IAAOJ, Scientific Science, 2015, 3(2), 29-37

MANOVA VARSAYIMLARI KARŞILANAMADIĞINDA ALTERNATİF BİR YÖNTEM OLARAK UZAKLIK ANALİZİ BİPLOT KULLANIMI*

B. Barış ALKAN[1], Cemal ATAKAN[2],

[1]Sinop Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, Sinop, Türkiye
[2]Ankara Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Ankara, Türkiye

E-posta: bbalkan@sinop.edu.tr

ÖZET

Çok değişkenli varyans analizi (MANOVA), g grup için ortalama vektörlerin eşitliği hipotezinin testi için kullanılır. MANOVA, varyans-kovaryans matrislerinin homojenliği ve çok değişkenli normallik varsayımlarına sahiptir. Bu varsayımlar ihlal edildiğinde MANOVA kullanılamaz. Bu çalışmada, MANOVA'ya alternatif bir yaklaşım olarak Gower ve Krzonowski (1999) tarafından önerilen ve uzaklık ölçülerine dayanan Uzaklık Analizi (UA) yönteminin teorik adımları yeniden gözden geçirilmiş ve gerçek bir veri kümesi üzerinden yöntemin önemi vurgulanmıştır.

Anahtar Kelimeler: Uzaklık analizi, biplot, çok değişkenli varyans analizi, temel koordinat analizi

USE OF DISTANCE ANALYSIS BIPLOT AS AN ALTERNATIVE METHOD IN CASE OF FAILURE TO PROVIDE THE MANOVA ASSUMPTIONS

ABSTRACT

Multivariate analysis of variance (MANOVA) is used for testing the null hypothesis of equal mean vectors for g groups. MANOVA assumes homogeneity of variance-covariance matrices and multivariate normality. We review theoretical steps of distance analysis method that is proposed by Gower and Krzonowski (1999) as an alternative approach to MANOVA. Moreover, the importance of the method is highlighted via a real data set.

Keywords: Distance analysis, biplot, manova, principal coordinates analysis

1. Giriş

Veri kümesinde heterojenlik, çarpıklık, kayıp veri gibi problemler söz konusuysa veya değişkenlerin sayısı gözlemlerin sayısından daha büyük ise, çok değişkenli tek yönlü varyans analizi (MANOVA) gibi geleneksel çok değişkenli teknikler, gruplar arasındaki farklılıklar belirleneceği zaman yüksek derecede kuşkulu sonuçlar verebilmektedir.

Böyle durumlarda tek yönlü MANOVA'ya alternatif bir çıkarımsal metot olarak Gower ve Krzonowski (1999) tarafından önerilen ve uzaklık ölçülerine dayanan Uzaklık Analizi (UA) yönteminin kullanımı araştırmacıların daha doğru ve güvenli bulgulara ulaşmalarını sağlayacaktır.

Gower ve Krzonowski (1999) çalışmalarında Temel Koordinat Analizi (TKA) ve Uzaklık Analizi (UA) olarak adlandırılan iki yaklaşımı ele almışlardır. TKA, Temel Bileşenler Analizine (TBA) paralel bir yaklaşım olup, analiz için gözlem çiftleri arasındaki uzaklıklardan oluşan matrisi kullanır. UA ise MANOVA'ya paralel bir yaklaşım olup, kareler ve çapraz çarpımların toplamlarından ziyade gözlemler arasındaki uzaklıkların parçalanması temeline dayanır (Fenty 2004).

*Bu çalışma 27-30 EKİM 2013 Tarihinde düzenlenen Uluslararası 8. İstatistik Kongresinde sunulan "MANOVA Varsayımlarının Sağlanmaması Durumunda Uzaklık Analizi Yönteminin Kullanımı" başlıklı sözlü bildirinin genişletilmiş halidir.

TKA, klasik metrik ölçekleme olarak da bilinmekte ve çok boyutlu ölçekleme (ÇBÖ) teknikleri ailesine aittir. Bu teknikler indirgenmiş bir uzayda çok değişkenli veri kümesinin görsel temsilini bulmak için tasarlanmıştır. Veri kümesinde yer alan değişkenler arasındaki benzerlikleri yansıtan bir koordinat sisteminde, dik eksenlere ait gözlemlerin bir grafiğini verir. Dik eksenler, indirgenmiş boyutlu uzayda (genellikle iki boyutlu) indirgemeden kaynaklanacak bilgi kaybını minimize edecek şekilde seçilir.
Temel varyans formülü, n tane gözlem noktası arasındaki karesel Öklid uzaklıkları toplamı olarak,

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2=\frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{t=1}^{n}\left(x_i-x_t\right)^2 \tag{1}$$

biçiminde yazılır.

