İL NÜFUS VE VATANDAŞLIK MÜDÜRLÜKLERİNİN İŞ YOĞUNLUĞUNA GÖRE HİBRİD KÜMELEME İLE SINIFLANDIRILMASI[*]

Yrd. Doç. Dr. Noyan AYDIN[**]

Ayşe Nur SEVEN[***]

ÖZ

Kamu Kurumları, ilgili görevlerini yerine getirebilmek için değişen nicelikte ve nitelikte personeli istihdam etmelidir. Ancak, istihdam edilen personel sayısının hizmetin verileceği ilin veya bölgenin nüfusu ile her zaman aynı paralellikte olmadığı görülmektedir. Bu durumda, personellerin iradi ya da gayriiradi olarak yer değiştirmesi, geçici nüfus sirkülasyonu ve e-hizmet uygulamasının artışı gibi faktörler etkili olmaktadır. Bu çalışmada, Türkiye'deki İl Nüfus ve Vatandaşlık Müdürlükleri, iş yoğunluklarına göre hibrid hiyerarşik k-ortalamalar kümeleme analizi ile sınıflandırılmıştır. Küme sayısına karar verirken silhouette endeksinden yararlanılmıştır. Analiz sonucunda, benzer iş yoğunluğuna sahip illerden oluşan altı farklı küme yapısı ortaya çıkarılmıştır. Elde edilen küme yapılarının geçerliliği ilgili istatistiki testler yardımıyla da desteklenmiştir.

Anahtar Kelime: *Hibrid Kümeleme, Hiyerarşik K-ortalamalar, Kamu Kurumları İş Yoğunluğu.*

JEL Sınıflandırması: *C38, Z18.*

CLASSIFICATION BASED ON WORK INTENSITY BY HYBRID CLUSTERING OF THE PROVINCIAL DIRECTORATES OF CIVIL REGISTRATION AND NATIONALITY

ABSTRACT

Public institutions must employ staff in varying number and qualification to fulfill their related responsibilities. However, it is observed that the number of staff are employed and the population of the province or region are not in the same parallels. In this case, some factors are effective: such as the displacement of the staff as voluntary or involuntary, temporary circulation of the population, the increase in electronic applications. In this study, the provincial directorates of civil registration and nationality in Turkey were classified according to their work intensity by hybrid hierarchical k-means clustering. Silhouette index was used to determine the number of clusters. According to the results, it was revealed six different cluster structure consisting of the provinces have a similar work intensity. The validity of the resulting cluster structure was supported with the help of relevant statistical tests.

Keywords: *Hybrid Clustering, Hierarchical K-means, Work Intensity of Public Institutions.*

JEL Classification: *C38, Z18.*

[*] Çalışma, halen danışmanlığı yürütülen 2.yazarın tez konusundan hareketle farklı bir yıl (2013) ve farklı bir kümeleme tekniği kullanılarak yapılmış, 15. Uluslararası EYİ Sempozyumu'nda (22-25 Mayıs 2014, Süleyman Demirel Üniversitesi, Isparta) bildiri olarak sunulmuştur.

[**] T.C. Dumlupınar Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Ekonometri Bölümü, noyan.aydin@dpu.edu.tr

[***] T.C. İçişleri Bakanlığı, Kütahya İl Nüfus ve Vatandaşlık Müdürlüğü, a.nur.seven@icisleri.gov.tr

1. GİRİŞ

Kamu kurum ve kuruluşlarının, kendi hizmet sahalarıyla ilgili olarak yerine getirmekle mükellef oldukları çeşitli görevleri bulunmaktadır. Bu görevlerin yerine getirilmesinde de değişen sayıda makine ve teçhizat ile insan gücüne ihtiyaç duymaktadırlar. Bu çerçevede, konuya insan gücü yani personel ihtiyacı açısından bakıldığında da bir insan kaynakları planlamasına ve dolayısıyla da mevcut iş yoğunluğuna göre gerekli kadro tahsisisin yapılmasına ihtiyaç duyulmaktadır. Bu durum da, iş yoğunluğu fazla olan kamu birimlerinde iş yoğunluğu daha az olanlara göre daha fazla sayıda personelin istihdam edilmesi gerekliliğini ortaya çıkarmaktadır. Bu doğrultuda, merkezi kamu otoritesinin ilgili birimlere gerekli personel atamalarını yapabilmesi aşamasında benzer iş yoğunluğuna sahip kamu hizmet birimlerinin sınıflandırılması ihtiyacı gündeme gelmektedir. Bu çalışmada, bir kamu kurumu olarak on bir ana hizmet kaleminde faaliyet gösteren İl Nüfus ve Vatandaşlık Müdürlükleri iş yoğunluklarına göre kümeleme analizi ile sınıflandırılmak istenmektedir.

Kümeleme analizi, doğal grupları (sınıfları) kesin olarak bilinmeyen birimleri ve/veya değişkenleri birbirleriyle benzer olan alt kümelere (gruplara, sınıflara) ayırmaya yardımcı olan yöntemler topluluğudur. Kümeleme analiziyle elde edilen sınıflandırma neticesinde, küme içi benzerlik (homojenlik) ve kümeler arası farklılık (hetorojenlik) en büyüklenmekte ve buradan hareketle ortak özellikler gösteren birimler ve/veya değişkenler hakkında genel tanımlamalar ve özet bilgiler elde edilmeye çalışılmaktadır (Özdamar, 1999). Ayrıca unutulmamalıdır ki, bu analizle birlikte birimler veya değişkenler arasındaki kesin bir ilişkinin var olup olmadığının araştırılmasından ziyade, eldeki çok değişkenli veri kümesine ilişkin açıklayıcı bir analizin ortaya koyulması amaçlanmaktadır.

Kümeleme analizi, kullanılan yöntemlere göre çeşitli şekillerde sınıflandırılabilmektedir. Buna göre kümeleme analizi temel olarak,

- Hiyerarşik kümeleme
- Hiyerarşik olmayan (parçalayıcı) kümeleme

analizi şeklinde ikiye ayrılmaktadır. Hiyerarşik kümeleme, kendi içinde birleştirici (agglomerative) ve ayrıştırıcı (divisive) kümeleme olarak ikiye ayrılmaktayken, parçalayıcı (partitioning) kümeleme analizi ise aşağıdaki gibi çeşitli alt gruplara ayrılabilmektedir (Singh ve Singh, 2012; Syal, Prasad ve Kumar, 2012):

- Merkez tabanlı kümeleme (k-ortalamalar, k-medyan, k-harmonik ve k-mod vb.)
- Araştırma (search) tabanlı kümeleme (genetik algoritma, genetik bulanık k-mod vb.)
- Yoğunluk (density) tabanlı kümeleme (DBSCAN, BRIDGE, DENCLUE vb.)
- Izgara (grid) tabanlı kümeleme (STING, GRIDCLUS, WAVECLUSTER vb.)
- Grafik tabanlı kümeleme (CACTUS, ROCK vb.)
- Model tabanlı kümeleme (Gaussian, EM, COOLCOT, STUCCO vb.)

Bu yöntemlerin bazı avantaj ve dezavantajlara sahip olması sebebiyle çeşitli hibrid (melez-karma) ve bulanık (fuzzy) kümeleme yöntemleri de geliştirilmiştir. Hibrid yöntemlerde, avantajlı oldukları yönlerinin kullanılıp dezavantajlı oldukları yönlerinin bertaraf edilebileceği en az iki kümeleme yönteminin bir arada kullanımı söz konusudur (Chen, Rhodes, Kline, ve Irvin, 2010).

Bu çalışmada, birleştirici hiyerarşik kümeleme ile parçalayıcı kümeleme yöntemlerinden biri olan k-ortalamalar (k-means) yönteminin uygun bir şekilde bir araya getirilmesiyle elde edilen *"hibrid hiyerarşik k-ortalamalar yöntemi"* kullanılarak Türkiye'deki İl Nüfus ve Vatandaşlık Müdürlükleri iş yoğunluğu bazında kümeleme analizi ile sınıflandırılacaktır. Analizin yapılmasında, SPSS 21 ve STATA 12 paket programlarından yararlanılacaktır.

2. LİTERATÜR

Genel olarak hibrid kümeleme ile ilgili olarak özellikle 2000'lerin ikinci yarısından itibaren pek çok çalışma yapılmıştır. Aşağıda hibrid kümeleme yöntemi kullanılarak yapılmış bazı çalışmalara ait ana hatlar yer almaktadır.

