n<p Boyutlu Biyolojik Verilerde Farklı Kümeleme Yöntemlerinin Karşılaştırmalı Olarak İncelenmesi*

İrfan ÖZTÜRK[1], Necati YILDIZ[2]
[1] HÜ, Teknik Bilimler Meslek Yüksek Okulu, Şanlıurfa
[2] Bingöl Üniv. Ziraat Fak. Zootekni Bölümü, Bingöl

Geliş Tarihi (Received): **22.08.2012** **Kabul Tarihi (**Accepted): **22.09.2012**

Özet: Araştırma, n<p boyutlu olan, 24 farklı Antepfıstığı (Pistacia vera L.) tipi ve bu tiplerden alınan 38 adet değişken üzerinde yapıldı. Ancak bu tip bir veri matrisine kümeleme analizine ilişkin bazı çok değişkenli test istatistiklerinin uygulanabilmesi için değişken sayısının (n≥p) azaltılması gerekmektedir. Değişken sayısının azaltılmasında Temel Bileşenler (Principal Component) analizi, Ayırma (Diskriminant) analizi ve Korelasyon analizinden yararlanılmıştır. Sözü edilen metotlarla indirgenen değişkenler, farklı kümeleme metotlarıyla karşılaştırmalı olarak incelendi. Sonuçta, kümelemede en uygun yöntemin Temel Bileşenler analizi ile birlikte kullanılan Ward metodunun olduğu saptandı. Küme sayısının belirlenmesinde ise en uygun ölçütün C_{max}, Wilks Lambda ve Hotelling Lawley İz istatistiklerinin olduğu belirlenmiştir.
Anahtar Kelimeler: Kümeleme Analizi, Sayısal Sınıflandırma, Temel Bileşenler Analizi, Ayırım Analizi, Korelasyon Analizi

A Comparison of Different Clustering Methods on the Biological Data with n<p Dimensions

Abstract : This research was carried out on 38 variables from 24 types of pistachio (Pistacia vera L.) with n<p dimensions. However, in order to apply the some multivariate test statistics of clustering analysis to this type data matrix, the number of variables (n≥p) must be decreased. Principal Component, Discriminant and analysis of correlation were used to decrease the number of variables. The reduced number of variables by using these mentioned methods, was comparatively evaluated by using different clustering methods. Present results shown that the most suitable method for clustering is the Ward method used together with Principle Components variable reducing technique. It was also found out that the most suitable measurement index to determine the number of clusters were Wilk's Lambda, C_{max} and Hotelling Lawley Trace statistics.
Keywords: Cluster Analysis, Numerical Taxonomy, Principal Component Analysis, Discriminant Analysis, Correlation Analysis

GİRİŞ

Türlerin alt bölümleri arasındaki ekolojik ve genetik etkilerin tümünün oluşturduğu varyasyon, türlere ait populasyon yapılarını ortaya koymaktadır (Jain,1975).

Morfolojik varyasyonların ıslah çalışmalarında saptanması, değişik tür ve çeşitler arasındaki taksonomik ilişkileri aydınlatmak bakımından büyük bir önem arz etmektedir. Çünkü yetiştirilen türler içerisinde bulunan varyasyonların bilinmesi ve bu varyasyonun dağılış durumu, ıslah programlarının uygulanması açısından çok önemlidir (Bliss, 1981).

Türler arasındaki varyasyonun saptanmasında sayısal sınıflandırma (Numerical Taxonomy) diğer bir ifade ile kümeleme analizi metotları önemli rol oynamaktadır. Bu metotlar ilk defa 1957'de Sneath tarafından organizmalardaki varyasyonu saptamak ve daha somut bir sınıflama yapmak maksadıyla kullanılmıştır. Bitkiler ve hayvanların doğru ve nicel olarak sınıflandırılmaları için uygun istatistik analizlerin kullanılması gerekmektedir.

Özellikle gen bankalarında korumaya alınan materyalin benzer ve benzer olmayan özellikleri ile ilgili kayıtlarını analiz etmek kolay değildir. Çünkü benzerliğin ve farklılığın saptanmasında genetik, agronomik, morfolojik, kimyasal vb. pek çok özellik dikkate alınmaktadır. Çok fazla sayıda özelliğin incelendiği durumlarda, her bir özelliğin ölçülebilir verilerle ifadesi, bu ifadelerin benzerliğini veya farklılığını ortaya koyacak bir dizi hesap işlemlerinin yapılmasını gerektirmektedir (Tan, 1983).