Eşitlik (1), uzaklık analizi ile varyans analizi arasındaki bağlantıyı vermektedir. MANOVA, çapraz çarpımlar ve kareler toplamları matrisinin parçalanmasını temel alır. UA ise, ikili gözlem çiftleri arasındaki uzaklıkları içeren bir matrisin parçalanmasını temel alır. UA'da tek varsayım herhangi iki i ve t gözlemleri arasında bir uzaklığın (d_{it}) tanımlanabilmesidir. Ayrıca, tüm i ve t gözlemleri için $d_{it}=d_{ti}$ olmalıdır. d_{it}'yi hesaplamak için farklı uzaklık ölçüleri vardır (Fenty 2004).

İki gözlem arasındaki benzerlik, Öklid uzayında gözlemlerin grafiği ile gösterilebilir. Bu grafiksel yaklaşımda benzer gözlemler arasındaki uzaklık, benzer olmayan gözlemler arasındaki uzaklıktan daha az olacaktır. En temel uzaklık ölçüsü, Öklid uzaklığıdır. j boyut üzerinden ölçülen i gözleminin değeri x_{ij} ile gösterilsin.
Bu durumda i ve t gözlemleri arasındaki uzaklık, d_{it},

$$d_{it}=\sqrt{\sum_{j=1}^{p}\left(x_{ij}-x_{tj}\right)^2} \tag{2}$$

ile verilir.

Öklid uzaklığı, Minkowski uzaklıkları olarak bilinen uzaklık ailesinin bir üyesidir (Coxon 1982). Bu form,

$$d_{it}^{(h)}=\sqrt[h]{\sum_{j=1}^{p}\left|x_{ij}-x_{tj}\right|^h} \tag{3}$$

ile ifade edilir. (3) eşitliğinde h=2 olarak alındığında Öklid uzaklığı elde edilir. Sadece Öklid uzaklığı, reel çözümlere ulaşacağından TKA ve UA'de önemli bir rol oynar.

Bu çalışmada TKA, UA ve Biplot yöntemi ele alınmıştır. UA ile Biplot'un birlikte ele alınmasıyla oluşacak UA Biplot yöntemi, grupların ayrışmasını ve örtüşmesini açıklamak için hem örneklerin hem de değişkenlerin saklı kalmış yönlerini ortaya koyan grafiksel bir yaklaşım sunar.

2. Materyal ve Metot

2.1. Temel Koordinat Analizi

TKA çok fazla bilinen ve kullanılan bir yöntem değildir. Ancak genetik (Kosaki ve ark 1996), antropoloji (Fox ve ark 1996), moleküler biyoloji (He ve Haymer 1999) gibi alanlarda uygulamaları yapılmıştır. TKA'nin gelişiminde Householder ve Young (1938), Gower (1966), Krzanowski ve Marriot (1994)'nin çalışmalarının katkıları büyüktür. Young ve Householder (1938), gözlemler arasındaki uzaklıkların bir kümesinin koordinatlarını yeniden yapılandırma fikrini ilk olarak ortaya atmışlardır.

$\mathbf{X}$, $n\times p$ boyutlu elemanları x_{ij}'ler olan bir veri matrisi olsun. Burada n gözlem sayısı, p değişken sayısıdır. x_{ij} elemanı, j. boyutta ölçülen i gözleminin değeridir. p boyutlu bir Öklid uzayında n gözlemin koordinatları $\mathbf{x}_i$ $(i=1,2,\ldots,n)$ ile verilir. Burada $\mathbf{x}_i=\left(x_{i1},\ldots,x_{ip}\right)^T$ olarak tanımlanmıştır. i ve t gözlemleri arasındaki Öklid uzaklığı,

$$\begin{aligned} d_{it}^2&=\left(\mathbf{x}_i-\mathbf{x}_t\right)^T\left(\mathbf{x}_i-\mathbf{x}_t\right)\\ &=\mathbf{x}_i^T\mathbf{x}_i+\mathbf{x}_t^T\mathbf{x}_t-2\mathbf{x}_i^T\mathbf{x}_t \end{aligned} \tag{4}$$

ile verilir.

Gözlemler arasındaki uzaklıklar, $\mathbf{K}_{n\times n}=\left(\mathbf{X}_{n\times p}\right)\left(\mathbf{X}_{n\times p}\right)^{T}$ matrisi yardımıyla açıklanabilir.

Burada $\mathbf{X}=\left[\mathbf{x}_1,\ldots,\mathbf{x}_n\right]^{T}$, $n\times p$ boyutlu koordinatlar matrisidir.

$\mathbf{K}$ matrisinin rankı,

$$\mathrm{Rank}\left(\mathbf{K}\right)=\mathrm{Rank}\left(\mathbf{X}\mathbf{X}^{T}\right)=\mathrm{Rank}\left(\mathbf{X}\right)=p$$

biçiminde elde edilir. $\mathbf{K}$ matrisi simetrik, pozitif yarı tanımlı, rankı p olan bir matris olup, p tane negatif olmayan özdeğere ve $n-p$ tane sıfır özdeğere sahiptir.

$\mathbf{K}$ bir iç çarpım matrisi olup,

$$\left[\mathbf{K}\right]_{it}=k_{it}=\mathbf{x}_i^{T}\mathbf{x}_t \tag{5}$$

şeklinde tanımlanır.

