLU A. (2022). Tabaka Sınırlarının Belirlenmesinde, Kümeleme Analiz Yöntemleri ve Tabaka Sınırı Belirleme Yöntemlerinin Karşılaştırılması. İstatistik Araştırma Dergisi, 12 (1), 68-81.

Comparison of Cluster Analysis Methods and Strata Boundary Determination Methods for Determination of Strata Boundaries

Abstract

In stratified sampling, the sampling frame is divided into non-overlapping strata. In most practical cases, this division occurs spontaneously, reflecting the natural state such as geographical regions, gender and age. Self-forming strata formed in this way may not be internally homogeneous according to the research variable. However, internal homogeneity of the strata can be achieved if the strata boundaries of the research variable are not strictly predetermined. Thus, internally homogeneous strata are formed according to the research variable and precision of an estimate is maximized. For this purpose, in this study, Cluster Analysis Methods and Optimum Strata Boundary Determination Methods were compared to ensure intra-stratum homogeneity. For this comparison, data sets with five different skewness values were derived. For each data set, strata boundaries were determined according to the decomposition methods. In order to increase the reliability of the estimations, independent samples were selected from each method for 1000 times. Thus, the Root Mean Squares Error (RMSE) of the mean estimator obtained from each method was calculated and it was concluded that the method with the smallest RMSE value had the optimum strata boundaries. Analyzes were made with the "NbClust" and "Stratification" packages in the R program. According to the results obtained in the study the strata boundaries obtained by the “Lavallee-Hidiroglou” Method for the simulations with the 3 lowest skewness values; “Average Clustering” Method for the fourth simulation; and “K-Means Clustering” Method for the simulation with the largest skewness value were determined as the optimum strata boundaries.

Keywords: Cluster Analysis, Optimum Strata Boundaries, Stratified Random Sampling

1. Giriş

Tabakalı Tesadüfi Örnekleme Yöntemi; işsizlik oranı, gelir eşitsizliği, genel mutluluk düzeyi, turizm istatistikleri, mali aracı kuruluş istatistikleri ve girişimlerde bilişim teknolojileri kullanım istatistiklerinin hesaplanması gibi ülkelerin istatistik kurumları tarafından yürütülen birçok parametre tahmininde kullanılan önemli bir örnekleme tekniğidir. Tabakalı örneklemede örnekleme çerçevesi, örtüşmeyen gruplara veya tabakalara bölünür. Oluşturulan tabakaların araştırma değişkenine göre içsel olarak homojen olması hedeflenir ve böylece tahminin hassasiyeti en üst düzeye çıkarılmış olur. Fakat çoğu pratik durumda, bu gibi optimum tabakaları oluşturmak çok zordur. Çünkü araştırmacılar coğrafi bölgeler (örn; Batı Karedeniz, Orta Anadolu, vb.), idari bölgeler (örn; iller, ilçeler, vb.), cinsiyet, yaş veya diğer doğal ölçütler (örn; kentsel-kırsal alan) gibi doğal durumu yansıtacak şekilde topluluğu tabakalaştırır (Khan, Reddy, & Rao, 2015). Fakat kullanılan değişkenin tabaka sınırları katı bir şekilde önceden belirlenmemişse, (örn; firmaların ciro değerleri ya da çalışan sayıları gibi) tabakaların içsel olarak homojenliği sağlanabilir. Bu nedenle, tahminlerin hassasiyetini arttırmak için Optimum Tabaka Sınırları (OTS) belirlenmelidir. OTS'nin belirlenmesinde temel husus tabakaların mümkün olduğunca içsel olarak homojen olması, yani tabaka varyanslarının belirli bir örnek dağılımı için mümkün olduğunca küçük olmasıdır.

Bu amaçla bu çalışmada sırasıyla Tabaka Sınırı Belirleme Yöntemleri ve Kümeleme Analiz Yöntemlerine ait literatür bilgisi verilmiş, sonrasında türetilen veri setleri üzerinde uygulama yapılmış ve sonuçlar sunulmuştur.

2. Tabaka Sınırı Belirleme Yöntemleri

Tabakalı Örnekleme Yönteminde oluşturulan tabakaların içsel olarak homojen olmaması sorunu birçok araştırmacı tarafından incelenmiş ve tahminlerin hassasiyetini arttıran Tabaka Sınırları Belirleme Yöntemleri geliştirilmiştir. Literatürde değişkenlerin tabaka sınırlarını belirlemek için birçok çalışma yapılmıştır. Genel olarak bu yöntemler yakınsama ve optimizasyon bazlı olmak üzere iki başlıkta toplanır (Hidiroglou & Kozak, 2017). Yakınsama bazlı yöntemlerden en yaygın olanları Dalenius ve Hodges (1957) tarafından önerilen Birikimli Kök Frekans Yöntemi ile Gunning ve Horgan (2004) tarafından önerilen Geometrik Tabakalama Yöntemidir. Mahalanobis (1952), Ekman (1959), Sethi (1963), Singh (1971) ve Thomsen (1976) çalışmaları ise diğer yakınsama örnekleri arasında gösterilebilir. Optimizasyon Bazlı Tabakalama Yöntemlerinin en yaygın kullanılanları ise Lavallee ve Hidiroglou (1988) ve Kozak (2004) tarafından ortaya atılan yöntemlerdir. Diğer yöntemler ise Sweet ve Sigman (1995), Rivest

(2002), Lednicki ve Wieczorkowski (2003), Keskintürk ve Er (2007), Verma, Kozak ve Zielinski (2007), Benedetti, Bee ve Espa (2010), Brito, Ochi, Montenegro ve Maculan (2010) ve Ballin ve Barcaroli (2013) tarafından önerilmiştir. Bu çalışmada literatürde en çok karşılaşılan ve karşılaştırılan aşağıdaki 3 yöntem kullanılmıştır.

2.1. Birikimli Kök Frekans Yöntemi

Birikimli Kök Frekansı Yöntemi en popüler tabaka sınırı belirleme yöntemlerinden biridir. Bu yöntemde ilgilenilen değişkenin toplam tahmin varyansı en aza indirgenerek tabaka sınırları belirlenir. Bu işlemde tabakalama değişkeninin tabakalama noktaları arasında yaklaşık olarak eşit dağıldığı varsayılır. Sınırlar, frekansların karekökünün kümülatif fonksiyonu üzerinde eşit aralıklar alınarak elde edilir (Dalenius & Hodges, 1957). Sınırları belirlemek için, aşağıdaki denklemlerin yinelemeli olarak çözülmesi gerekir:

$$\frac{S_h^2+(\mu_h-\bar{X}_h)^2}{S_h} = \frac{S_{h+1}^2+(\mu_{h+1}-\bar{X}_h)^2}{S_{h+1}}, \quad h = 1,2,\dots,L-1 \qquad (2.1)$$

Burada h tabaka indisi, L tabaka sayısı, μ_h kitle ortalaması, $\bar{X}_h$ *h* tabakasının kitle ortalaması ve S_h^2 h tabakasındaki tabakalama değişkeninin kitle varyansıdır.