Chen, Tai, Harrison ve Pan (2005), DNA(gen) çiplerinin sınıflandırılması amacıyla yeni bir hibrid kümeleme yöntemi olan *"hybrid k-means"* tekniğine çalışmalarında yer vermişlerdir. Bu çalışmada, ortalama (average-within groups) küme oluşturma algoritmasını kullanan birleştirici hiyerarşik kümeleme yöntemi ile k-ortalamalar algoritmasını kullanan ve hiyerarşik olmayan (partitional-parçalayıcı) kümeleme yöntemleri, korelasyon ve Öklid uzaklık ölçütlerine göre sentezlenmiş; hibrid kümelemenin kümeleme kalitesi ve aykırı gözlemler açısından iki klasik yönteme kıyasla daha iyi sonuçlar verdiği ortaya koyulmuştur.

Vijaya, Murty ve Subramanian (2006), protein dizilerinin sınıflandırılması için birleştirici hibrid hiyerarşik kümeleme olarak adlandırılan "*bottom-up hybrid hierarchical clustering (BHHC)"* tekniğini ilgili çalışmalarında kullanmışlardır. İki aşamalı olan bu yöntemde, birleştirici hiyerarşik kümeleme tekniği ile k-medyan parçalayıcı kümeleme tekniklerinden faydalanılmış ve karşılaştırmalı olarak etkin bir sonuç elde edildiği görülmüştür.

Cao, Liang ve Bai (2009), kategorik veriler için yapmış oldukları hibrid kümeleme çalışmasında, başlangıç küme merkezlerinin seçiminde k-mod ve bulanık k-mod algoritmaları kullanılmış; görgül sonuçlar neticesinde, başlangıç küme merkezlerinin rassal olarak seçildiği k-ortalamalar tekniğine kıyasla daha etkin sonuçlar elde edildiği görülmüştür.

Murugesan ve Zhang (2011), bazı haber kaynaklarından elde edilmiş olan belgelerin sınıflandırılmasında hibrid bisect k-ortalamalar kümeleme tekniğini ilgili çalışmada kullanmışlardır. Kümelemenin ilk aşamasında, bisect parçalayıcı k-ortalamalar tekniği ile kümeleme başlangıç merkezleri elde edilmiş, daha sonra bu centroidler ortalama (average-within grup) bağlantı algoritmasını kullanan birleştirici hiyerarşik kümeleme tekniğinde yeniden kümeleme işlemine tabi

tutulmuştur. Hem hibrid bisect k-ortalamalar tekniği ile hem de klasik bisect k-ortalamalar tekniği ile elde edilen kümeleme sonuçları, bazı küme doğrulama ölçütleri ile karşılaştırılmış ve hibrid yöntemin daha iyi sonuçlar verdiği görülmüştür.

Syal vd. (2012), Goodall uzaklık ölçütü temelinde kategorik veriler için bir hibrid kümeleme çalışması yapmıştır. Çalışmanın ilk aşamasında k-ortalamalar parçalayıcı kümeleme tekniğinde başlangıç küme merkezleri olarak kullanılmak üzere k-mode algoritması uygulanmıştır. Daha sonra k-ortalamalar tekniği ile elde edilmiş alt kümelerin belirli sayıdaki en büyük gözlemleri birleştirilerek birleştirici hiyerarşik küme analizi uygulanmış ve etkin sonuçlar elde edilmiştir.

3. METODOLOJİ

3.1 Uzaklık Ölçüsünün Seçimi

Çok değişkenli istatistiksel yöntemlerden olan ve n gözlem ile p değişkeni içeren kümeleme analizinde, amacımız doğrultusunda uzaklık (benzemezlik) veya benzerlik ölçülerinden yararlanarak çözüm süreci başlatılmaktadır. Bu doğrultuda amacımız birbirine yakın gözlemlerin ve farklı gözlem gruplarından oluşacak kümelerin belirlenmesi ise, n x n boyutlu matris aracılığıyla uzaklık ölçülerinden; benzer şekilde amacımız değişkenlerin kümelenmesi ise, p x p boyutlu matris aracılığıyla benzerlik ölçülerinden yararlanılır. Çok sayıda uzaklık ve benzerlik ölçüsü bulunmakta olup, söz konusu ölçünün seçiminde eldeki verinin türü dikkate alınmaktadır. Örneğin veriler, oransal ya da aralıklı ölçekle elde edilmiş ise Öklid, kare Öklid, Chebychev ve Mahattan City-Block gibi uzaklık ölçülerinden veya Pearson korelasyon katsayısı ve Kosinüs benzerlik ölçüsü gibi benzerlik ölçülerinden yararlanılır. Diğer taraftan veriler, sıklık (sayma sayısı) şeklinde ölçülmüşse ki-kare veya phi-kare; ikili (binary) şeklinde ölçülmüş ise de Öklid, kare-Öklid, Lance-Williams, büyüklük farkları veya durum uzaklık ölçüsü gibi uzaklık ölçülerinden yararlanılmaktadır (Alpar, 2011).

81 ile ait 11 değişkenin yer aldığı bu çalışmada, gözlemlerin birbirine olan yakınlığı/uzaklığı çerçevesinde bir kümeleme analizi yapılacağı için bir uzaklık (benzemezlik) ölçüsüne ihtiyaç duyulmaktadır. Diğer taraftan, kümeleme analizinde, büyük varyansa sahip değişkenlerin kümelerin belirlenmesindeki belirleyici etkisinin bertaraf edilebilmesi için de, farklı ölçü birimlerindeki değişkenler dönüştürme işlemine tabi tutulmalı yani standardize edilmeli; aynı ölçü birimine sahip olanların ise logaritmaları alınarak değişkenlikteki aşırı farklılıkların bertaraf edilmesi yoluna gidilmelidir (Fraley ve Raftery, 1998). Bu çalışmada yer alan veriler aynı ölçü birimine sahip ve aşırı sağa çarpık sıklık sayılarından yani kesikli verilerden oluştuğu için z-dönüşümü gibi standardize yöntemleri kullanılamamıştır. Bu sebeple ve de değişkenlere ait verilerdeki değişim aralıkları ile varyansların da büyük farklılıklar arz etmesi sebebiyle verilerin logaritmaları kullanılmış; böylece, dolaylı olarak veriler oransal-sürekli veriler haline gelmiştir. Bu doğrultuda, hibrid kümelemenin ilk aşaması olan hiyerarşik kümelemede, sürekli veriler için uygun olan uzaklık ölçülerinden ikisi olan

Öklid ve kare-Öklid uzaklık ölçülerinden sırasıyla ortalama ve WARD algoritmalarının kullanıldığı analiz sürecinde yararlanılmıştır.

X_{ik}, i. gözlemin k. değişken değeri; X_{jk}, j. gözlemin k. değişken değeri ve p, değişken sayısı olmak üzere, Öklid ve kare-Öklid uzaklık ölçülerinin formülleri aşağıda verilmiştir:

Öklid Uzaklık Ölçüsü : $d_{ij} = \sqrt{\sum_{k=1}^{p}(X_{ik} - X_{jk})^2}$ (1)

Kare-Öklid Uzaklık Ölçüsü : $d_{ij} = \sum_{k=1}^{p}(X_{ik} - X_{jk})^2$ (2)

Diğer taraftan hibrid kümelemenin ikinci aşamasında kullanılacak k-ortalamalar yönteminde, kümeler arası değişkenliğin en büyük ve kümeler içi değişkenliğin ise en küçük olacak şekilde gözlemlerin önceden belirli olan sayıda kümelere atanmasında da genelde Öklid uzaklığı kullanıldığı için verilerin sürekli veya oran ölçekli tipte olması zaten gerekli olmaktaydı.

3.2 Kümeleme Yönteminin Seçimi

Uzaklık ya da benzerlik ölçüleri ve dolayısıyla matrisleri temelinde gözlemleri veya değişkenleri kümelemede kullanılan yöntemler genel olarak bulanık (fuzzy) ve bulanık olmayan (non-fuzzy) yöntemler olarak ayrılmakta; bulanık olmayan kümeleme yöntemleri de temelde hiyerarşik (aşamalı) ve hiyerarşik olmayan (aşamasız) şeklinde sınıflandırılmaktadır.