Eşitlik (4),

$$\frac{1}{n}\sum_{i=1}^{n}d_{it}^{2}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^{T}\mathbf{x}_i+\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_t^{T}\mathbf{x}_t-\frac{2}{n}\sum_{i=1}^{n}\mathbf{x}_i^{T}\mathbf{x}_t \tag{6}$$

biçiminde yazılabilir.

Eğer veri kümesi sıfır ortalama ve birim varyansa göre standartlaştırılmış ise,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^{T}\mathbf{x}_t=0 \tag{7}$$

olacağından (6) eşitliği,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}d_{it}^{2}&=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^{T}\mathbf{x}_i+\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_t^{T}\mathbf{x}_t\\&=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^{T}\mathbf{x}_i+\mathbf{x}_t^{T}\mathbf{x}_t\end{aligned} \tag{8}$$

olarak yazılabilir.

Benzer şekilde,

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n}d_{it}^{2}&=\frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_i^{T}\mathbf{x}_i+\frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t^{T}\mathbf{x}_t\\&=\mathbf{x}_i^{T}\mathbf{x}_i+\frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t^{T}\mathbf{x}_t\end{aligned} \tag{9}$$

elde edilir.

Eşitlik (4) kullanılarak,

$$\begin{aligned}-2\mathbf{x}_i^{T}\mathbf{x}_t&=d_{it}^{2}-\mathbf{x}_i^{T}\mathbf{x}_i-\mathbf{x}_t^{T}\mathbf{x}_t\\\mathbf{x}_i^{T}\mathbf{x}_t&=-\frac{1}{2}\left(d_{it}^{2}-\mathbf{x}_i^{T}\mathbf{x}_i-\mathbf{x}_t^{T}\mathbf{x}_t\right)\end{aligned} \tag{10}$$

elde edilir. Buradan da,

$$\left[\mathbf{K}\right]_{it}=k_{it}=\mathbf{x}_i^{T}\mathbf{x}_t=-\frac{1}{2}\left(d_{it}^{2}-\mathbf{x}_i^{T}\mathbf{x}_i-\mathbf{x}_t^{T}\mathbf{x}_t\right) \tag{11}$$

biçiminde yazılabilir.

Eşitlik (8) ve (9) kullanılarak,

$$\mathbf{x}_t^T\mathbf{x}_t = \frac{1}{n}\sum_{i=1}^{n} d_{it}^2 - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{x}_i \tag{12}$$

$$\mathbf{x}_i^T\mathbf{x}_i = \frac{1}{n}\sum_{t=1}^{n} d_{it}^2 - \frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t^T\mathbf{x}_t \tag{13}$$

eşitlikleri elde edilir.

Eşitlik (11)'de (12) ve (13) eşitlikleri yerine yazılıp düzenlendiğinde,

$$\left[\mathbf{K}\right]_{it} = k_{it} = \mathbf{x}_i^T\mathbf{x}_t = -\frac{1}{2}\left(d_{it}^2 - \frac{1}{n}\sum_{t=1}^{n} d_{it}^2 - \frac{1}{n}\sum_{i=1}^{n} d_{it}^2 + \frac{2}{n}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{x}_i \right) \tag{14}$$

elde edilir.

Eşitlik (9) kullanılarak,

$$\begin{aligned} \frac{1}{n}\sum_{t=1}^{n} d_{it}^2 &= \mathbf{x}_i^T\mathbf{x}_i + \frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t^T\mathbf{x}_t \\ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n} d_{it}^2 &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{x}_i + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n}\mathbf{x}_t^T\mathbf{x}_t \\ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n} d_{it}^2 &= \frac{2}{n}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{x}_i \end{aligned} \tag{15}$$

eşitliği elde edilir.

Eşitlik (14)'deki $\frac{2}{n}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{x}_i$ ifadesi yerine (15)'de verilen karşılığı kullanılırsa,

$$\left[K\right]_{it} = k_{it} = \mathbf{x}_i^T\mathbf{x}_t = -\frac{1}{2}\left(d_{it}^2 - \frac{1}{n}\sum_{t=1}^{n} d_{it}^2 - \frac{1}{n}\sum_{i=1}^{n} d_{it}^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n} d_{it}^2 \right) \tag{16}$$

elde edilir. Burada, $\left[\mathbf{D}\right]_{it} = \frac{1}{2}d_{it}^2$ alınırsa, (16) eşitliği

$$\left[\mathbf{K}\right]_{it} = k_{it} = \mathbf{x}_i^T\mathbf{x}_t = -\left(D_{it} - D_{i\bullet} - D_{\bullet t} + D_{\bullet\bullet}\right) \tag{17}$$

biçiminde yazılabilir. Burada, $D_{i\bullet} = \frac{1}{n}\sum_{t=1}^{n} D_{it}$, $D_{\bullet t} = \frac{1}{n}\sum_{i=1}^{n} D_{it}$, $D_{\bullet\bullet} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n} D_{it}$ olduğundan, (17)

eşitliği,

$$\left[\mathbf{K}\right]_{it} = k_{it} = \mathbf{x}_i^T\mathbf{x}_t = -\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\mathbf{D}\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\right) \tag{18}$$

şeklinde yazılabilir ve $\mathbf{1} = (1,1,\ldots,1)^T$, n tane 1'den oluşan vektör, $\mathbf{I}$ ise $n \times n$ boyutlu birim matristir.