Yöntemin işlem adımları aşağıda verilmiştir:

1. Veri küçükten büyüğe sıralanır,
2. Veri sınıflara ayrılır. Sınıf sayısı (J) tabaka sayısından fazla olmalıdır ($J > L$),
3. Her sınıftaki frekans hesaplanır ($f_i, i = 1,2,\dots,J$),
4. Her sınıfın frekansının karekökü hesaplanır,
5. Frekansların kareköklerinin toplamı hesaplanır ($\sum_{i=1}^{J}\sqrt{f_i}$),
6. Frekansların kareköklerinin toplamı, tabaka sayısına bölünür ($Q = \frac{1}{L}\sum_{i=1}^{J}\sqrt{f_i}$),
7. Hesaplanan Q değeri her tabakanın üst sınırı olur $(Q, 2Q, \dots, (L-1)Q, LQ)$.

2.2. Geometrik Yöntem

Bu tabaka sınırı belirleme yöntemi, ilgilenilen yardımcı değişkenin değişim katsayısını her tabakada eşitleyerek tabaka sınırlarının belirlenmesi fikrine dayanmaktadır. Sadece ilgili değişken x ve tabaka sayısı (L) gerektiren sınırların hesaplanması, geometrik bir ilerlemeye dayanır (Gunning & Horgan, 2004).

$$b_h = m\left(\frac{M}{m}\right)^{\frac{h}{L}}, \quad h = 0,1,\dots,L-1 \qquad (2.2)$$

Bu sınırlar, m ve M'nin sırasıyla x'lerin minimum ve maksimum değerleri olduğu yerlerdir. Ortaya çıkan tabakalar $[b_h; b_h + 1)$ için $h = 0,1,\dots,L-1$ şeklinde gösterilebilir. Örnek olarak; tabaka sayısı 4 (L=4), kitledeki en küçük değer 5 (m=5) ve en büyük değer 50000 (M=50000) ise tabaka sınırları;

$$b_h = m\left(\frac{M}{m}\right)^{\frac{h}{L}} = 5((\frac{50000}{5})^{1/4})^h = 5 * 10^h, \qquad (h = 0,1,2,3,4)$$

ile hesaplanır. Buna göre tabaka sınırları sırasıyla 5-50;51-500;501-5000;5001-50000 olacaktır.

Gunning ve Horgan (2004) algoritmanın normal dağılımlar için çalışmadığına dikkat çekmiştir. Daha genel olarak, yöntem herhangi bir simetrik dağılım için çalışmamaktadır. Bununla birlikte sınırlar geometrik olarak arttığından, minimum değeri sıfıra yakın değişkenler için yöntemin iyi çalışmayacağını da belirtmişlerdir. Ayrıca en yüksek değerlerin nadir olduğu yüksek derecede sağa çarpık dağılımlar için en iyi sonuçları elde etmeyi beklemişlerdir.

2.3. Lavallee-Hidiroglou Yöntemi

Lavallee ve Hidiroglou (1988) çalışmasında önerilen tabakalama yöntemi, kitlenin tam-sayım ve örnekleme tabakalarına ayrıştırılması özelliğiyle daha önce tanımlanmış olan yöntemlerden farklılık göstermektedir. Tam-sayım tabakasında tabakadaki tüm birimler kesin olarak seçilirken, örnekleme tabakalarında birimler yerine

koymadan basit tesadüfi örnekleme yöntemi kullanılarak seçilir. Yöntemin amacı, tahmin edicinin verilen değişim katsayısına (CV) ve örnek dağıtım şekline göre örnek hacmini minimize edecek tabaka sınırlarını hesaplamaktır (Lavallee & Hidiroglou, 1988). Buradaki problem, verilen örnek dağıtım yöntemini kullanarak istenen değişim katsayısı kısıtında minimum örnek hacmini elde edecek tabaka sınırlarının $y_1 < b_1 < b_2 < \cdots < b_{L-1} < y_N$ hesaplanmasıdır (1. tabaka sınırı: $y_1 \cdots b_1$, 2. tabaka sınırı: $b_1 + 1 \cdots b_2$ olacak şekilde). Bu yöntemde toplam örnek hacmi (n) aşağıdaki formülle tabakalara dağıtılır.

$$n = N_L + \frac{N(\sum_{h=1}^{L-1} W_h^2 S_h^2 / \dot{A}_h)}{N(c\bar{X})^2 + \sum_{h=1}^{L-1} W_h S_h^2} \tag{2.3}$$

Burada;

$\dot{A}_h$; örnek dağıtım yöntemi,

W_h; örnekleme ağırlığı (N_h/N),

N_L; tam-sayım tabakasına dağıtılan birimlerin sayısı,

c; güven seviyesi,

$\bar{X}$; kitle ortalaması,

S_h^2; h tabakasının kitle varyansı olarak ifade edilmiştir.

$\dot{A}_h$ formülü ise aşağıdaki gibidir:

$$\dot{A}_h = \frac{N_h^{2q_1} \bar{X}_h^{2q_2} S_h^{2q_3}}{\sum_{h=1}^{L} N_h^{2q_1} \bar{X}_h^{2q_2} S_h^{2q_3}} \tag{2.4}$$

Burada, $0 \leq 2q_1 \leq 1, 0 \leq 2q_2 \leq 1$ ve $0 \leq 2q_3 \leq 1$'tür. Bu genel formül bilinen bütün dağıtım yöntemlerini göstermektedir. Aşağıdaki tabloda hangi kombinasyonun hangi dağıtım yöntemini temsil ettiği ve istenen dağıtım yöntemini seçmek için gerekli q değerleri gösterilmiştir (Lavallee & Hidiroglou, 1988).

Tablo 1. $\dot{A}_h$ ile Dağıtım Yöntemleri Arasındaki İlişki

Dağıtım Yöntemi	$2q_1$	$2q_2$	$2q_3$
Neyman Dağıtım	1	0	1
X-Oransal Dağıtım	1	1	0
N-Oransal Dağıtım	1	0	0
Güç Dağıtım	p	p	0

Buna göre algoritmanın işleyişi aşağıdaki adımlarda verilmiştir:

1. Kitle küçükten büyüğe sıralanır,
2. Öncelikle geçici tabaka sınırları belirlenir $b_0 < b'_1 < b'_2 < \cdots < b'_{L-1} < b_L$,
3. Bu tabaka sınırları için tek tek tabaka ağırlığı, ortalaması ve varyansı hesaplanır,
4. Başlangıç tabaka sınırları aşağıdaki formülle yeni hesaplananlarla değiştirilir ($b''_1, \ldots, b''_{L-1}$),

$$b''_h = \frac{\beta'_h + \sqrt{\beta'^2_h - 4\alpha'_h \gamma'_h}}{2\alpha'_h}, h = 1,2, \ldots, L-1 \tag{2.5}$$

Burada yer alan α, β, γ terimleri, örnek hacmi hesabı için oluşturulan formüldeki 2. dereceden reel kök hesabındaki katsayılardır. Bu hesap $\alpha_h b_h^2 + \beta_h b_h + \gamma_h = 0$ şeklinde gösterilebilir,

5. Ardışık iki hesaplama aynı olana veya ihmal edilebilir miktarlarda farklılık gösterene kadar 3. ve 4. adım tekrarlanır.