3.2.1 Hiyerarşik Kümeleme Yöntemi

Hiyerarşik kümeleme yöntemi kendi içinde birleştirici (agglomerative) ve ayrıştırıcı (divisive) kümeleme yöntemleri olarak ikiye ayrılmaktadır. Birleştirici yöntemlerde, tüm gözlemler başlangıç düzeyinde ayrı birer küme olarak ele alınmakta ve daha sonra uzaklık veya benzerlik ölçülerine göre en yakın veya en çok benzeyen gözlemler birleşerek bir küme oluşturmaktadır. Bu doğrultudaki her aşamada gözlem sayısı bir azalarak tüm gözlemler tek bir kümede birleşinceye kadar kümeleme işlemi devam etmektedir. Bu çerçevede ve farklı uzaklık veya benzerlik ölçüleri ile farklı küme bağlantı tekniklerine göre bir araya gelen gözlemlerin oluşturduğu kümeler, ağaç diyagramları (dendogram) veya buz saçaklarına benzeyen (icicle plot) grafikler ile gözlemlenebilmektedir. Ayrıştırıcı kümeleme yönteminde ise, birleştirici yöntemdeki süreç tersine işleyerek başlangıçta tüm gözlemlerin oluşturduğu tek bir küme, benzemezlik veya uzaklık ölçüleri temelinde her gözlem tek bir kümeyi temsil edene kadar devam etmektedir. Hiyerarşik kümeleme yöntemlerinde küme sayısının ne olacağına ilişkin önsel bir bilgiye ihtiyaç duyulmazken, hiyerarşik olmayan kümeleme yöntemlerinde ise, oluşturulacak küme sayısının önceden bilinmesi gerekmektedir. Hiyerarşik olmayan kümeleme yöntemlerindeki bu kısıtlamaya karşın, her gözlem için uzaklık veya benzerlik matrislerinin hesaplanmasına gerek duyulmadığı için daha büyük (n>300-400 gibi) veri setlerine uygulanabilmesindeki kolaylık ve aykırı (outlier) değerlere daha az duyarlı olmaları bu yöntemleri avantajlı kılmaktadır (Alpar, 2011).

Bu çalışmada kullanılacak hibrid kümeleme yönteminin ilk aşaması olan hiyerarşik kümeleme yöntemi, tek bağlantı, tam bağlantı, ortalama bağlantı, centroid, medyan ve WARD gibi farklı çözüm algoritmaları tercih edilerek uygulanabilmektedir. Kümeleme analizinin kalitesi, geçerliliği ve oluşacak küme sayıları, seçilen kümeleme algoritmasına, değişkenler veya gözlemler arasında hangi uzaklık ya da benzerlik ölçütünün kullanılacağına ve de seçilen değişkenlere (bir değişken eklemeye veya silmeye) karşı çok duyarlıdır (Templ, Filzmoser ve Reimann, 2008). Bu algoritmalardan hangisinin tercih edileceği istenilen kümeleme yapısına bağlı olduğu için, tercih edilecek küme bağlantı algoritmasına karar vermek de büyük önem taşımaktadır. Bahsi geçen algoritmaların özellikleri kısaca aşağıda verilmiştir (Milligan, 1980; Everitt, Landau, Leese ve Stahl, 2011; Rao ve Srinivas, 2008); Hubert, 1974):

- *Tam Bağlantı-En Uzak Komşuluk (Complete Linkage-Farthest Neighbor)*: Ortalama bağlantı algoritması gibi daha belirgin bir yapı ortaya çıkarabilirse de, birleştirmenin (agglomeration) ilk aşamalarında homojen kümeler oluşturmaya eğilimlidir. Büyük veri kümeleri için genelde uygun olmayıp, özellikle belirgin olmayan, küçük ve yoğun küme yapısına sahip veriler için uygundur.

- *Tek Bağlantı-En Yakın Komşuluk (Single Linkage-Nearest Neighbor)*: Zincirleme etkisine (chain effect) sahiptir, yani giderek uzayan küme yapıları oluşturmaya eğilimlidir ve aykırı değerlerden de çok etkilenir. Ancak küme yapıları birbirlerinden çok ayrık bir yapıda ise, bu algoritma başarılı sonuçlar verebilir.

- *WARD*: Küme içi varyansın (değişkenliğin) en küçük, kümeler arası varyansın en büyüklenmesi amaçlanır. Centroid ve medyan bağlantı tekniklerinin karma ve ağırlıklı halidir. Daha az ve aynı sayıda gözlem içeren kümeler oluşturmaya eğilimli ve aykırı değerlere de duyarlıdır.

- *Merkezi Bağlantı (Centroid)*: İki küme arasındaki uzaklık, bu kümelerin kendi merkezleri arasındaki uzaklık olarak ele alınır. Aykırı değerlerden daha az etkilenir.

- *Medyan Bağlantı*: İki küme arasındaki uzaklık, bu kümelerin merkezleri arasındaki uzaklığın eşit ağırlıklı olarak hesaplanmasıyla elde edilir. Aykırı değerlerden daha az etkilenir.

- *Ortalama Bağlantı (Average Linkage-Within Groups)*: Kümeler küçük varyanslar ile birbirine bağlıdır ve tek bağlantı ile tam bağlantı arası sonuçlar vermesi ve aykırı gözlemlerden en az etkilenen algoritma olması sebebiyle çok tercih edilir.

- *Ortalama Bağlantı (Average Linkage-Between Groups)*: Kümeler arası ortalama bağlantı algoritması da tek bağlantı algoritması gibi zincirleme etkiye sahiptir.

Kuiper ve Fisher (1975), Blashfield (1976), Milligan (1980), Hands ve Everitt (1987), Milligan ve Cooper (1988), Ferreira ve Hitchcock (2009)'ın ilgili çalışmalarında, en iyi kümeleme sonuçlarının WARD ve Ortalama Bağlantı (*Average Linkage-Within Groups*) algoritmaları ile elde edildiği Rand endeksi, Cohen istatistiği ve Cophenetic korelasyon katsayısı gibi kriterler ışığında gösterilmiştir.

Ayrıca bu çalışmalarda, merkezi bağlantı ve ortalama bağlantı algoritmalarının kullanılmasıyla görece daha eşit sayıda gözleme sahip kümelerin oluştuğu; tam bağlantı ve WARD algoritmalarının kullanılmasıyla ise görece eşit olmayan sayıda gözlem içeren kümelerin oluştuğu görülmüştür. Diğer taraftan ortalama bağlantı algoritmasının aykırı gözlemlere en az duyarlı algoritma olduğu ve de WARD algoritmasının aykırı gözlemlere duyarlı olduğu da bu çalışmalarda ifade edilmiştir.

Yukarıda yer alan bilgiler ve kullanılan verilerdeki aykırı gözlemlerin de varlığı dikkate alınarak hibrid kümelemenin ilk aşaması olan hiyerarşik kümelemede, WARD ve ortalamalar algoritmaları ayrı ayrı kullanılarak elde edilen küme merkezleri ikinci aşamada k-ortalamalar kümeleme yönteminde kullanılmış ve böylece oluşan küme yapıları için karşılaştırma yapabilme imkânı elde edilmiştir. Bu doğrultuda, WARD ve ortalama bağlantı küme bağlantı tekniklerine ilişkin matematiksel formüller aşağıda yer almıştır.

k, l, m kümeler; N, küme birim sayısı olmak üzere m. kümenin j. küme ile uzaklığı:

$$\textit{Ortalama Bağlantı} : d_{mj} = (N_k d_{kj} + N_l d_{lj})/N_m \tag{3}$$

$$\textit{WARD} : d_{mj} = \left[\left((N_j + N_k)d_{kj} + (N_j + N_l)d_{lj} - N_j d_{kl}\right)/(N_j + N_m)\right] \tag{4}$$

3.2.2 K-Ortalamalar Kümeleme Yöntemi

K-ortalamalar (k-means) yöntemi (Mac Queen, 1967), hiyerarşik olmayan-parçalayıcı (partitioning) kümeleme analizinde kullanılan yöntemlerin en çok bilineni ve de özellikle büyük veri kümeleri için en yaygın olarak kullanılanıdır. K-ortalamalar yönteminde, en az küme sayısı 2 ve en çok küme sayısı da gözlem sayısı kadar olabilmekte; gözlemler, kümeler arası değişkenliğin en fazla ve kümeler içi değişkenliğin en az olabilmesi adına her gözlem için gözlemlerle küme merkezleri arasındaki Öklid uzaklığının karelerinin toplamını minimum kılacak şekilde sınıflandırılmaktadır.

K, küme sayısı; X_{ij}, i. kümeye ait j. birim; m_i, i. kümenin merkezi; n_i, i. kümenin gözlem sayısı olmak üzere en küçüklenecek amaç fonksiyonu (Mac Queen, 1967) aşağıda verilmiştir:

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left\|X_{ij} - m_i\right\|^2 \tag{5}$$

Bu yöntemdeki süreç, sayısı daha önce belirlenmiş olan kümelere rasgele birer başlangıç merkezinin seçilmesiyle başlamakta ve her bir gözlemin yakınlık/benzerlik düzeyine göre bu kümelere atanmasıyla yinelemeli bir şekilde devam etmektedir. Burada dikkat edilmesi gereken husus şudur ki: gözlemlerin uzaklık veya benzerlik ölçülerine göre bir kümeye atanmasından sonra bu gözlem ya da gözlemlerin bir daha yer değiştiremeyecek olmasıdır. Bu sebepledir ki, hiyerarşik yöntemlerde yinelemeli bir süreç yokken, k-ortalamalar yönteminde başlangıç merkezlerinin seçim süreci yinelemeli hale gelmektedir. Burada ise, hiyerarşik olmayan yöntemlerin rassal küme merkezleri seçimi sebebiyle her denemede farklı sonuçlar verebiliyor olması sakıncası ortaya çıkmaktadır.