$\mathbf{K}$ matrisi simetrik ve pozitif yarı tanımlı bir matristir. Bu nedenle $\mathbf{K}$ matrisine spektral ayrışım uygulandığında,

$$\mathbf{K}_{n\times n} = \mathbf{V}_{n\times n}\mathbf{\Lambda}_{n\times n}\left(\mathbf{V}_{n\times n}\right)^T \tag{19}$$

elde edilir. Burada $\mathbf{\Lambda} = \mathrm{diag}\left(\lambda_1, \lambda_2 \ldots, \lambda_p, \ldots, \lambda_n\right)$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq \ldots \geq \lambda_n$, $\mathbf{K}$ matrisinin özdeğerlerinin oluşturduğu köşegen matris olup, $\mathbf{V} = \left[\mathbf{v}_1, \ldots, \mathbf{v}_n\right]$ ise $n \times n$ boyutlu $\mathbf{v}_i^T\mathbf{v}_i = 1$ olan özvektörler matrisidir.

$\mathbf{K}$ matrisi $n-p$ tane sıfır özdeğere sahip olduğundan (19) eşitliği,

$$\mathbf{K}_{n\times n} = \mathbf{V}_{n\times p}\mathbf{\Lambda}_{p\times p}\left(\mathbf{V}_{n\times p}\right)^{T} \tag{20}$$

biçiminde yeniden yazılabilir.

Eşitlik (20) yeniden düzenlenirse,

$$\mathbf{K}_{n\times n} = \left(\mathbf{V}_{n\times p}\mathbf{\Lambda}^{1/2}_{p\times p}\right)\left(\mathbf{V}_{n\times p}\mathbf{\Lambda}^{1/2}_{p\times p}\right)^{T} \tag{21}$$

olarak yazılabilir ve burada $\mathbf{X} = \mathbf{V}_{n\times p}\,\mathbf{\Lambda}^{1/2}_{p\times p}$ dir. Böylece gözlemlerin koordinatlarını veren p-boyutlu koordinat matrisi, $\mathbf{X}_{n\times p} = \mathbf{V}_{n\times p}\mathbf{\Lambda}^{1/2}_{p\times p}$ eşitliği ile bulunabilir. İki boyutlu bir grafiksel yaklaşım için koordinatlar ise, $\mathbf{X}_{n\times 2} = \mathbf{V}_{n\times 2}\mathbf{\Lambda}^{1/2}_{2\times 2}$ eşitliği ile elde edilir.

Başka bir deyişle, gözlemlerin koordinatları, $\mathbf{XX}^{T}$ matrisinin özdeğerleri ve ilişkili özvektörlerinin belirlenmesi ve daha sonra özdeğerlerin karekökünün oluşturduğu matris ile özvektörler matrisinin çarpılması ile bulunur. $\mathbf{X}$ veri kümesinin değişkenleri arasındaki uzaklık matrisinin tanımlanması ile $\mathbf{X}$'in indirgenmiş boyutta sütun değişkenlerinin grafiği bulunabilir (Park ve ark 2008).

Gower (1966), TBA yöntemi ile karışmaması için bu yönteme TKA adını vermiştir. Gower, TKA'nin TBA'nın bir eşi olduğunu, benzer çözümler içerdikleri, ancak TBA'da değişkenler arasındaki ilişkilerin $p\times p$ boyutlu varyans-kovaryans veya korelasyon matrisi temel alınırken, TKA'nın gözlemler arasındaki birlikteliklerin $n\times n$ tipinde bir uzaklık matrisini temel aldığını vurgulamıştır.

2.2. Uzaklık Analizi

n tane birimin g tane grup içine $n_1, n_2, \ldots, n_g$ büyüklüklerinde bölündüğünü varsayalım. Gruplandırma $n\times g$ boyutlu $\mathbf{G}$ matrisi ile gösterilebilir. i birimi r grubunda yer alıyorsa $g_{ir}=1$, diğer durumda $g_{ir}=0$ biçimindedir. $\mathbf{N} = \mathrm{diag}\left(n_1,\ldots,n_g\right)$, $n_1, n_2, \ldots, n_g$ grup birim sayılarından oluşan köşegen bir matristir. $\mathbf{m} = \left(n_1,\ldots,n_g\right)$ ise, grup birim sayılarını içeren bir vektördür.

Grup ortalamalarının koordinatları ise,

$$\bar{\mathbf{X}}_{g\times p} = \mathbf{N}^{-1}_{g\times g}\left(\mathbf{G}_{n\times g}\right)^{T}\mathbf{X}_{n\times p} \tag{22}$$

eşitliği ile elde edilir.