3. Kümeleme Analiz Yöntemleri

Kümeleme Analizi temel olarak benzer birimlerin bir araya getirilmesi yaklaşımını kullanmaktadır. Kümeleme Analizi, gruplanmamış verileri benzerliklerine göre gruplandırarak araştırmacıya özetleyici bilgiler sunar. Kümeleme analizindeki amaç, benzer olarak bir araya getirilen birimlerden oluşan kümelerde; küme içi değişimin minimum, kümeler arası değişimin ise maksimum olmasını sağlamaktır. Literatürde birçok kümeleme analiz yöntemi bulunmaktadır ve genel olarak hiyerarşik (aşamalı) ve hiyerarşik olmayan (aşamalı olmayan) yöntemler biçiminde ikiye ayrılır.

3.1. Hiyerarşik Kümeleme Yöntemleri

Hiyerarşik Kümeleme Yönteminde verideki n sayıda gözlemin her biri bir küme olarak düşünülür. Bu kümelere birleştirme işlemleri uygulanır ve sonuçta tek bir küme oluşana kadar bu işlemler devam eder. Hiyerarşik kümeleme yönteminin genel adımları aşağıda verilmiştir (Servi, 2009):

1. Küme sayısı n olarak alınır. $K_1,K_2,\ldots..,K_n$ ile ifade edilen bu n sayıda küme için benzerlik matrisi hesaplanır.
2. Benzerlik matrisinde n küme sayısı, i = 1,2,….,n , j = 1,2,…,n ve i ≠ j olmak üzere tüm *min(d(Ki,Kj))* uzaklığına sahip iki küme belirlenir. Bu iki küme birleştirilerek yeni bir küme oluşturulur.
3. Benzerlik matrisi yeni oluşan küme de dikkate alınarak güncellenir.
4. Tek bir küme elde edilinceye kadar 2. ve 3. adımlar tekrarlanır.

Hiyerarşik kümelemede benzerlik matrisinin hesaplanması tanımlanan uzaklık fonksiyonuna bağlıdır. Literatürde birçok uzaklık fonksiyonu tanımı vardır. Uzaklık ölçüleri arasındaki fark bu çalışmanın konusu olmadığı için bu çalışmada uzaklık ölçüsü olarak sıklıkla kullanılan "Öklid Uzaklığı" kullanılmıştır. Öklid uzaklığı, gözlem vektörleri arasındaki farkların kareleri toplamının karekökünün alınması ile $d(x_i, x_j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$ formülü ile hesaplanır (Erişoğlu, 2011). Bu formülde p değişken sayısıdır.

Bu çalışmada hiyerarşik kümeleme analiz yöntemlerinden aşağıdaki 6 tanesi kullanılacaktır:

i. Tek Bağlantı Kümeleme Yöntemi: Florek, Lukaszewicz, Steinhaus ve Zubrzycki (1951) çalışmasında ilk kez önerilen bu yöntemde iki farklı kümedeki en yakın gözlemlere göre belirlenen uzaklık değeri, iki küme arasındaki uzaklık olarak tanımlanır. Tek bağlantı kümeleme aynı zamanda "en yakın komşuluk" olarak da adlandırılır. Bu yöntemde *x* biriminin y ve z birimlerinden oluşan yeni kümeye olan uzaklığı,

$$\mathrm{d}(\mathrm{x},\{\mathrm{y},\mathrm{z}\}) = \min(\mathrm{d}(\mathrm{x},\mathrm{y}),\mathrm{d}(\mathrm{x},\mathrm{z})) \quad (3.1)$$

eşitliği ile hesaplanır.

ii. Tam Bağlantı Kümeleme Yöntemi: ilk olarak Sorensen (1948) tarafından önerilen bu yöntemde iki farklı kümedeki birbirine en uzak gözlem çifti arasındaki uzaklık değeri, iki küme arasındaki uzaklık olarak alınır. Tam bağlantı kümeleme yöntemi aynı zamanda "en uzak komşuluk" olarak da tanımlanır. Bu yöntemde *x* biriminin y ve z birimlerinden oluşan yeni kümeye olan uzaklığı,

$$\mathrm{d}(\mathrm{x},\{\mathrm{y},\mathrm{z}\}) = \mathrm{maks}(\mathrm{d}(\mathrm{x},\mathrm{y}),\mathrm{d}(\mathrm{x},\mathrm{z})) \quad (3.2)$$

eşitliği ile hesaplanır.

iii. Ortalama Bağlantı Kümeleme Yöntemi: Ayrı gruplarda yer alan gözlem çiftleri arasındaki ortalama uzaklık iki küme arasındaki uzaklık olarak alınır. Her biri sırasıyla n ve m adet birimden oluşan K_i ve K_j kümeleri arasındaki uzaklık

$$\left(\mathrm{K_i},\mathrm{K_j}\right) = \frac{1}{\mathrm{nm}}\sum_{\mathrm{i=1}}^{\mathrm{n}}\sum_{\mathrm{j=1}}^{\mathrm{m}}\mathrm{d}\left(\mathrm{x_i},\mathrm{x_j}\right) \quad (3.3)$$

eşitliği ile tanımlanır.

iv. Merkezi Bağlantı Kümeleme Yöntemi: İki küme arsasındaki uzaklık kümelerin kendi merkezleri arasındaki uzaklık olarak alınır. K_i kümesinin merkezi $\bar{x}$ ve K_j kümesinin merkezi $\bar{y}$ olarak alındığında K_i kümesinin K_j kümesine olan uzaklığı merkez bağlantı yöntemine göre

$$\mathrm{d}\left(\mathrm{K_i},\mathrm{K_j}\right) = \mathrm{d}(\bar{\mathrm{x}},\bar{\mathrm{y}}) \quad (3.4)$$

şeklinde hesaplanır.

v. Medyan Bağlantı Kümeleme Yöntemi: Merkez bağlantı kümeleme tekniğinde iki küme bağlandığında yeni oluşan kümenin merkezi, bağlanan kümelerden birim sayısı fazla olan kümenin merkezine yakın olmaktadır (Erişoğlu, 2011). Bu durumu ortadan kaldırmak için medyan bağlantı tekniği önerilmiştir. Bu yönteme göre K_i kümesinin, K_j ve K_l kümelerinin bağlanması ile elde edilen $\{K_j \cup K_l\}$ kümesine olan uzaklığı

$$d\left(K_i, (K_j \cup K_l)\right) = \frac{1}{2} d\left(K_i, K_j\right) + \frac{1}{2} d(K_i, K_l) - \frac{1}{4} d(K_j, K_l) \quad (3.5)$$

eşitliği ile hesaplanır.

vi. Ward Bağlantı Kümeleme Yöntemi: Artırımlı (incremental) hata kareler toplamı yöntemi olarak da adlandırılan Ward'ın hiyerarşik kümeleme yönteminde amaç, artan sınıf içi hata kareler toplamının minimum yapılmasıdır. Bu yönteme göre K_i kümesinin, K_j ve K_l kümelerinin bağlanması ile elde edilen $\{K_j \cup K_l\}$ kümesine olan uzaklığı

$$d\left(K_i, (K_j \cup K_l)\right) = \frac{(n_i+n_j)d\left(K_i,K_j\right)+(n_i+n_l)d(K_i,K_l)-n_i d(K_j,K_l)}{n_i+n_j+n_l} \quad (3.6)$$

şeklinde hesaplanır (Ward, 1963).