Yöntemin bir diğer sakıncası da, başlangıç küme merkezlerinin aykırı değerlerden seçilerek az sayıda aykırı değeri içeren küçük kümeler oluşturma eğiliminde olmasıdır (Alpar, 2011). Bu sebeple araştırmacı, eğer veri setinde aykırı gözlemler varsa ya onları kümeleme analizi öncesi çıkarmalı ya da ortalama kümeleme algoritmasını da karşılaştırma yapabilmek için tercih etmelidir.

3.2.3 Hibrid Kümeleme Yöntemi

Hibrid kümeleme, hiyerarşik ve k-ortalamalar gibi mevcut klasik kümeleme algoritmalarının aşağıda özetlenen bazı olumsuz yönlerinden kurtulmak (veya azaltmak) amacıyla iki veya daha fazla kümeleme yönteminin uygun şekilde birleştirilerek bir arada kullanılmasıyla (sentezlenmesiyle) oluşturulan melez bir yöntemdir (Gan, Ma ve Wu, 2007). Bu olumsuz yönler;

- Aşamalı kümeleme yöntemlerinde kullanılan algoritmaların hangi çözüm aşamasında durması gerektiğine ilişkin kesin bir ölçüt bulunamaması,
- Aşamalı kümeleme yöntemlerinde bir birimin, gözlemin belirli bir kümeye dâhil edilmesinden sonra iyileştirme adına başka bir kümeye yeniden atanamaması,
- K-ortalamalar algoritması gibi parçalayıcı kümeleme yöntemlerinde küme sayısının analiz öncesinde bilinmesinin gerekmekte olması,
- K-ortalamalar kümeleme algoritmasında rassal olarak seçilecek başlangıç küme sayısı ve küme merkezlerinin sonuçları etkilemekte olması,
- K-ortalamalar kümeleme algoritmasının aykırı (outlier) değerlere duyarlı olması,
- K-ortalamalar kümeleme algoritmasında rastgele seçilen küme merkezlerinin birbirine çok yakın olabilmesi,

şeklinde ifade edilebilir.

Birleştirici hiyerarşik kümeleme ile k-ortalamalar kümeleme yöntemlerinin yukarıda bahsi geçen olumsuz yönlerinin bertaraf edilebilmesi adına geliştirilmiş olan ve bu çalışmada da kullanılacak olan *"hibrid hiyerarşik k-ortalamalar"* yönteminde, hiyerarşik kümeleme yöntemiyle elde edilecek başlangıç küme merkezleri (centroidler) parçalayıcı kümeleme yöntemlerinden k-ortalamalar algoritmasını başlatmak için kullanılmaktadır. Ayrıca, bu yöntemde başlangıç küme sayısına karar verilirken hiyerarşik kümelemedeki dendogramlar kullanılabileceği gibi *"Silhouette Index"* gibi küme doğrulama araçlarından da istifade edilebilmektedir. Hibrid yöntemin bu aşamaları,

- Kümeleme analizine tabi tutulacak değişkenler ile birimlerin belirlenmesi,
- Analizde kullanılacak verilerde değişim aralıklarının ve varsa aykırı gözlemlerin belirlenerek verilere uygun dönüşüm işleminin yapılması (z-skor, logaritma almak vb.),
- Analizde kullanılacak verilerin aralık, sayma sayısı veya kategorik özelik taşıması durumunu da göz önünde bulundurarak uygun küme bağlantı algoritması (WARD, ortalama bağlantı, en yakın

komşu vb.) ile uzaklık (birimler kümelenecekse) veya korelasyon (değişkenler kümelenecekse) ölçüsünün seçilerek matrislerin hesaplanması,

- Verilerin, küme doğrulama araçlarının gösterdiği ya da önsel olarak doğruluğuna kanaat getirilen sayıda (ağaç diyagramına veya ilgili çalışmalara dayanarak) kümeye ayrılacak şekilde hiyerarşik kümeleme ile sınıflandırılarak küme merkezlerinin belirlenmesi,
- Hiyerarşik kümeleme ile elde edilmiş küme merkezlerinin başlangıç küme merkezleri olarak seçilerek verilerin bu defa k-ortalamalar kümeleme algoritması ile yeniden sınıflandırılması ve daha önceden belirlenen sayıda kümenin elde edilmesi,
- Elde edilen kümeleme analizi verilerinin küme doğrulama araçları ile değerlendirilmesi,

olarak verilmektedir (Chipman ve Tibshirani, 2006; Chen vd. 2005).

3.3 Küme Sayısının Belirlenmesi

Kümeleme analizinde, uygun küme sayısının bulunması önemli bir sorundur. Özellikle k-ortalamalar gibi bazı kümeleme teknikleri analizin en başında küme sayısının belirlenmesini gerekli kılmaktadır. Diğer taraftan, bir veri seti için farklı kümeleme algoritmalarının seçilmesi sonucu farklı küme yapıları oluşabileceği için, kümeleme kalitesinin sorgulanması da büyük önem taşımaktadır. Bu çerçevede, aynı veri kümesinden elde edilebilecek farklı küme yapılarının anlamlı olup olmadığının sınanması kümeleme geçerliliği/doğrulanması olarak adlandırılmış ve böylece kümeleme kalitesinin değerlendirilebilmesi için çeşitli doğrulama/geçerlilik ölçütleri (validation tools) geliştirilmiştir. Böylece eldeki gözlemler ya da değişkenler için doğal küme yapısının ve küme sayısının ne olması gerektiği ortaya koyulabilmekte ve alınan kararlar kümeleme geçerlilik ölçütleri ile desteklenebilmektedir (Bolshakova ve Azuaje, 2003).

İçsel ve dışsal kriter (internal and external criteria) olarak iki farklı tipte küme geçerlilik ölçütü vardır. Dışsal kriterlerde (Rand, Jaccard, Hubert, vb.), veri hakkında sahip olunan ön bilgi ile kümeleme algoritması sonunda elde edilen kümeleme yapısı karşılaştırılır; kümeleme sonunda elde edilen gözlemlerin küme etiketleri ile daha önceden bilgi sahibi olunan gözlemlerin kategori etiketleri karşılaştırılır. İçsel kriterler (Calinski-Harabasz, Hartigan, Shilhouette, vb.) ise, veri seti ile kümeleme yapısı arasındaki uyumun belirlenmesinde sadece veri setindeki doğal yapıyı ve nicel değerleri göz önünde bulundurarak kümeleme sonuçlarını değerlendirir. İçsel kriterlerin çoğu, kümeler içi kareler toplamını veya kümeler arası kareler toplamını temel alarak değerlendirmeyi yapmaktadır (Theodoridis ve Koutroumbas, 2006).

Kümeleme analiziyle ilgili çalışmalarda daha çok içsel kriterlerin küme geçerlilik ölçütü olarak kullanıldığı görülmektedir. Bunun sebebi olarak, içsel kriterlerin dışsal kriterlere göre doğal ve anlamlı küme yapıları oluşturmada daha başarılı sonuçlar vermesi gösterilebilir. Örneğin Rendon, Abundez, Arizmendi ve Quiroz'un (2011) çalışmasında, içinde silhouette indeksinin de yer aldığı bazı içsel

küme geçerlilik ölçütleriyle dışsal küme geçerlilik ölçütlerini k-ortalamalar tekniği üzerinde uygulamışlardır. Çalışma sonucunda da içsel kriterlerin daha başarılı doğal küme yapıları oluşturdukları ortaya konmuştur.

Bu çalışmada, küme sayısının belirlenmesinde içsel küme geçerlilik ölçütlerinden olan ortalama silhouette genişliği endeksi (Rousseeuw, 1987) kullanılacaktır. Bu endeksin nasıl elde edildiği aşağıda gösterilmektedir.

i: kümedeki birim sayısı $(i = 1, \ldots\ldots, m)$

j: küme sayısı $(j = 1, \ldots\ldots, c)$

$a(i)$: X_j kümesindeki i. birim ile diğer tüm X_j kümesi elemanları arasındaki ortalama uzaklık,

$b(i)$: X_j kümesindeki i. birim ile X_j kümesi dışındaki diğer tüm kümeler arasındaki ortalama uzaklıkların en küçük olanı

olarak X_j kümesindeki i. birim için silhouette genişliği olan $s(i)$ aşağıdaki gibi hesaplanmaktadır:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad ve \quad -1 \leq s(i) \leq 1 \tag{6}$$

$s(i)$'den hareketle j. küme için ortalama silhouette genişliği olan $s(j)$ ise aşağıdaki gibi hesaplanmaktadır:

$$s(j) = \frac{1}{m}\sum_{i=1}^{m} s(i) \tag{7}$$

Tüm kümeleri kapsayan genel ortalama silhouette genişliği endeksi değeri de aşağıdaki gibi hesaplanmaktadır:

$$S(ij) = \frac{1}{c}\sum_{j=1}^{c} s(j) \tag{8}$$

Burada, $a(i)$, $b(i)$'den ne kadar küçükse $s(i)$ o kadar +1'e yaklaşacak, $a(i)$, $b(i)$'den ne kadar büyükse de $s(i)$ o kadar -1'e yaklaşacaktır. Sonuç olarak en büyük silhouette genişliği genel ortalamasına sahip olan küme yapısı, verilerin en iyi sınıflanmasını göstermektedir. Diğer taraftan ortalama silhouette endeksinin 0,5'in üzerinde olması kümelemenin makul düzeyde başarılı olduğunu ve 0,7'nin üzerinde olması ise çok güçlü olduğunu (bkz. Tablo 1) göstermektedir (Ng ve Han, 1994).