Gower ve Krzanowski (1999), toplam karesel uzaklıklar toplamını $(\mathbf{T})$, grup-içi $(\mathbf{W})$ ve gruplar-arası $(\mathbf{B})$ karesel uzaklıklar toplamlarının toplamı olarak $\mathbf{T = W + B}$ biçiminde ayrıştırmışlardır.

Burada,

$$\mathbf{T} = \frac{1}{n}\mathbf{1}^{T}\mathbf{D}\mathbf{1} \tag{23}$$

$$\mathbf{W} = \sum_{r=1}^{g}\frac{1}{n_r}\mathbf{1}_r^{T}\mathbf{D}_{rr}\mathbf{1}_r \tag{24}$$

$$\mathbf{B} = \frac{1}{n}\mathbf{m}^{T}\,\bar{\mathbf{\Delta}}\,\mathbf{m} \tag{25}$$

dir.

$\mathbf{D}$ matrisi g^2 tane alt matrise bölünür, $\mathbf{D}_{rs}$ $(r,s=1,\ldots,g)$. Burada $n_r\times n_s$ boyutlu $\mathbf{D}_{rs}$, s. gruptaki her bir gözlem ve r. gruptaki her bir gözlem arasındaki 2 ile bölünmüş karesel uzaklıkları içerir. Eşitlik (24)'deki $\mathbf{D}_{rr}$, r gruptaki gözlemlerin tüm çiftleri arasındaki uzaklıkları içeren $\mathbf{D}$ matrisinin alt matrisidir.

Eşitlik (25)'deki $\bar{\Delta}$, $g \times g$ boyutlu simetrik bir matris olup, r ve s grup ortalamaları arasındaki 2 ile bölünmüş karesel uzaklığın (r,s)'inci elamanını verir (Gardner ve ark 2005).

2.3. Uzaklık Analizi ile Biplot Kullanımı

Uzaklık analizi ile Biplot birlikte oluşturulabilir. UA Biplot, değişken sayısı, gözlem sayısından fazla ise etkileyici boyut indirgeme özelliklerine sahiptir (Gardner ve ark 2009). Büyük veri kümeleri için MANOVA varsayımlarının sağlanması zor olduğundan, UA kullanımı bu tür veri kümeleri için önemlidir. UA, grup ortalamaları arasındaki uzaklıkları maksimize eder. Grup yapısının grafiksel olarak incelenmesi için UA Biplot kullanılır.

Grup ortalamaları için $\bar{\mathbf{X}}$ koordinatları,

$$-\left(\mathbf{I}_g - \frac{\mathbf{1}\mathbf{1}^T}{g}\right)\mathbf{D}^*\left(\mathbf{I}_g - \frac{\mathbf{1}\mathbf{1}^T}{g}\right) = \bar{\mathbf{X}}\bar{\mathbf{X}}^T \tag{26}$$

eşitliğinin spektral ayrışımından bulunabilir. Burada, $\mathbf{D}^* = \mathbf{N}^{-1}\mathbf{G}^T\mathbf{D}\mathbf{G}\mathbf{N}^{-1}$'dir. Bu değer TKA kullanılarak elde edilebilir (Gardner ve ark 2005). Eksenlerin orijini grup genişlikleri ile ağırlıklandırılmamış grup ortalamaları noktalarının merkezindedir.

Alternatif olarak ağırlıklandırılmış TKA,

$$-\left(\mathbf{I}_g - \frac{\mathbf{1}\mathbf{m}^T}{n}\right)\mathbf{D}^*\left(\mathbf{I}_g - \frac{\mathbf{m}\mathbf{1}^T}{n}\right) = \bar{\mathbf{X}}\bar{\mathbf{X}}^T \tag{27}$$

eşitliği ile de ifade edilebilir.

Bu durumda, grafiksel gösterimde, eksenlerin orijini n gözlem noktalarının merkezindedir. İki boyutlu grafiksel yaklaşımda, gözlem noktalarını ve onlara karşılık gelen grup ortalamalarını göstermek için,

$$-\frac{1}{2}\left(\bar{\mathbf{X}}^{*T}\bar{\mathbf{X}}^*\right)^{-1}\bar{\mathbf{X}}^{*T}\left(\mathbf{I}_g - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\left(\mathbf{d}_i - \mathbf{d}_0\right) \tag{28}$$

eşitliği kullanılır (Gardner ve ark 2005).

Eşitlik (28)'deki $\bar{\mathbf{X}}^*$, (27) eşitliğinin spektral ayrışımından bulunan $\bar{\mathbf{X}}$'nın iki boyutta yaklaşımını gösterir. Burada, $\mathbf{d}_i$ vektörü, grafik üzerinde g grup ortalamasının her birine i. gözlem noktasının karesel uzaklıklarını içerir. $\mathbf{d}_0$ ise, her bir grup ortalamasının sırasıyla eksenlerin orijinine olan karesel uzaklığıdır.