3.2. Hiyerarşik Olmayan Kümeleme Yöntemi (K-Ortalamalar Kümeleme Yöntemi)

K-Ortalamalar Kümeleme Yöntemi kümeleme analizinde en yaygın olarak kullanılan bir kümeleme algoritmasıdır. K–Ortalamalar Kümeleme Yöntemi, hata kareler toplamının minimum yapılmasına dayalı olarak verideki en uygun parçalanmayı bulmayı amaçlayan adımsal optimizasyon yöntemidir (Erişoğlu, 2011). Bu yöntem, Hiyerarşik Kümeleme Yöntemlerinin aksine ilk aşamada oluşturulacak küme sayısının belirlenmesi ile başlar. Bu sebeple verinin parçalanması ve bu parçalanmaya göre küme merkezlerinin oluşturulması kümeleme sonucuna oldukça etki eder. Bu etkinin azaltılması için başlangıçta verinin nasıl parçalanacağına ilişkin genel bir yöntem yoktur. Bu amaçla Hiyerarşik Kümeleme Yöntemleri ile küme sayısına karar verip, K–Ortalamalar Kümeleme Yöntemi daha sonra uygulanabilir. Bu yöntemdeki aşamalar aşağıdaki gibi sıralanabilir:

1. Veri hakkında bilgi varsa bu bilgi ışığında yoksa rassal olarak başlangıç küme merkezleri seçilir. Temelde bu merkezler birbirinden uzak olan bir dizi gözlemdir.

2. Her birim merkeze olan uzaklığına göre tanımlanmış olan en yakın kümeye atanır.

3. Bu atama sonucu yeni oluşan küme merkezleri bulunur.

4. Her birimden her bir merkeze olan uzaklık tekrar hesaplanır ve kümede olmayan gözlemler en yakın kümelere taşınır.

5. Küme merkezleri nispeten kararlı kalana kadar bu süreç devam ettirilir.

4. Bulgular

Ekonomik ve işyeri araştırmalarının çoğunda, hedef değişkenlerin dağılımları (örnek: İşletmelerin cirosu, hane halkı gelirleri, vb.) yaygın olarak birçok küçük ve birkaç büyük birime sahip çarpık dağılımlara benzemektedir (Khan, Reddy, & Rao, 2015). Bu sebeple bu çalışmada gerçek bir duruma örnek olması açısından farklı çarpıklık değerlerine sahip beş farklı kitle türetilmiş ve tabaka sınırlarını belirlemek ve tahminleri en iyileştirmek için yukarıda açıklanan Kümeleme Analiz Yöntemleri ve Tabaka Sınırı Belirleme Yöntemleri karşılaştırılmıştır. Bu karşılaştırma yapılırken 7 farklı kümeleme analizi ve 3 farklı tabaka sınırı belirleme yöntemi olmak üzere, 10 farklı ayrıştırma yöntemi kullanılmıştır. Bu ayrıştırma yöntemleriyle elde edilen tabakalar belirlendikten sonra oransal dağıtım yöntemiyle örnek birimler tabakalara dağıtılmıştır. Tahminlerin güvenilirliğini arttırmak için her yöntemden 1000 defa örnek seçilmiştir. Böylece her yöntemden elde edilen ortalama tahmin edicisine ilişkin Kök Hata Kareler Ortalamaları (KHKO) elde edilmiş ve en küçük KHKO değerini veren yöntemin tabaka sınırlarının optimum sınırlar olduğu sonucuna ulaşılmıştır.

Bu bölümde aşağıdaki notasyonlar kullanılmıştır:

$k = 1, \dots, 10$	$k : Ayrıştırma\ Yöntem\ İndisi$
$j = 1, \dots, 1000$	$j :$ Örnek Sayısı
$h = 1, \dots, H$	$h : Tabaka\ Sayısı$
$i = 1, \dots, n_{kh}$	$n_{kh} : k.\ Yöntem\ ve\ h.tabaka\ için$ örnek hacmi

x_{kjhi} : k. ayrıştırma yöntemi, j. tekrar ve h. tabakadaki i. birimin değeri

$\bar{x}_{kj} = \frac{1}{N}\sum_{h=1}^{H_k} N_{kh} \frac{\sum_{i=1}^{n_{kh}} x_{kjhi}}{n_{kh}}$: $k.Yöntem\ ve\ j.tekrar\ için\ ortalama\ tahmin\ edicisi\ (TE)$

$\bar{x}_k = \frac{\sum_{j=1}^{1000} \bar{x}_{kj}}{1000}$: $k.Yöntem\ için\ ortalama\ tahmin\ edicisi$

$V(\bar{x}_k) = \frac{\sum_{j=1}^{1000} (\bar{x}_{kj} - \bar{x}_k)^2}{1000}$: $k.\ Yöntem\ için\ ortalamanın\ varyansı$

$\sqrt{V(\bar{x}_k)} = \sqrt{\frac{\sum_{j=1}^{J} (\bar{x}_{kj} - \bar{x}_k)^2}{J}}$: $k.Yöntem\ için\ ort.\ TE\ standart\ hatası$