Tablo 1. Ortalama Silhouette Genişliği Endeksi ve Küme Yapısı İlişkisi

Ortalama Silhouette Genişliği (Average Silhouette Width)	Küme Yapısı Hakkında Yorum
0,71 - 1	Güçlü
0,51-0,70	Yeterli
0,26-0,50	Zayıf
≤ 0,25	Başarısız

4. UYGULAMA VE BULGULAR

İl Nüfus ve Vatandaşlık Müdürlüklerinin iş yoğunluklarına göre hibrid kümeleme analizi ile sınıflandırılacağı bu çalışmada, kurumun hizmet verdiği 11 ana hizmet kalemine ait işlem/hizmet sayıları 11 ayrı değişken olarak; 81 İl Nüfus ve Vatandaşlık Müdürlüğü ise uygulama birimleri olarak 2013 yılı verilerinden hareketle analize dâhil edilecektir. Bu değişkenler aşağıda yer almaktadır:

Tablo 2. Kümeleme Analizinde Yer Alacak Değişkenler

1	Doğum	7	Nüfus Kayıt Örneği
2	Ölüm	8	Yerleşim Yeri Belgesi
3	Evlenme	9	Nüfus Cüzdanı Düzenleme
4	Boşanma	10	Vatandaşlık İşleri
5	Kayıt Düzeltme	11	Diğer (Pasaport, Miras, Evlatlık vb.)
6	Adres Beyan		

Hibrid kümeleme analizine geçmeden önce küme sayısına karar verebilmek için Ortalama Silhouette Genişliği Endeksi'nden faydalanılmış ve elde edilen endeks değerleri de aşağıda verilmiştir:

Tablo 3. Ortalama Silhouette Genişliği Endeksi Değerleri

Küme Sayısı	2	3	4	5	6	7	8
Silhouette Endeksi	***0,94***	***0,83***	***0,72***	***0,65***	***0,53***	0,48	0,43

Uygun küme sayısının belirlenmesinde endeks değerinin 0,5'ten büyük olması gerektiğinden çalışmanın yöntem bölümünde bahsedilmişti. Buna göre en uygun küme sayısının 2 olduğu, ancak 6 kümeye kadar da seçim yapılabileceği Tablo 3'te görülmektedir. Çalışmada yer alan değişkenlerin bir kamu kurumuna ait olması sebebiyle, iş yoğunluğu bazında personel atamaları gerçekleştirilirken sayısal gerçeklikler kadar eş ve hastalık nedenli tayinler ile kadro pozisyonu gibi unsurlar da göz önünde bulundurulmalıdır. Bu sebeple kamu otoritesinin karar alanını geniş tutabilmesi çerçevesinde küme sayısının 6 olmasının uygun olacağını düşünmekteyiz. Bu doğrultuda ve 6 küme sayısından hareketle elde edilen hibrid k-ortalamalar kümeleme analizi sonuçları Tablo 4'te yer almaktadır.

Tablo 4'e bakıldığında her kümeye ait sütunda tek bir alan ve ek olarak iki bölmeli alanın olduğu görülecektir. Tek parça olan alan, hem ortalama hem de WARD algoritmaları aracılığıyla elde edilmiş kümeleme analizi sonuçlarına göre aynı kümede yer alan illeri içermektedir. İki bölmeli alanda ise, sol tarafta sadece ortalamalar algoritmasına göre ilgili kümede yer alan iller; sağ tarafta ise, sadece WARD algoritmasına göre ilgili kümede yer alan iller görülmektedir.

Tablo 4. Hibrid K-Ortalamalar Kümeleme Analizi Sonuçları

1.Küme		2.Küme		3.Küme		4.Küme		5.Küme		6.Küme	
Ort.	**WARD**	**Ort.**	**WARD**	**Ort.**	**WARD**	**Ort.**	**WARD**	**Ort.**	**WARD**	**Ort.**	**WARD**
13 İl	**17 İl**	**27 İl**	**20 İl**	**9 İl**	**8 İl**	**26 İl**	**22 İl**	**5 İl**	**11 İl**	**1 İl**	**3 İl**
Adana Antalya Bursa Diyarbakır Gaziantep Hatay Kayseri Kocaeli Konya Mersin Şanlıurfa		Adıyaman Afyonkarahisar Aksaray Çorum Elazığ Erzurum Eskişehir Giresun Malatya Mardin Muğla Ordu Osmaniye Sakarya Sivas Tekirdağ Tokat Trabzon Yozgat Zonguldak		Ağrı Batman Bitlis Hakkâri Muş Siirt Şırnak		Amasya Bolu Burdur Çanakkale Düzce Edirne Erzincan Iğdır Isparta Karaman Kastamonu Kırıkkale Kırklareli Kırşehir Kütahya Nevşehir Niğde Rize Sinop Uşak		Ardahan Bayburt Gümüşhane Kilis Tunceli		İstanbul	
Ankara İzmir	Manisa K.Maraş Samsun Aydın Balıkesir Denizli	Aydın Balıkesir Denizli K.Maraş Manisa Samsun Van		Bingöl Kars	Van	Artvin Bartın Bilecik Çankırı Karabük Yalova	Bingöl Kars		Artvin Bartın Bilecik Çankırı Karabük Yalova		Ankara İzmir

Bu çalışmanın asıl amacı her ne kadar illerin iş yoğunluklarının neden farklılık arz ettiğini ortaya koymak olmasa da, aynı kümede yer alan iller ile farklı kümede yer alan illerin varsa sosyoekonomik açıdan benzerliklerinin ve farklılıklarının da genel olarak ortaya çıkarılabilmesi yararlı olabilecektir. Bu açıdan ilgili kümeler, illerin genel anlamda sosyoekonomik yapısının birer göstergesi olarak nitelendirilebilecek olan toplam nüfustaki payları, okuma yazma oranları ve gayri safi katma değere sağladıkları katkıları itibariyle de değerlendirilecektir. İllerin nüfusunun azlığı ya da çokluğu potansiyel olarak kurumun iş yüküne doğrudan etki edebilecek bir göstergedir. Örneğin nüfus ne kadar fazla ise, potansiyel olarak ölüm, doğum ve nüfus cüzdanı işlemleri de o ölçüde fazla olmaktadır. Benzer şekilde, illerin genel ekonomiye sağladıkları katkının büyüklüğü de hem illerin iş potansiyeli anlamında nüfusunu etkilemekte hem de yurtdışı seyahatlerin fazlalığı sebebiyle pasaport vb. işlemlerin sayısını da artırmaktadır. Diğer taraftan temel eğitim seviyesinin bir göstergesi olan okuryazarlık oranının da bir gösterge olarak kabul edilmiş olmasındaki sebep ise, okuryazar olmayan kişilerin genel olarak nüfus işlemlerini geciktirebilmeleri ya da uzun süreyle ihmal edebilmeleridir. Ayrıca, okuryazarlık seviyesi pasaport vb. işlemlerdeki talebi de dolaylı olarak etkilemektedir.

Yukarıda bahsedilen gerekçelerin de etkisiyle ve daha geniş bir bakış açısı sunabilmesi amacıyla, oluşan küme yapılarının niceliksel yapısı hem iş yoğunlukları hem de ilgili sosyoekonomik göstergeler bazında aşağıda yer alan Tablo 5 ile Tablo 6'da verilmiştir.