Eğer ağırlıklandırılmış TKA kullanılırsa (28) yerine,

$$-\frac{1}{2}\left(\bar{\mathbf{X}}^{*T}\bar{\mathbf{X}}^*\right)^{-1}\bar{\mathbf{X}}^{*T}\left(\mathbf{I}_g - \frac{1}{n}\mathbf{1}\mathbf{m}^T\right)\left(\mathbf{d}_i - \mathbf{d}_0\right) \tag{29}$$

eşitliği kullanılır. UA biplot elde etmek için, orijinal değişkenler ile ilgili bilgi, ölçeklendirilmiş eksenler şeklinde eklenmektedir (Gardner ve ark 2005).

3. UYGULAMA

Ham petrol ve ürünleri günümüzde bir çok alanda kullanılan çok önemli bir enerji kaynağı olarak kabul edilir. Petrol karbon ve hidrojen elementlerinden meydana gelen hidrokarbonlardan oluşur. Karbon ve hidrojen elementleri çok çeşitli ve karmaşık molekül yapıları ortaya çıkarır.

Petrolde, birçok metal element yer alır. En bol bulunan elementler vanadyum ve nikeldir. Petroldeki vanadyum ve nikel miktarlarının (5-40g/ton) petrolün asfalten içeriğinin artması ile arttığı bilinmektedir (Uysal 2006).

Bu uygulamada, Gerrild ve Lantz (1969) tarafından elde edilen ham petrol örneklerinin analizini içeren veri kümesinin Johnson ve Wichern (2002)'de verilen düzenlenmiş şekli kullanılmıştır. Veri kümesi, üç kumtaşı bölgesinden (Bölge I: Wilhelm, Bölge II: Aşagi-Mulinia, Bölge III: Yukarı-Mulinia) alınan 56 ham petrol örneklerinin vanadyum, demir, berilyum, doymuş hidrokarbonlar, aromatik hidrokarbonlar kimyasal özelliklerini içermektedir. Söz konusu değişkenler ile direkt petrol kalitesini belirlemeden ziyade petrolün kökeni (denizel vs.) ve oluşum koşulları gibi yorumlar yapılabilmektedir. Bunun yanında direkt olarak petrol kalitesi, API (American Petroleum Institute) denilen bir indeksle belirlenmekte bunun hesabı ise farklı yöntemlerle elde edilen parametrelere göre yapılmaktadır (Uysal 2006).

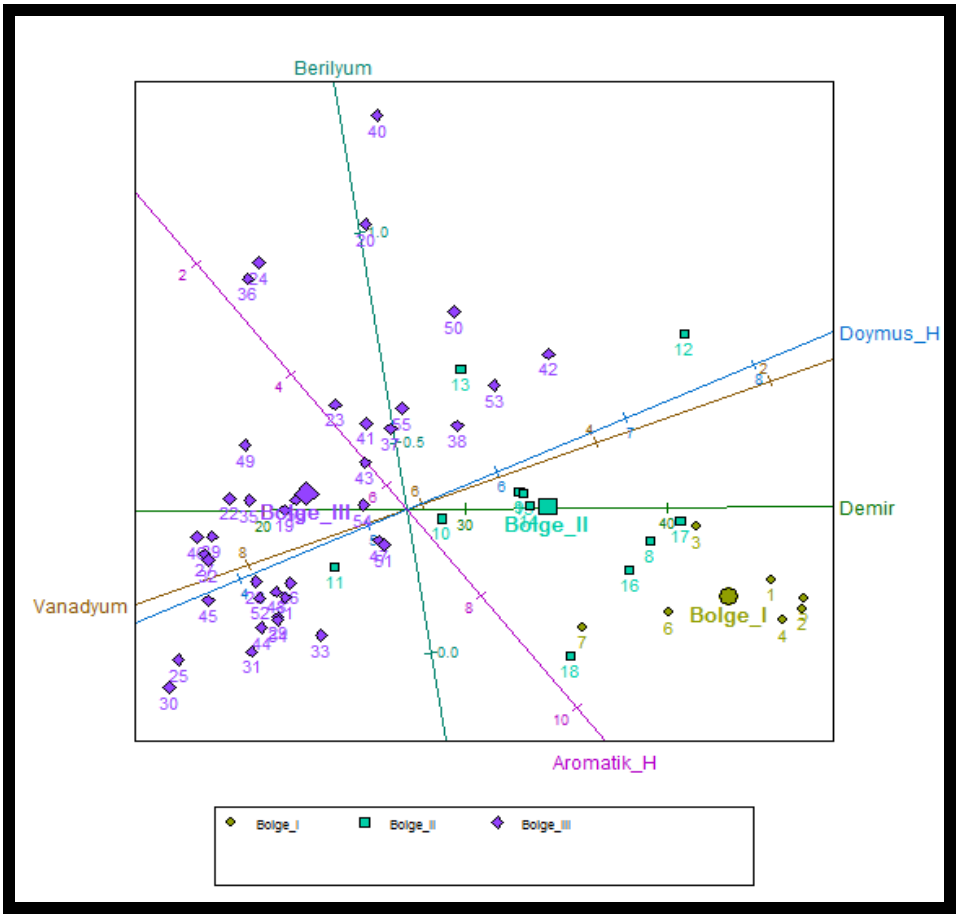
Veri kümesine R programında yer alan kütüphane ve fonksiyonların kullanımı ile UA Biplot uygulanmış ve Şekil 1'de verilen grafiksel yaklaşım elde edilmiştir.