İncelenen özellik sayısı arttıkça, benzerlik veya ayrılığı ortaya koyan indekslerin hesaplama işlemleri karmaşık ve zaman alıcı bir hal almaktadır. Çünkü klasik istatistiksel yöntemler, karakter adedi ve karşılaştırılan örnek sayısı arttıkça yetersiz kalmaktadır. Bu sorun, fazla sayıdaki karakteri sınıflandırma indeksleri kapsamında formüle eden kümeleme analizinin ortaya konması ile büyük ölçüde çözüme kavuşmuştur.

* Doktora tezinden alınmıştır.
*Sorumlu yazar: Öztürk, İ., ozirfan23@yahoo.com

Kümeleme analizi sadece hayvan ve bitki genetiği kaynaklı materyalin değerlendirilmesinde kullanılmakla kalmayıp, ıslah çalışmalarında da genetik varyasyonu ortaya koyma, hat seçimi, ırk saptama, hibrit indeksleme, konukçu - patojen - vektör ilişkilerini belirleme, toprak sınıflamaları için saptanan özelliklerden yararlanarak toprak haritaları yapmak amacıyla da kullanılmaktadır (Tan, 1983).

Günümüzde Biyoloji, Ziraat ve Tıp alanında yapılan araştırmalardan elde edilen birçok özelliği kapsayan karmaşık veriler, çok değişkenli analiz yöntemleri ile değerlendirilmektedir (Johnson and Wichern, 1992).

Çok değişkenli analiz yöntemlerinden en yaygın olarak kullanılanları; Çok Değişkenli Varyans analizi (Multivariate Analysis of Variance), Faktör analizi (Factor Analysis), Temel Bileşenler analizi (Principal Component Analysis), Ayırım analizi (Discriminant Analysis), Uyum analizi (Correspondence Analysis), Çok boyutlu ölçekleme (Multidimensional Scaling) ve Kümeleme analizi (Cluster Analysis) olarak sıralanabilir (Anderberg, 1973; Özdamar,1999).

Veri matrisindeki birim ve/veya değişkenler hakkında ayrıntılı şekilde bilgi edinilmek isteniyorsa, bunların en çok benzer olanları bir araya getirilerek sınıflandırılırlar. Sınıflandırma sonucunda aynı grup içindeki birimlerin (bireyler) benzerliği fazla iken diğer gruplardakilerle olan benzerlikleri daha azdır (Anderberg, 1973).

Kümeleme analizinden en yüksek yararı elde etmek için, aşağıdaki hususlara dikkat edilmesi gerektiğini Anderberg, (1973) belirtmiştir;

a) Verilerde küme yapısının olmaması veya tek bir küme yapısının oluşması gibi özel durumlarla karşılaşılabileceğini unutmamak gerekir.

b) Aynı veri setine farklı kümeleme yöntemleri uygulandığında farklı sonuçlar verebilir. Bu sebepten tek bir kümeleme yöntemine bağlı kalmayıp birden fazla yöntemin denenmesi daha güvenilir sonuçlar verecektir.

c) Sınıflandırma sonucunda, karmaşık bir veri yapısı ile karşılaşılabilir

d) Kümeleme yöntemleri hipotez testleri için amaçtır.

Çalışmanın amacı verileri en uygun kümelere ayırmaktır. Ancak veri setinde bireylere ait çok fazla değişken (p) bulunması; verileri kümelere ayırmada, yorumlamada, elde edilen kümelere çok değişkenli istatistik analizlerinin uygulanmasında, anlamsız değişkenlerin bireyleri yanlış kümelerde sınıflandırması gibi sıkıntılar ortaya çıkarabilecektir. Bu nedenle öncelikle değişkenlerin boyutu azaltılarak birimlerin kümelenmesine çözüm arandı.

Ayrıca bireylerin çok boyutlu uzaydaki dağılımına bağlı olarak farklı kümeleme yöntemlerine göre farklı sonuçlar elde edilebileceğinden, sınıflandırma yapılırken birden fazla kümeleme yöntemi kullanılarak sonuçlar değerlendirecektir.