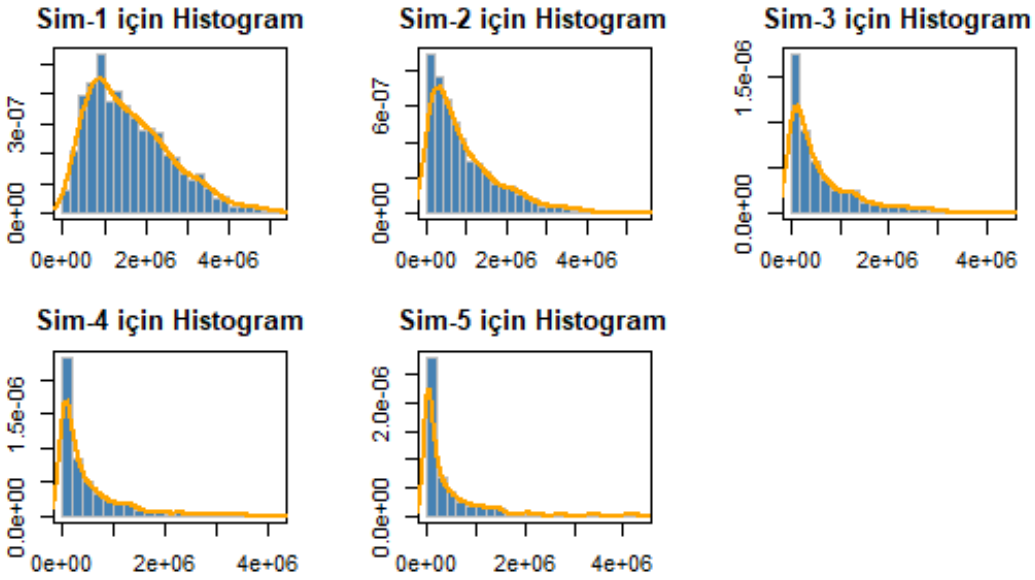
$KHKO_k = \sqrt{\bar{V}(\bar{x}_k) + (\bar{x}_k - \bar{X})^2}$: $k.Yöntem\ için$ KHKO

KHKO formülünde yer alan $\bar{X}$ kitle ortalamasıdır. Uygulamada k indisi ile gösterilen 10 ayrıştırma yönteminin 7'si kümeleme analizi 3'ü tabaka sınırı belirleme yönteminden oluşmaktadır ve aşağıdaki listede sıralanmıştır:

1. K-Ortalamalar Kümeleme Yöntemi (K-Ort)
2. Medyan Bağlantı Kümeleme Yöntemi (Medyan)
3. Merkezi Bağlantı Kümeleme Yöntemi (Merkezi)
4. Ortalama Bağlantı Kümeleme Yöntemi (Ortalama)
5. Tam Bağlantı Kümeleme Yöntemi (Tam)
6. Tek Bağlantı Kümeleme Yöntemi (Tek)
7. Ward Bağlantı Kümeleme Yöntemi (Ward)
8. Birikimli Kök Frekans Yöntemi (Cumrootf)
9. Geometrik Yöntem (Geo)
10. Lavallee-Hidiroglou Yöntemi (LH)

Uygulamada j indisi ile gösterilen tekrar sayısı örnek tahminlerinin hassasiyetini arttırmak için önsel olarak 1000 ile sınırlandırılmıştır. 100, 500 ve 1000 tekrarla denemeler yapılmış ve 1000 tekrarın yeterli olduğuna karar verilmiştir.

Veri setleri R paket programında standart tanımı *rbeta (n, shape1, shape2, ncp=0)* şeklinde olan *rbeta* formülü ile yaratılmıştır. Bu formül beta dağılımına sahip n birimden oluşan rassal veri yaratmak için kullanılır. Beta dağılımı [0,1] aralığında iki tane pozitif şekil parametresi (shape1 ve shape2) ile ifade edilmiş bir sürekli olasılık dağılımı ailesidir (Kennedy, 1988). Bu formülde gözlem sayısı "n", beta dağılım parametreleri "shape1", "shape2" ve opsiyonel olan merkeziyetsizlik parametresi (non-centrality parameter) "ncp" ile gösterilmektedir. Uygulamadaki amacımız farklı çarpıklığa sahip veriler türetmek olduğu için, shape1 parametresine sırasıyla 2, 1, 0.5, 0.25 ve 0.2 değerleri; shape2 parametresine ise bütün veri setlerinde sabit olacak şekilde 10 değeri atanmıştır. Ayrıca tabaka sınırı belirleme yöntemlerinden olan Geometrik Yöntem minimum değeri sıfıra yakın olan verilerde iyi sonuçlar vermediği için beta dağılımı ile türetilen veriler 10 000 000 değeri ile çarpılmıştır. Bu şekilde elde edilmiş 5 ayrı veri setine ait histogram görselleri ve özet istatistik bilgiler Şekil 1 ve Tablo 2'de özetlenmiştir.



Şekil 1. Oluşturulan Veri Setlerine Ait Histogram Görselleri

Tablo 2. Oluşturulan Veri Setlerine Ait Temel İstatistiki Göstergeler

	Sim-1	Sim-2	Sim-3	Sim-4	Sim-5
Birim Sayısı	1000	1000	1000	1000	1000
Medyan	1 455 129.95	671 381.80	362 454.47	231 387.71	156 643.89
Ortalama	1 642 935.46	955 930.56	628 164.23	484 826.27	397 109.25
Min	86 777.06	960.38	19.84	0.46	0.01
Maks	5 160 506.73	5 291 443.90	4 232 040.98	4 136 807.14	4 282 776.29
Standart Sapma	1 005 070.57	889 196.69	710 045.05	651 775.55	601 677.32
Çarpıklık	0.82	1.48	1.70	2.24	2.86

Yukarıda histogram görsellerinden ve tablodaki çarpıklık değerlerinden de görüldüğü gibi çarpıklık değeri simülasyon birden beşe doğru artmaktadır. Bunun amacı, farklı çarpıklık değerleri için yöntemlerin performanslarını karşılaştırmaktır.

Literatürde hem türetilmiş hem de gerçek veri setleri kullanılarak yapılan çalışmalar vardır. Gunning & Horgan, 2004 çalışmasında farklı çarpıklık değerine sahip 4 veri seti üzerinden uygulamalarını yapmışlardır. Bu veri setleri; bir İrlanda firmasının borçlarına ait muhasebe kayıtları, Amerika Birleşik Devletleri'ndeki şehir nüfusları, 4 yıllık Amerika Birleşik Devletleri kolejlerindeki öğrenci sayıları ve Amerika Birleşik Devletleri'ndeki büyük ticari bankaların kaynaklarıdır. Khan, Reddy, & Rao, 2015 ise kendi çalışmalarında Fiji'deki çiftçilerin şeker kamışı üretim verilerini elde edebilmek için yardımcı değişken olarak hektar cinsinden şeker kamışı ekim alanı verisini kullanmıştır. Hidiroglou & Kozak, 2017 çalışmasında hem gerçek veri seti üzerinden bir uygulama yapmış hem de Slanta & Krenzke, 1996 çalışmasına atıfta bulunarak, bu çalışma bulgularını paylaşmıştır. Slanta & Krenzke, 1996 ise kendi çalışmalarında 5 farklı örnek hacmine sahip F(x)=(j-1/2)/N formülü ile türetilmiş veri seti üzerinde, Lavallee ve Hidiroglou yöntemini performansını sınamışlardır (Slanta & Krenzke, 1996).