Şekil 1, bölgeleri optimal bir şekilde ayırmayı ve gözlem noktaları arasındaki uzaklıkları korumayı amaçlamaktadır. Burada, her bir eksen değişkenlerin orijinal birimlerinde derecelendirilmiştir.

Tablo 1 incelendiğinde, Berilyum ve Doymuş Hidrokarbonlar değişkenlerinin UA Biplot grafiğinde iyi temsil edildiği, Aromatik Hidrokarbonlar değişkenin ise çok iyi temsil edilmediği görülmektedir. Bu durum, Aromatik Hidrokarbonlar değişkeni üzerinden yapılacak kestirimlerin çok fazla tutarlı olmayacağı yorumunu beraberinde getirmektedir. Düşük eksen kestirimlerine sahip değişkenlerden elde edilecek kestirimlerin daha yorumlanabilir olması için üç boyutlu bir grafiksel gösterim düşünülebilir. Şekil 1'de gösterilen gözlem noktaları tüm bölge içi varyasyonun iki boyutta bir yaklaşımıdır. İki boyutlu bu grafiksel yaklaşım ile toplam varyansın %65'i açıklanmaktadır.

Şekil 1, sadece bölge ortalamalarının yer aldığı UA biplot grafiğini verir. Bu grafiğin beş değişken için üç bölge ortalamasını ayırdığı görülmektedir. Grafiksel yaklaşımda, bölge ortalama noktalarından eksenlere dik izdüşümler çizilerek elde edilecek uzaklıklara göre eksenler üzerinden bölge ortalamaları yorumlanır.

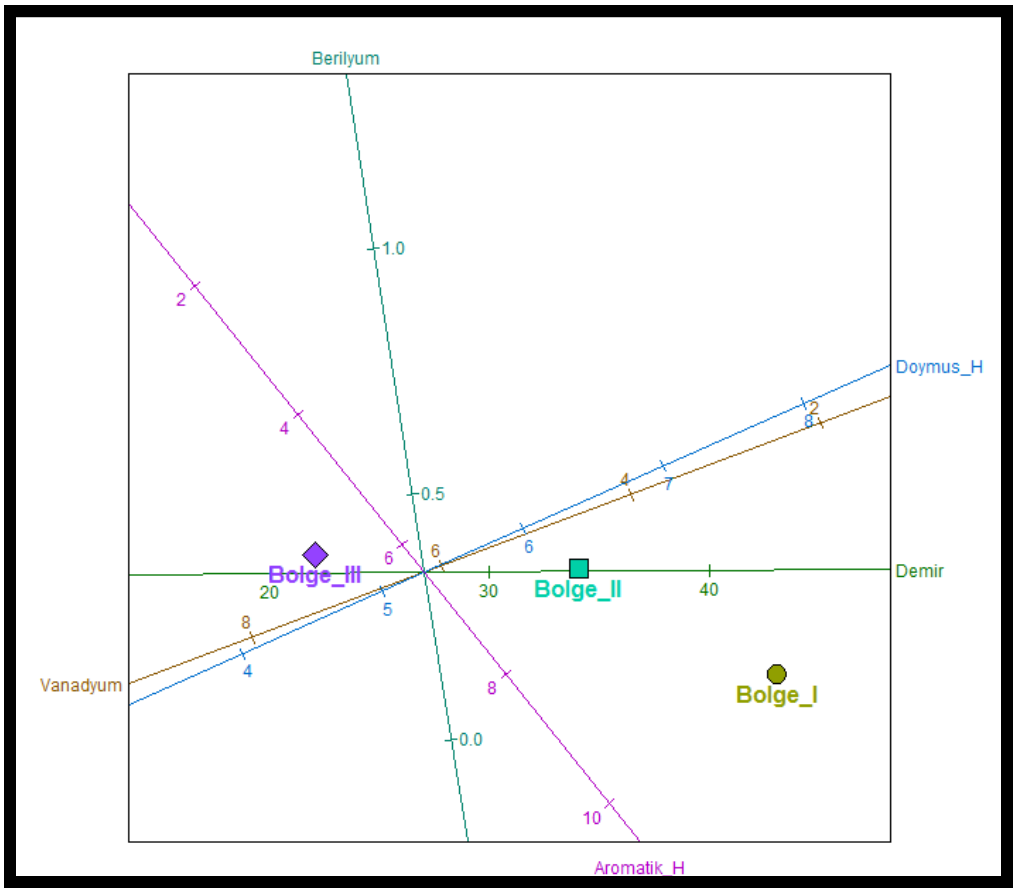
Buna göre Şekil 2'de verilen grafiksel yaklaşım incelendiğinde Bölge II ve Bölge III ortalamalarının Berilyum değişkeni açısından fazla bir farklılık göstermediği görülmektedir. Aromatik_H, Vanadyum, Doymuş_H, Demir değişkenleri Bölge I ve Bölge II'ye göre Bölge III'de farklı bir yapı sergilemektedir. Demir, Aromatik_H ve Doymuş_H değişkenleri bakımından bölge ortalamaları arasındaki farklılıklar incelendiğinde Bölge I'de bu değişkenlerin etkilerinin fazla olduğu Şekil 2'den yorumlanabilmektedir. Vanadyum değişkeni Bölge III ortalaması üzerinde etkiliyken, Doymuş_H değişkeni Bölge I ve Bölge III ortalamaları üzerindeki etkilidir. Bu üç bölge ortalaması düşünüldüğünde Demir oranı bakımından en zengin Bölge I iken, en fakir Bölge III olarak görülmektedir.