Tablo 5. Genel Değerlendirme – Ortalamalar Yöntemi

Değişken / Küme	1	2	3	4	5	6
Toplam GSYH Katma Değer (%)	41,1	24,8	1,4	10,6	0,6	21,5
Ortalama Okuma Yazma Oranı (%)	96,1	95,0	91,2	95,7	93,7	97,4
Toplam Nüfus İçi Pay (%)	38,7	27,6	4,5	10,0	0,7	18,5
Toplam İş Yükü (%)	35,2	30,4	4,8	12,3	1,1	16,2
Ortalama İş Yükü	1.526.533	635.304	301.667	266.044	128.313	9.107.156

Tablo 6. Genel Değerlendirme – WARD Yöntemi

Değişken / Küme	1	2	3	4	5	6
Toplam GSYH Katma Değer (%)	35,5	14,7	1,5	8,7	2,8	36,8
Ortalama Okuma Yazma Oranı (%)	96,1	94,6	91,2	95,4	94,8	97,7
Toplam Nüfus İçi Pay (%)	34,5	17,3	5,2	9,2	2,3	30,4
Toplam İş Yükü (%)	33,6	19,7	5,3	11,1	3,4	27,1
Ortalama İş Yükü	1.114.230	553.885	371.842	283.274	172.964	5.083.317

Tablo 4'ü Tablo 5 ve Tablo 6 ile birlikte değerlendirdiğimizde her iki algoritmaya göre elde edilen kümelerin özellikleri aşağıda özetlenmiştir:

❖ Ortalamalar Algoritmasına Göre:

1. Küme: İstanbul dışındaki büyükşehirlerin çoğunu içeren bu küme, toplam katma değer (%41,1), toplam nüfus (%38,7) ve toplam iş yükündeki (%35,2) paylar açısından 1. sırada; ortalama okuma yazma oranı (%96,1) ve ortalama iş yükü (1.526.533) açısından ise 2. sırada yer almaktadır.

2. Küme: Az sayıda büyükşehrin yanı sıra çoğunluğu orta gelişmişlik düzeyindeki illerden oluşan bu küme, toplam katma değer (%24,8), toplam nüfus (%27,6) ve toplam iş yükündeki (%30,4) paylar açısından 2. sırada; ortalama okuma yazma oranında (%95) 4. sırada ve ortalama iş yükü (635.304) açısından ise 3. sırada yer almaktadır.

3. Küme: Doğu ve güneydoğudaki az gelişmiş illerden oluşan bu küme, toplam katma değer (%1,4), toplam nüfus (%4,5) ve toplam iş yükündeki (%4,8) paylar açısından 5. sırada; ortalama okuma yazma oranında (%91,2) ile 6. ve son sırada bulunmakta olup, ortalama iş yükü (301.667) açısından ise 4. sırada yer almaktadır.

4. Küme: Alt-orta seviyede gelişmiş illerden oluşan bu küme, toplam katma değer (%10,6), toplam nüfus (%10) ve toplam iş yükündeki (%12,3) paylar açısından 4. sırada; ortalama okuma yazma oranında (%95,7) ile 3. ve ortalama iş yükü (266.044) açısından ise 5. sırada yer almaktadır.

5. Küme: Çoğunluğu doğu ve kuzeydoğudaki az gelişmiş illerden oluşan bu küme, toplam katma değer (%0,6), toplam nüfus (%0,7), toplam iş yükündeki paylar (%1,1) ve ortalama iş yükü

(128.313) açısından 6. ve son sırada; ortalama okuma yazma oranında ise (%93,7) ile 5. sırada bulunmaktadır.

6. Küme: Sadece en gelişmiş il olan İstanbul'dan oluşan bu küme, toplam katma değer (%21,5), toplam nüfus (%18,5) ve toplam iş yükündeki (%16,2) paylar açısından 3. sırada; ortalama okuma yazma oranı (%97,4) ve ortalama iş yükü (9.107.156) açısından ise 1. sırada yer almaktadır.

❖ WARD Algoritmasına Göre:

1. Küme: İstanbul, Ankara, İzmir dışındaki çoğu büyükşehri içeren bu küme, toplam katma değerdeki pay (%35,5), ortalama iş yükü (1.114.230) ve ortalama okuma yazma oranı (%96,11) açısından 2. sırada; toplam nüfus (%34,5) ve toplam iş yükündeki paylar açısında ise 1. sırada yer almaktadır.

2. Küme: Az sayıda büyükşehrin yanı sıra çoğunluğu orta gelişmişlik düzeyindeki illerden oluşan bu küme, toplam katma değer (%14,7), toplam nüfus (%17,3) ve toplam iş yükündeki (%19,7) paylar açısından 3. sırada; ortalama okuma yazma oranında (%94,6) 5. sırada ve ortalama iş yükü (553.885) açısından ise 3. sırada yer almaktadır.

3. Küme: Doğu ve güneydoğudaki az gelişmiş illerden oluşan bu küme, toplam katma değer (%1,5) payı ve ortalama okuma yazma oranı (%91,2) açısından 6. ve son sırada; toplam nüfus (%5,2) ve toplam iş yükündeki (%5,3) payı açısından 5. sırada ve ortalama iş yükü (371.842) açısından ise 4. sırada yer almaktadır.

4. Küme: Çoğunluğu alt-orta seviyede gelişmiş illerden oluşan bu küme, toplam katma değer (%8,7), toplam nüfus (%9,2) ve toplam iş yükündeki (%11,1) paylar açısından 4. sırada; ortalama okuma yazma oranında (%95,4) 3. ve ortalama iş yükü (283.274) açısından ise 5. sırada yer almaktadır.

5. Küme: Çoğunluğu kuzey bölgelerindeki az gelişmiş illerden oluşan bu küme, toplam katma değerdeki (%2,8) pay açısından 5. sırada; ortalama okuma yazma oranında (%94,8) 4. sırada; toplam nüfus (%2,3), toplam iş yükündeki (%3,4) paylar ile ortalama iş yükü (172.964) açısından ise 6. ve son sırada yer almaktadır.

6. Küme: Sadece en gelişmiş iller olan İstanbul, Ankara ve İzmir'den oluşan bu küme, toplam katma değerdeki pay (%36,8), ortalama okuma yazma oranı (%97,7) ve ortalama iş yükü (5.083.317) açısından 1. sırada; toplam nüfus (%30,4) ve toplam iş yükü payı (27,1) açısından ise 2. sırada yer almaktadır.

Yukarıda yer alan değerlendirmelerden de anlaşılacağı üzere, her iki algoritmaya göre elde edilmiş hibrid kümeleme sonuçları genel itibariyle birbirine benzemekte, kümelerde yer alan illerin büyük çoğunluğunun aynı olduğu görülmektedir. Ancak, WARD algoritmasının uç değerlere çok daha fazla duyarlı olması neticesinde bazı iller iş yoğunluğu bazında farklı kümelerde yer almıştır. Örneğin, ortalamalar algoritmasına göre elde edilen küme yapısına göre birinci kümede yer alan Manisa, Maraş, Samsun, Denizli, Aydın ve Balıkesir ikinci kümeye; Ankara ve İzmir ise sadece İstanbul'un yer aldığı

altıncı kümeye atanmıştır. Burada belirtilmesi gereken son bir husus da, ilgili yazında bu çalışmada olduğu gibi uç değerli gözlemlere sahip birimlerin analizden çıkarılması gerektiği yönünde yapılan eleştirilere yöneliktir. İş yükü anlamında logaritması alınmasına rağmen halen bazı uç değerli gözlemlere sahip olduğu görülen İstanbul, Ankara ve İzmir'in çıkarılması ve 5 küme sayısı ile analizin yapılması durumunda bile küme yapılarının her iki algoritmaya göre de değişmediği görülmektedir.

Kümeleme analizinde, küme içi farklılıkların en az ve de kümeler arası farklılıkların ise en çok olması amaçlanmaktadır. Bu çerçevede yapılan kümeleme analizi neticesinde oluşan kümelerin ve dolayısıyla küme ortalamalarının birbirlerinden yeterince farklı bir yapı ortaya koyup koyamadıklarının belirtilmesi gerekmektedir. Özellikle ortalama silhouette genişliği endeksine göre en iyi küme sayısının 2 olmasına rağmen daha önce belirtilen gerekçelerle küme sayısının 6 olarak ele alındığı bu çalışmada, kümeler arası farklılıkların istatistiki çerçevede test edilmesi daha da önemli hale gelmektedir.

Kümeleme analizi neticesinde elde edilen 6 kümenin istatistiki olarak birbirinden farklı olup olmadığının test edilmesinde kullanılacak varyans analizinden önce değişkenlerin normal dağılıp dağılmadığının ve de varyansların homojen olup olmadıklarının da test edilmesi gereklidir. Değişkenlere ait ham veriler yerine logaritmaları alınmış veriler kullanılmış olup, iki algoritmaya göre de tüm değişkenlere ait verilerin normal dağılıma uygun bir dağılım sergilediği görülmektedir.