Kümeleme analizi için R paket programında bulunan "*NbClust*" paketi kullanılmıştır (Charrad, Ghazzali, Boiteau, & Niknafs, 2014). Kümeleme analizi için kullanılan formülün standart formu aşağıdaki gibidir:

NbClust(data, diss= NULL, distance= "euclidean", min.nc= 2, max.nc=15, method = "NULL", index = "all", alphaBeale = 0.1) (4.1)

Burada "data" kullanılacak veri setini, "diss" benzerlik matrisini, "distance" uzaklık ölçüsünü, "min.nc" oluşturulabilecek en küçük küme sayısını, "max.nc" oluşturulabilecek en büyük küme sayısını, "method" kullanılacak analiz yöntemini, "index" kullanılacak kümeleme sayısı belirleme kriterini ve "alphaBeale" ise küme sayısı belirleme kriterlerinden biri olan Beale endeksi için önem düzeyini ifade etmektedir, varsayılan ayarında değeri 0.1'dir. Benzerlik matrisi tanımlanmamışsa, uzaklık ölçüsü olarak "euclidean", "maximum", "manhattan", "canberra", "binary", "minkowski" uzaklık ölçülerinden biri seçilmelidir, varsayılan ayar Öklid uzaklık ölçüsüdür. Uygulanacak kümeleme analiz yöntemleri ise "ward.D", "ward.D2", "single", "complete", "average", "mcquitty", "median", "centroid", "kmeans" seçenekleri arasından seçilmelidir. Küme sayısı belirleme kriterleri ise Charrad, Ghazzali, Boiteau ve Niknafs (2014)'de tanımlanmış 30 farklı yöntemden oluşmaktadır. Bu yöntemlerin içinden bir tane ya da tamamı belirleme yöntemi olarak seçilebilir. Örneğin "index=CCC" yazılırsa kübik kümeleme kriterine (Cubic Clustering Criterion) göre en uygun küme sayısı seçilir. "index=all" seçeneğinde 30 yöntem için de uygun küme sayısı belirlenir, bu durumda ise en çok yöntem tarafından seçilen küme sayısı, en uygun küme sayısı olarak belirlenmiş olur.

Bu çalışmada kümeleme analizi için gerçekleştirilen uygulamada, uzaklık ölçüsü olarak Öklid Uzaklığı, küme büyüklükleri için minimum değer olarak 2, maksimum değer olarak 10 değerleri, küme sayısı belirleme kuralları ise "index=all" seçeneği seçilmiştir. Böylece bu küme değerleri arasında uygun küme sayıları daha önceden tanıtılan 7 kümeleme analiz yöntemi için hesaplanmıştır.

Tabaka sınırı belirleme yöntemleri için ise R paket programında bulunan "*stratification*" paketi ile uygulama gerçekleştirilmiştir. Bu pakette yer alan formüllerin varsayılan ayarları aşağıda gösterilmiştir. Birikimli Kök Frekans Yöntemi için 4.2, Geometrik Yöntem için 4.3 ve Lavalle-Hidiroglou Yöntemi için 4.4 formül geçerlidir.

strata.cumrootf(x, n = NULL, CV = NULL, Ls = 3, certain = NULL, alloc = list(q1 = 0.5, q2 = 0, q3 = 0.5), rh = rep(1, Ls), model = c("none", "loglinear", "linear", "random"), model.control = list(), nclass = NULL) (4.2)

strata.geo(x, n = NULL, CV = NULL, Ls = 3, certain=NULL, alloc = list(q1 = 0.5, q2 = 0, q3 = 0.5), rh = rep(1, Ls), model = c("none", "loglinear", "linear", "random"), model.control = list()) (4.3)

strata.LH(x, n = NULL, CV = NULL, Ls = 3, certain = NULL, alloc = list(q1 = 0.5, q2 = 0, q3 = 0.5), takenone = 0, bias.penalty = 1, takeall = 0, rh = rep(1, Ls), model = c("none", "loglinear", "linear", "random"), model.control = list(), initbh = NULL, algo = c("Kozak", "Sethi"), algo.control = list()) (4.4)

Burada "x" tabakalama değişkeni vektörünü, "n" örnek hacmini, "CV" değişim katsayı değerini, "Ls" tabaka sayısını, "certain" vektör x'te yer alan ve örnekte seçilmesi kesin olan birimleri, "alloc" örnek birimlerin tabakalara dağıtım yöntemini belirleyen katsayıları, "rh" her tabakadaki beklenen cevaplama oranını, "model" tabaklama değişkeni X ile araştırma değişkeni Y arasındaki farklılığı tanımlayacak modeli, "model.control" model parametrelerini, "nclass" *Birikimli Kök Frekans Yöntemi* için sınıf sayısını, "takenone" örnek birim seçilmeyecek tabaka sayısını, "takeall" tabakadaki bütün birimlerin örnek seçileceği tabaka sayısını, "bias.penalty" "takenone=1" olduğunda araştırma tahmin edicisinin beklenen hata kareler ortalamasının hesabında kullanılan 0 ile 1 arasında değer alan ceza parametresini, "initbh" başlangıç tabaka sınırlarını, "algo" seçilecek optimizasyon algoritmasını ve "algo.control" ise optimizasyon algoritmasını kontrol edecek parametreleri göstermektedir. Bu formüllerde yer alan birçok seçenek, tabakalama değişkeni ile araştırma değişkeninin farklı olduğu durumlar için sunulmuştur. Bu çalışmada böyle bir ayrım yapılmadığı için sadece kullanılan alternatiflere ait bilgiler verilmiştir. Formüllerde "n" ya da "CV" den bir tanesi seçilmelidir. Örnek hacmi seçilmişse değişim katsayı değeri, değişim katsayısı seçilmişse örnek hacmi seçilmez. Bu çalışmada CV değer 0,01 alınmıştır. Örnek birimlerin tabakalara dağıtımı için Hidiroglou ve Srinath (1993)'ta açıklanan 3 sayısal değer yer almalıdır. Buna göre;

- q1= 0,5 ve q2=q3=0 ise oransal dağıtım,
- q1=q2=p/2 ve q3=0 ise güç dağıtım,
- q1=q3=0,5 ve q2=0 ise Neyman dağıtım yöntemini ifade etmektedir.