Şekil 1. Ham petrol veri kümesi için UA Biplot grafiği

Tablo 1. UA Biplot grafiğinde eksenlerin kestirimleri

Vanadyum	Demir	Berilyum	Doymuş Hidrokarbonlar	Aromatik Hidrokarbonlar
0.60	0.59	0.83	0.79	0.40



Şekil 2. Bölge ortalamaları için UA Biplot grafiği

4. SONUÇ

Bu çalışmada, Temel Koordinat Analizi (TKA), Uzaklık Analizi (UA) ve Biplot yöntemleri incelenmiş ve gerçek bir veri kümesi üzerinden MANOVA varsayımlarının sağlanmaması durumunda uzaklık analizi kullanımının önemi vurgulanmıştır. Uzaklık analizi, dağılımsal varsayımlara göre değişmemektedir. Üstelik küçük örneklemlerde ve grup varyans-kovaryans matrislerinin heterojenliğinden etkilenmez. Ayrıca bir hipotez testi gerektiği zaman permütasyon test prosedürleri kullanılabilir. UA biplot, mevcut istatistiksel çıkarım prosedürleri ile birlikte ele alınması durumunda, veri analizi için daha da önemli bir araca dönüşür. Uzaklık analizi sonucu elde edilen grafiksel yaklaşımda, grup farklılıkları daha az belirgin olduğunda, gruplar arasındaki örtüşmenin derecesinin belirlenmesi ve ilgili gruplar arasındaki farklılıkların istatistiksel önemini saptamak için bootstrap veya permütasyon hipotez testi gibi parametrik olmayan çıkarım teknikleri uygulanmaktadır.

Kaynaklar

Coxon, A. P. M. 1982. The User's Guide to Multidimensional Scaling. London: Heinnemann.

Fox, C. L., A. G. Martin, and S. V. Civit. 1996. Cranial variation in the Iberian Peninsula and the Balearic Islands: Inferences about the history of the population. American Journal of Physical Anthropology 99(3): 413–428.

Fenty, J.2004. Analyzing distances. The Stata Journal. 4(1):1-26.

Gardner S, Le Roux NJ, Rypstra T, Swart JPJ 2005. Extending a Scatterplot for Displaying Group Structure in Multivariate Data: A Case Study, ORiON, 21(2), 111-124.

Gardner-Lubbe S, Roux N J, Maunders H, Shah V, Patwardhan S. 2009. Biplot methodology in exploratory analysis of microarray data, Statistical Analysis and Data Mining, 2(2), 135 – 145.

Gerrild P.M. and Lantz R.J. 1969.Chemical Analysis of 75 Crude Oil Samples from Pliocene Sand Units, ELK Hills Oil Field,. U.S. Geological Survey Open-File Report, California.

Gower, J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika, Vol. 53, pp. 325–338.

Gower J.C. and Krzanowski W.J. 1999. Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance, Applied Statistics, Vol. 48, pp. 505-519.

Householder, A.S. and Young, G. 1938. Matrix approximation and latent roots, American Mathematical Monthly, Vol. 45, pp. 165-171.

He, M. and Haymer, D. S. 1999. Genetic relationships of populations and the origins of new infestations of the Mediterranean fruit fly, Molecular Ecology, Vol. 8(8), pp. 1247-1257.

Johnson, R.A. and Wichern, D.W. 2002. Applied Multivariate Statistical Analysis. Prentice-Hall, New Jersey.

Kosaki K., Jones M. C. and Stayboldt, C. 1996. Zimmer phocomelia: Delineation by principal coordinate analysis, American Journal of Medical Genetics, Vol. 66(1), pp. 55-59.

Krzanowski, W.J. and Marriott, F.H.C. 1994. Multivariate Analysis, Part I: Distributions, Ordination and Inference. London: Arnold.

Uysal, A. 2006. Ham petrol fraksiyonlarının biyolojik bozunma sonrası fizikokimyasal özelliklerinin değişimi, Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü, Isparta.