Tablo 7. Shapiro-Wilk Normallik Testi (p değerleri)

Kümeleme Algortiması	**Ortalamalar (Average)**						**WARD**					
Değişken / Küme Sayısı	**1**	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**
Doğum	,50	,88	,84	,96	,18	,40	,50	,88	,84	,96	,18	,40
Ölüm	,69	,19	,78	,58	,44	,11	,69	,19	,78	,58	,44	,11
Evlenme	,41	,19	,77	,63	,21	,18	,41	,19	,77	,63	,21	,18
Boşanma	,65	,78	,84	,44	,40	,09	,65	,78	,84	,44	,40	,09
Kayıt Düzeltme	,14	,25	,43	,15	,90	,46	,14	,25	,43	,15	,90	,46
Adres Beyan	,35	,82	,09	,34	,37	,42	,35	,82	,09	,34	,37	,42
Nüfus Kayıt Örneği	,59	,75	,10	,94	,33	,57	,59	,75	,10	,94	,33	,57
Yerleşim Yeri Belgesi	,44	,16	,80	,69	,33	,26	,44	,16	,80	,69	,33	,26
Nüfus Cüzdanı Düzenleme	,22	,34	,43	,62	,12	,26	,22	,34	,43	,62	,12	,26
Vatandaşlık	,77	,45	,32	,16	,49	,73	,77	,45	,32	,16	,49	,73
Diğer	,27	,82	,73	,10	,61	,74	,27	,82	,73	,10	,61	,74

Not: Shapiro-Wilk istatistiğine ait p değerinin 0,05'ten büyük olması, ilgili kümedeki değişkene ait verilerin normal dağılıma uygun bir dağılım sergilediğini göstermektedir.

Diğer taraftan, Tablo 8 ve Tablo 9'da verilen 11 değişkene ait ham verilerin her iki kümeleme algoritmasına göre hesaplanmış olan küme ortalamalarına bakıldığında, tüm değişkenlere ait küme ortalamalarının birbirlerinden belirgin bir biçimde farklılık gösterdikleri çıplak gözle dahi rahatça görülebilmektedir. Ancak, yine de bu durumun istatistiki olarak sınanması gerekmektedir.

Tablo 8. Değişkenlere Göre Küme Ortalamaları (Ortalamalar "Average" Yöntemi)

Küme / İşlemSayısı	Doğum	Ölüm	Evlenme	Boşanma	Kayıt Düzeltme	Adres Beyan	NKÖ	Yerleşim Yeri Belgesi	Nüfus Cüzdanı Düzenleme	Vatandaşlık	Diğer
1.Küme	42.666	11.850	18.626	4.566	1.023	167.419	473.786	219.839	227.468	515	358.775
2.Küme	13.325	4.693	6.454	1.236	329	58.312	232.446	122.091	70.730	425	125.263
3.Küme	10.287	1.398	3.261	149	477	23.210	107.985	53.135	40.386	102	61.277
4.Küme	4.071	1.891	2.227	487	79	23.883	94.859	55.312	24.222	143	58.870
5.Küme	1.904	674	935	107	52	10.905	52.144	21.748	9.335	177	30.332
6.Küme	238.071	63.837	112.941	29.094	5.488	1.124.236	2.531.447	1.736.221	1.500.964	1.525	1.763.332

Tablo 9. Değişkenlere Göre Küme Ortalamaları (WARD Yöntemi)

Küme / İşlemSayısı	Doğum	Ölüm	Evlenme	Boşanma	Kayıt Düzeltme	Adres Beyan	NKÖ	Yerleşim Yeri Belgesi	Nüfus Cüzdanı Düzenleme	Vatandaşlık	Diğer
1.Küme	31.065	8.593	13.324	2.880	699	111.497	376.362	179.173	157.574	483	232.580
2.Küme	11.216	3.927	5.485	995	266	51.663	203.951	105.986	59.225	429	110.742
3.Küme	13.867	1.770	4.311	158	665	29.140	137.367	66.378	53.466	70	64.650
4.Küme	4.705	2.020	2.474	512	108	24.806	101.328	56.692	26.727	171	63.731
5.Küme	2.219	986	1.200	212	44	15.500	65.509	36.856	13.265	113	37.060
6.Küme	123.060	38.555	60.980	17.541	2.963	627.141	1.379.892	855.438	794.255	957	1.182.535

Klasik varyans analizinin yapılabilmesi için kümelerde yer alan değişkenlere ait verilerin varyanslarının homojen bir dağılım sergilemesi gerekmektedir. Bu doğrultuda gerçekleştirilen *Levene Testi* (varyansların homojenliği testi) sonuçları Tablo 10'da verilmiştir. Tablodaki sonuçlar bize, ortalamalar (average) algoritmasına göre 6 kümede yer alan 11 değişkene ait verinin de homojen bir varyans yapısına sahip olduğunu göstermektedir. Ancak, WARD kümeleme algoritmasına göre ise Doğum, Evlenme, Yerleşim Yeri Belgesi, Nüfus Cüzdanı Düzenleme ve Vatandaşlık değişkenlerine ait verilerin homojen olmayan bir varyans yapısına sahip oldukları görülmektedir.

Tablo 10. Varyansların Homojenliği Testi

Yöntem /	**Ortalamalar (Average)**		**WARD**	
Değişken	**Levene Stat.**	**Sig.**	**Levene Stat.**	**Sig.**
Doğum	,412	,799	3,356	***,009***
Ölüm	,459	,766	1,007	,420
Evlenme	,120	,975	4,062	***,003***
Boşanma	1,429	,233	,976	,438
Kayıt Düzeltme	,205	,935	1,110	,362
Adres Beyan	1,803	,137	1,610	,168
Nüfus Kayıt Örneği	1,936	,113	1,625	,164
Yerleşim Yeri Belgesi	,380	,822	3,135	***,013***
Nüfus Cüzdanı Düzenleme	,163	,956	3,044	***,015***
Vatandaşlık	1,606	,181	2,381	***,046***
Diğer	1,871	,121	,620	,685

Not: Levene istatistiğine ait p değerinin 0,05'ten küçük olması, varyansların homojen olmadığını göstermektedir.

Levene testine göre, ortalamalar (average) algoritması ile elde edilen küme yapısındaki tüm değişkenlere ait veriler homojen bir varyansa sahip oldukları için klasik varyans analizi olan ANOVA'nın uygulanmasında bir sakınca yoktur. Bu doğrultuda ANOVA ile elde edilen sonuçlar Tablo 11'de yer almıştır. % 5 anlamlılık düzeyinde yapılan ANOVA testi, tüm değişkenler açısından küme ortalamalarının istatistiki olarak birbirinden farklı olduğunu göstermektedir (sig.<0,05).

Tablo 11. Ortalamalar Yöntemi – Varyans Analizi (F testi)

Değişken	**F**	**Sig.**
Doğum	***121,79***	,000
Ölüm	72,95	,000
Evlenme	***136,67***	,000
Boşanma	68,02	,000
Kayıt Düzeltme	81,78	,000
Adres Beyan	96,12	,000
Nüfus Kayıt Örneği	86,50	,000
Yerleşim Yeri Belgesi	51,35	,000
Nüfus Cüzdanı Düzenleme	***128,36***	,000
Vatandaşlık	13,78	,000
Diğer	62,74	,000

Not: F değeri daha büyük olan değişkenin, kümelerin oluşmasındaki belirleyicilik etkisi daha fazladır. Evlenme, Nüfus Cüzdanı Düzenleme ve Doğum İşlemleri değişkenleri en belirleyici olanlardır.

Levene testine göre, WARD algoritması ile elde edilen küme yapısına göre bazı değişkenlere ait veriler homojen bir varyansa sahip olmadıkları için klasik varyans analizi olan ANOVA sadece homojen varyansa sahip olan değişkenler için uygulanmış, homojen varyans varsayımını yerine getiremeyen değişkenler için ise *Welch Testi* uygulanmıştır. Bu doğrultuda ANOVA ile elde edilen sonuçlar Tablo 12'de ve Welch ile elde edilen sonuçlar ise Tablo 13'te yer almıştır. %5 anlamlılık düzeyinde yapılan her iki testin sonucuna göre, ilgili değişkenler açısından küme ortalamalarının istatistiki olarak birbirinden farklı olduğu görülmektedir (sig.<0,05).

Tablo 12. WARD Yöntemi – Varyans Analizi (F testi)

Değişken	F	Sig.
Ölüm	81,10	,000
Boşanma	78,13	,000
Kayıt Düzeltme	71,71	,000
Adres Beyan	90,68	,000
Nüfus Kayıt Örneği	74,13	,000
Diğer	71,23	,000

Not: Levene testi sonucunda varyansları homojen olan değişkenler için varyans analizi (ANOVA) uygulanmıştır.

Tablo 13. WARD Yöntemi – Welch Testi

Değişken	F	Sig.
Doğum	***91,11***	,000
Evlenme	***103,93***	,000
Yerleşim Yeri Belgesi	44,19	,000
Nüfus Cüzdanı Düzenleme	***98,33***	,000
Vatandaşlık	15,94	,000

Not: Levene testi sonucunda varyansları homojen olmayan değişkenler için Welch testi uygulanmıştır. F değeri daha büyük olan değişkenin, kümelerin oluşmasındaki belirleyicilik etkisi daha fazladır. Evlenme, Nüfus Cüzdanı Düzenleme ve Doğum İşlemleri değişkenleri en belirleyici olanlardır.