Tabaka sınırı belirleme yöntemlerinin dezavantajı tabaka sayısının önceden belirlenmesidir. "Ls" yani tabaka sayısı bu sebeple dolu olmalıdır. Bu çalışmada yukarıda adı geçen 3 tabaka sınırı belirleme yöntemi için 3'ten 7'ye kadar tabaka sayıları önsel olarak belirlenmiş ve veri setleri 5 farklı tabaka sayısı için tabakalara ayrılmıştır. Bu işlemler sonucunda 10 ayrıştırma yöntemi için (7'si kümeleme 3'ü tabaka sınırı belirleme yöntemi) tabaka büyüklükleri belirlenmiştir. Bu aşamadan sonra örnek hacmi belirlenip, örnekler tabakalara dağıtılıp veri setlerine ilişkin ortalama tahminlerinin KHKO'ları karşılaştırılmıştır. Bu uygulama için örnek hacim büyüklüğü 200 olarak belirlenmiştir. Örneklerin tabakalara dağıtımı oransal dağıtım yöntemiyle yapılmıştır. Yukarıdaki yöntemler için örneklerin tabakalara dağılımı belirlendikten sonra örnek seçim işlemi Tabakalı Basit Tesadüfi Örnekleme Yöntemiyle yapılmış, daha önce de değinildiği üzere örnek tekrar sayısı 1000 olarak belirlenmiştir. Sonuç olarak her bir yöntemde 1000 defa örnek seçilmiş ve her bir yönteme ait ortalama tahmini ve ortalamanın varyans tahminlerine ulaşılmıştır. Bu 10 yöntemi birbiri ile kıyaslamak için hesaplanan KHKO değerleri Tablo 3'te verilmiştir:

Tablo 3. Yöntemlerin KHKO Değerleri

Yöntem	Sim-1	Sim-2	Sim-3	Sim-4	Sim-5
Ortalama	31 113.80	24 162.47	18 705.67	5 905.44	8 239.33
Merkezi	30 329.63	16 160.24	18 818.21	8 800.44	12 586.51
Tam	25 994.18	19 412.70	8 911.38	16 376.18	6 634.93
K-Ortalamalar	34 985.42	22 557.16	16 424.74	15 175.73	5 970.40
Medyan	22 273.20	15 056.90	18 941.25	21 072.09	15 034.48
Tek	61 580.44	45 483.98	24 152.47	28 521.64	28 028.16
Ward	25 279.13	13 669.21	16 237.64	11 772.29	9 231.71
cumrootf_3	25 585.63	22 662.02	17 576.92	17 047.95	17 460.59
cumrootf_4	19 762.29	18 006.49	12 772.06	13 338.31	15 434.84
cumrootf_5	16 310.49	14 401.40	10 439.38	10 386.40	11 249.25
cumrootf_6	13 582.09	11 572.44	9 356.08	9 329.73	10 908.58
cumrootf_7	11 756.70	11 040.84	7 486.96	8 180.40	8 854.51
Geo_3	40 128.71	46 157.35	40 203.60	38 610.95	35 606.67
Geo_4	31 710.79	38 756.59	36 688.03	35 514.85	35 714.71
Geo_5	25 678.77	32 962.36	32 613.52	33 475.64	33 305.12
Geo_6	21 735.49	27 582.34	27 279.27	30 168.55	29 867.17
Geo_7	18 690.07	24 479.32	24 781.22	27 248.03	26 280.88
LH_3	24 814.91	22 628.00	16 660.09	15 004.08	14 628.92
LH_4	18 502.59	16 950.56	12 282.63	11 805.17	10 903.23
LH_5	15 736.37	13 489.36	9 720.74	8 738.27	8 754.91
LH_6	13 444.24	11 853.11	8 934.04	7 579.56	6 909.30
LH_7	11 240.72	10 335.59	7 252.25	6 604.77	6 311.94

Tablo 3'te oransal dağıtım ile örnek dağıtımı yapılan, kümeleme ve tabaka sınırı belirleme yöntemlerine göre tabaka sınırı belirlenen beş farklı çarpıklığa sahip veri seti için elde edilen ortalama tahminine ilişkin KHKO'ları verilmiştir. Kırmızı ile boyanmış hücreler her veri seti için en düşük %10'luk değerleri göstermektedir. Buna göre;

- Çarpıklık derecesi 0.82 olan 1. veri setinde; Lavallee-Hidiroglou Yöntemi 11240.72 değeri ile en küçük KHKO değerine sahiptir. Bu sebeple 1. veri seti için optimum tabaka sayısı 7 olarak belirlenmiştir. Buna göre 1. Veri seti için optimum tabaka sınırları; Lavallee-Hidiroglou Yönteminde 7 tabaka için hesaplanan sınırlardır. Bu sınırlar; 0-700000, 700000-1200000, 1200000-1750000, 1750000-2400000, 2400000-3100000, 3100000-3900000, 3900000-ve üzeri şeklinde hesaplanmıştır.

- Çarpıklık derecesi 1.48 olan 2. veri setinde; Lavallee-Hidiroglou Yöntemi 10335.59 değeri ile en küçük KHKO değerine sahiptir. Bu sebeple 2. veri seti için optimum tabaka sayısı 7 olarak belirlenmiştir. Buna göre 2. Veri seti için optimum tabaka sınırları; Lavallee-Hidiroglou Yönteminde 7 tabaka için hesaplanan sınırlardır. Bu sınırlar; 0-350000, 350000-750000, 750000-1200000, 1200000-1750000, 1750000-2400000, 2400000-3300000, 3300000-ve üzeri şeklinde hesaplanmıştır.

- Çarpıklık derecesi 1.7 olan 3. veri setinde; Lavallee-Hidiroglou Yöntemi 7252.25 değeri ile en küçük KHKO değerine sahiptir. Bu sebeple 3. veri seti için optimum tabaka sayısı 7 olarak belirlenmiştir. Buna göre 3. Veri seti için optimum tabaka sınırları; Lavallee-Hidiroglou Yönteminde 7 tabaka için hesaplanan sınırlardır. Bu sınırlar; 0-220000, 220000-520000, 520000-850000, 850000-1300000, 1300000-18500000, 1850000-2600000, 2600000-ve üzeri şeklinde hesaplanmıştır.

- Çarpıklık derecesi 2.23 olan 4. veri setinde; Ortalama Bağlantı Kümeleme Yöntemi 5905.44 değeri ile en küçük KHKO değerine sahiptir. Bu sebeple 4. veri seti için optimum tabaka sayısı 9 olarak belirlenmiştir. Buna göre 4. Veri seti için optimum tabaka sınırları; Ortalama Bağlantı Kümeleme Yöntemi ile hesaplanan sınırlardır. Bu sınırlar; 0-220000, 220000-620000, 620000-1000000, 1000000-1500000, 1500000-2100000, 2100000-2400000, 2400000-3100000, 3100000-4100000, 4100000-ve üzeri şeklinde hesaplanmıştır.

- Çarpıklık derecesi 2.86 olan 5. veri setinde ise; K-Ortalamalar Kümeleme Yöntemi 5970.4 değeri ile en küçük KHKO değerine sahiptir. Bu sebeple 5. veri seti için optimum tabaka sayısı 7 olarak belirlenmiştir. Buna göre 5. Veri seti için optimum tabaka sınırları; K-Ortalamalar Kümeleme Yöntemi ile hesaplanan sınırlardır. Bu sınırlar; 0-150000, 150000-400000, 400000-700000, 700000-1100000, 1100000-18000000, 1850000-2800000, 2800000-ve üzeri şeklinde hesaplanmıştır.