5. SONUÇ

Bu çalışmada, Türkiye'deki İl Nüfus ve Vatandaşlık Müdürlüklerinin hibrid kümeleme analizi aracılığıyla iş yoğunluklarına göre sınıflandırılması yapılmıştır. Elde edilen sonuçlar, kurumun ilgili görevlerini yerine getirebilmek için gereksinim duyacağı sayıda personeli istihdam edebilmesi yönünde alacağı kararlarda etkili olabilecektir. Analizde, hem fiili olarak uygulama alanı bulabilecek hem de küme içi benzerliği (homojenliği) ve kümeler arası farklılığı (hetorojenliği) maksimize edebilecek küme yapıları ortaya çıkarabilmek hedefi doğrultusunda, küme sayısı silhouette endeksi yardımıyla 6 olarak belirlenmiş ve karşılaştırma yapabilmek amacıyla da hem WARD hem de ortalamalar algoritmalarına göre sonuçlar elde edilmiştir. Ayrıca illerdeki iş yoğunluklarının farklı olmasına sebep olabilecek sosyoekonomik unsurların varlığı da göz önünde bulundurularak, elde

edilen küme yapılarının değerlendirilmesinde toplam nüfustaki pay, okuma yazma oranı, toplam katma değerdeki pay göstergelerinden de yararlanılmıştır.

Analizde, sırasıyla 1., 2., 3., 4., 5. ve 6. kümelerdeki il sayıları WARD algoritmasına göre 17, 20, 8, 22, 11 ve 3 olarak; ortalamalar algoritmasına göre ise, 13, 27, 9, 26, 5 ve 1 olarak bulunmuştur. Küme sayılarındaki farklılıklar, WARD algoritmasının uç değerli gözlemlere ortalamalar algoritmasına kıyasla daha hassas olmasından ileri gelmektedir. Bunun en açık delili olarak, WARD'a göre 6. kümede İstanbul, Ankara ve İzmir illerinin birlikte yer almasına karşın, ortalamalar yönteminde 6. kümede sadece İstanbul'un yer almış olması gösterilebilir. Elde edilen kümelerin genel yapılarına bakıldığında ise, 1. kümenin toplam nüfustaki pay, okuma yazma oranı, toplam katma değerdeki pay, toplam iş yükündeki pay ve ortalama iş yükü göstergeleri açısından 3 büyük metropol il dışındaki büyükşehirlerden oluştuğu görülmektedir. 2. kümede ise, söz konusu göstergeler açısından orta sırada yer alan orta gelişmişlik düzeyindeki iller yer almaktadır. 3. kümede, tüm göstergeler açısından son sıralarda yer alan doğu ve güneydoğudaki en az gelişmişlik düzeyine sahip olan iller yer almaktadır. 4. kümede ise, alt-orta sınıf olarak nitelendirilebilecek düzeydeki iller yer almaktadır. 5. küme, doğu ve kuzeydoğuda yer alan az gelişmiş illeri içermekte ve son küme olan 6. kümede ise, tüm göstergeler açısından aslan payına sahip olan ülkenin en büyük metropol illeri İstanbul, Ankara ve İzmir yer almaktadır. Elde edilen küme yapılarına göre 11 değişken çerçevesinde elde edilen küme ortalamalarının istatistiki olarak birbirlerinden yeterince farklı olup olmadıkları ANOVA ve Welch testleri ile sorgulanmış, küme yapılarının istatistiki olarak birbirlerinden farklı oldukları yani küme içi homojen ve kümeler arası da hetorojen bir yapı sergiledikleri ortaya çıkarılmıştır.

KAYNAKÇA

Alpar, R. (2011) "Uygulamalı Çok Değişkenli İstatistiksel Yöntemler", Ankara, Detay Yayıncılık.

Blashfield, R.K. (1976) "Mixture Model Tests of Cluster Analysis: Accuracy of Four Agglomerative Hierarchical Methods", Psychological Bulletin, 83(3): 377-388.

Bolshakova, N. ve Azuaje, F. (2003) "Cluster Validation Techniques for Genome Expression Data", Signal Processing, 83(4): 825-833.

Cao, F., Liang, J. ve Bai, L. (2009) "A New Initialization Method for Categorical Data Clustering", Expert Systems with Applications, 36(7): 10223-10228.

Chen, B., Tai, P.C., Harrison, R. ve Pan, Y. (2005) "Novel Hybrid Hierarchical-K-means Clustering Method (HK-means) for Microarray Analysis", IEEE Computational Systems Bioinformatics Conference, California-USA.

Chen, B., Rhodes, C., Kline, C. ve Irvin, L. (2010) "Protein Sequence Motif Information Generated by Fuzzy-Hybrid Hierarchical K-means Clustering Algorithm", International Conference on Bioinformatics & Computational Biology (BIOCOMP'10) Conference, Nevada-USA.

Chipman, H. ve Tibshirani, R. (2006) "Hybrid Hierarchical Clustering with Applications to Microarray Data, Biostatistics", 7(2): 286-301.

Everitt, B. S., Landau, S., Leese, M. Ve Stahl, D. (2001) "Hierarchical Clustering in Cluster Analysis", John Wiley & Sons, 5th Edition, Kings College, London, UK.

Ferreira, L. ve Hitchcock, D.B. (2009) "A Comparison of Hierarchical Methods for Clustering Functional Data", Comminications in Statistics – Simulation & Computation, 38(9): 1925-1949.

Fraley, C. ve Raftery, A.E. (1998) "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis", The Computer Journal, 41(8): 578-588.

Gan, G., Ma, C. ve Wu, J. (2007) "Data Clustering Theory, Algorithms and Applications", Philadelphia, Asa-Siam Series.

Hands, S, Everitt, B. (1987) "A Monte Carlo Study of the Recovery of Cluster Structure in Binary Data by Hierarchical Clustering Techniques", Multivariate Behavioral Res., 22(2): 235-243.

Hubert, L. (1974) "Approximate Evaluation Techniques for the Single-Link and Complete-Link Hierarchical Clustering Procedures", Journal of the American Statistical Association, 69(347): 698-704.

Kuiper, F.K. ve Fisher, L. (1975) "A Monte Carlo Comparison of Six Clustering Procedures", Biometrics, 31(3): 777-783.

MacQueen, J. (1967) "Some Methods for Classification and Analysis of Multivariate Observations", In Proceedings of The Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1: 281-297.

Milligan, G. W. (1980) "An Exemination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms", Psychometrika, 45(3): 325-342.

Milligan, G.W. ve Cooper, M.C. (1988) "A Study of Standardization of Variables in Cluster Analysis", 5(2): 181-204.

Murugesan, K. ve Zhang, C. (2011) "Hybrid Bisect K-means Clustering Algorithm", Business Computing and Global Informatization International Conference, Shanghai-China.

Ng, R. T. ve Han, J. (1994) "Efficient and Effective Clustering Methods for Spatial Data Mining", 20th International Conference on Very Large Data Bases, Santiago de Chile-Chile.

Özdamar, K. (1999) "Paket Programlar ile istatistiksel Veri Analizi 2", Eskişehir: Kaan Kitabevi.

Rao, A.R. ve Srinivas, V.V. (2008) "Regionalization of Watersheds: An Approach Based on Cluster Analysis", Water Science and Technology Library, Vol. 58, Springer, USA.

Rendon, E., Abundez, I., Arizmendi, A. ve Quiroz, M. (2011) "Internal versus External Cluster Validation Indexes", International Journal of Computers and Communications, 5(1): 27-34.

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

Singh, N. ve Singh, D. (2012) "Performance Evaluation of K-Means and Heirarichal Clustering in Terms of Accuracy and Running Time", (IJCSIT) International Journal of Computer Science and Information Technologies, 3(3): 4119-4121.

Syal, R., Prasad, G.V.S.R. ve Kumar, V.V. (2012) "A Novel Hybrid Clustering Algorithm: Integrated Partitional and Hierarchical Clustering Algorithm for Categorical Data", International Journal Computer Science & Emerging Technologies, 3(5): 138-146.

Templ, M., Filzmoser, P. ve Reimann, C. (2008) "Cluster Analysis Applied to Regional Geochemical Data: Problems and Possibilities", Applied Geochemistry, 23(8): 2198-2213.

Theodoridis, S. ve Koutroumbas, K. (2006) "Pattern Recognition", 3rd Ed., London, Academic Press.

Vijaya, P.A., Murty, M.N. ve Subramanian, D.K. (2006) "Efficient Bottom-up Hybrid Hierarchical Clustering Techniques for Protein Sequence Classification", Pattern Recognition, 39(12): 2344-2355.