5. Tartışma ve Sonuç

Bu çalışmada 1000 birimden oluşan farklı çarpıklık değerlerine sahip beş farklı kitle türetilmiştir. Tabakalı Tesadüfi Örneklemede kullanılmak üzere bu kitlelerin tabaka sınırlarını belirlemek ve tahminleri en iyileştirmek için Kümeleme Analiz Yöntemleri ve Tabaka Sınırı Belirleme Yöntemleri karşılaştırılmıştır. Bu karşılaştırma yapılırken 7 farklı kümeleme analizi, 3 farklı tabaka sınırı belirleme yöntemi olmak üzere 10 farklı ayrıştırma yöntemi kullanılmıştır. Tabakalar belirlendikten sonra oransal dağıtım yöntemiyle örnek birimler tabakalara dağıtılmıştır. Tahminlerin güvenilirliğini arttırmak için her yöntemden 1000 defa örnek seçilmiş, bu örneklerin ortalama tahminlerine ilişkin KHKO değerleri karşılaştırılmış ve sonuçları sunulmuştur.

Lavallee-Hidiroglou Yöntemi bir optimizasyon yöntemi olduğu için 5 veri setinin 3'ünde en başarılı ayrıştırmayı sağlamıştır. En düşük çarpıklığa sahip ilk üç veri setinde Lavallee-Hidiroglou Yöntemi, optimum tabakalamayı sağlamışken, 4. veri setinde Ortalama Bağlantı Kümeleme Yöntemi, en çarpık veri setinde ise K-Ortalamalar Kümeleme Yöntemi en iyi tabakalamayı yapmıştır. Çarpıklık değerleri arttıkça kümeleme analiz yöntemlerinin başarısının artması, çarpık dağılıma sahip veri setlerinde, kümeleme analiz yöntemlerinin de tabaka sınırı belirleme çalışmalarında kullanılabileceğini göstermektedir. Ayrıca özellikle hiyerarşik kümeleme yöntemlerinde küme sayısının önsel olarak belirlenmemesi, tabaka sınırı belirleme yöntemlerine kıyasla bir üstünlük olarak değerlendirilebilinir. Buna rağmen hiyerarşik kümeleme yöntemlerinin uygulama mantığı gereği büyük veri setleri için maliyetli olmaları bir dezavantaj olarak kabul edilmelidir.

Sonuç olarak bu çalışma, çeşitli çarpıklık dağılımlarına göre tabakalı tesadüfi örneklemede kullanılan tabaka sınırlarının belirlenmesini karşılaştıran öncü bir çalışmadır. Bu kapsamda özellikle hiyerarşik olmayan K-ortalamalar yöntemine ait bulgular dikkat çekicidir. Çalışmanın bundan sonraki aşamalarında, hanehalkı/iş yeri düzeyinde farklı tahmin edicilere göre model denemeleri ve performans karşılaştırmaları yapılabilir. Bu makale yazarın doktora tez çalışmasının bir kısmını oluşturmaktadır ve tez çalışmasında hem gerçek veri seti üzerinde hem de diğer tabaka dağıtım yöntemleri için çalışma gerçekleştirilmiştir.

Kaynakça

Ballin , M., & Barcaroli, G. (2013). Joint determination of Optimal Stratification and Sample Allocation Using Genetic Algorithm. *Survey Methodology*, 369-393.

Benedetti, R., Bee, M., & Espa, G. (2010). A Framework for Cut-Off Sampling in Business Survey Design. *Journal of Official Statistics*, 651-671.

Brito, J., Ochi, L., Montenegro, F., & Maculan, N. (2010). An Iterative Local Search Approach Applied to the Optimal Stratification Problem. *Int. Trans. Oper. Res.*, 753-764.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 1-36.

Dalenius, T., & Hodges, J. (1957). The Choice of Stratification Points. *Scand. Actuar. Journal*, 198-203.

Ekman, G. (1959). An Approximation Useful in Univariate Stratificaiton. *Ann. Math. Stat.*, 219-229.

Erişoğlu, M. (2011). Uzaklık Ölçülerinin Kümeleme Analizine Olan Etkilerinin İncelenmesi ve Geliştirilmesi.

Florek, K., Lukaszewicz, J., Steinhaus, H., & Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum*, 282-285.

Gunning, P., & Horgan, J. (2004). A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. *Survey Methodology*, 159-166.

Hidiroglou, M., & Kozak, M. (2017). Stratification of Skewed Populations: A Comparison of Optimisation-based Versus Approximate Methods. *International Statistics Review*, 1-19.

Hidiroglou, M., & Srinath, K. (1993). Problems Associated with Designing Subannual Business Surveys. *Journal of Business & Economic Statistics*, 397-405.

Kennedy, D. (1988). A Note On Stochastıc Search Methods For Global Optimization. *Advances in Applied Probability*, 476-478.

Keskintürk, T., & Er, S. (2007). A Genetic Algorithm Approach to Determine Stratum Boundaries and Sample Sizes of Each Stratum in Stratified Sampling. *Comput. Stat. Data Anal.*, 53-67.

Khan, M. G., Reddy, K. G., & Rao, D. (2015). Designing Stratified Sampling in Economic and Business Surveys. *Journal of Applied Statistics*, 1-20.

Kozak, M. (2004). Optimal Stratification Using Random Search Method in Agricultural Surveys. *Stat. Transition*, 797-806.

Lavallee, P., & Hidiroglou, M. A. (1988). On the Stratification of Skewed Populations. *Survey Methodology*, 33-43.

Lednicki, B., & Wieczorkowski, R. (2003). Optimal Stratification and Sample Allocation Between Subpopulations and Strata. *Stat. Transition*, 287-306.

Mahalanobis, P. (1952). Some Aspects of the Design of Sample Surveys. *The Indian Journal of Statistics*, 1-7.

Rivest, L. (2002). A Generalization of the Lavallee and Hidiroglou Algorithm for Stratification in Business Surveys. *Survey Methodology*, 191-198.

Servi, T. (2009). Çok Değişkenli Karma Dağılım Modeline Dayalı Kümeleme Analizi.

Sethi, V. (1963). A Note on the Optimum Stratification of Populaitons for Estimating the Population Means. *Australian J. Stat.*, 20-33.

Singh, R. (1971). Approximately Optimum Stratification on the Auxiliary Variable. *J. Am. Stat. Assoc.*, 829-833.

Slanta, J., & Krenzke, T. (1996). Applying the Lavallée and Hidiroglou Method to Obtain Stratification Boundaries for the Census Bureau's Annual Capital Expenditure Survey. *Survey Methodology*, 65-75.

Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, 1-34.

Sweet, E., & Sigman, R. (1995). Evaluation of Model-Assisted Procedures for Stratifying Skewed Populaitons Using Auxiliary Data. *U.S. Bureau of the Census*, 1-9.

Thomsen, I. (1976). A Comparison of Approximately Optimal Stratification Given Proportional Allocaiton with Other Methods of Stratification and Allocaiton. *Metrika*, 15-25.

Verma, M. R., Kozak, M., & Zielinski, A. (2007). Modern Approach to Optimum Stratification. *Review and Perspectives*, 223-250.

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