MATERYAL ve METOT

Bu çalışmanın verileri, Şanlıurfa TİGEM Ceylanpınar Tarım İşletmesinde Acar, (1997) tarafından yapılan araştırmadan elde edilmiştir. Ölçümler, 24 farklı erkek Antepfıstığı (Pistacia vera L.) tipleri üzerinde gerçekleştirilmiş olup, fıstık ağaçlarını türlerine ayırmak amacıyla 38 adet farklı karakterden ölçüm alınmıştır.

Ancak bilindiği üzere kümeleme analizi ile birlikte bazı Çok Değişkenli test istatistiklerinin uygulanabilmesi için veri matrisinin en azından (n≥p) boyutunda olması gerekmektedir. Değişken sayısının (p) birim (fıstık tipleri) sayısından (n) fazla olduğu durumlarda, (n<p durumunda) verileri analiz edebilmek için, değişken sayısının bazı boyut indirgeme yöntemleri ile azaltılması gerekmektedir. Bu nedenle araştırmada Korelasyon analizi, Temel Bileşenler analizi ve Stepwise Diskriminant analizi ile değişkenlerin boyutu (n≥p)'ye indirgenerek gerçek değişkenlerle elde edilen kümelemeye en yakın boyut indirgeme yöntemi belirlenmeye çalışıldı. Bununla birlikte yukarıda sözü edilen kümeleme yöntemlerinden en yaygın olarak kullanılanları, değişken boyutu indirgenmiş 24 farklı fıstık ağacı tipi üzerine uygulanarak en iyi küme yapısını ortaya koyan yöntem tespit edilmeye çalışıldı.

Değişken boyutu indirgeme işleminden sonra farklı kümeleme yöntemlerine göre elde edilen kümeler, küme sayısını belirlemede kullanılan C_{max}, k, ve M_{min} ölçütleri ile Wilk's Lambda ve Hotelling Lawley İz istatistikleri kullanılarak test edilip en iyi küme yapısı belirlenecektir.

Kümeleme analizinde kullanılan yöntemlerin teorik temelleri, karşılaştırmalarda kolaylık sağlamak amacıyla aşağıda verildi.

Sayısal sınıflandırma analizinde nicel veriler için kullanılan uzaklık ölçütü, Minkowsky ölçütü olup aynı zamanda genel uzaklık ölçüsüdür (Anderberg, 1973).

x_i ve x_j (p) tane karaktere göre ölçülmüş *i.* ve *j.* bireyler olarak tanımlandığında, Minkowsky ölçütü;

$$d_a(x_i, x_j) = \left[\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^a \right]^{1/a} \quad , \quad \lambda \geq 1 \text{ için;}$$

(*i, j* = 1,2, . . . , n) olarak verilebilir.

Bu ölçüt Minkowsky uzaklık ölçütünün, λ=2 olduğu durumdaki özel hali Öklid uzaklık ölçütünü verir (Anderberg, 1973; Tatlıdil, 1992).

$$d_2(x_i, x_j) = \left[\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^2 \right]^{1/2}$$

Mariott (1974) ve Kendall (1975)'ın Öklid uzaklığında gözlem değerleri standardize edildikten sonra kullanıldığından, karakterlerin farklı ölçü birimleri ile ölçülmüş olmasından etkilenmediğini bildirmektedir (Özdamar, 1988). Bu üstünlüğünden dolayı, kümeleme analizinde sıklıkla kullanılan uzaklık ölçüsü Öklid olduğunu, ayrıca gruplar arası hata kareler ortalamasını minimize ettiği için genellikle tercih edildiğini bildirmiştir (Tatlıdil, 1992; Huang, 2001).

Kümeleme kriterleri

Test istatistikleri	Wilk's Lambda	Hotelling Lawley İz	C_{max}	M_{min}	k
Formüller	$\Lambda = \frac{\lvert W \rvert}{\lvert W+B \rvert}$	$T_o^2 = \sum_{r=1}^{p} \lambda_r$	$C = \left[\frac{iz(B)}{k-1}\right] \Big/ \left[\frac{iz(W)}{N-k}\right]$	$M = k^2 \lvert W \rvert$	$k \cong \left(\frac{n}{2}\right)^{1/2}$
Kritik değer	$i^2_{p(k-1),\sim}$	$i^2_{p(k-1),\sim}$			

Kümeleme Yöntemleri

1-*Tek Bağlantı (en yakın komşu) kümeleme yöntemi*

d*(u,v)w* = min {d*uw*, d*vw*}

2-*Tam Bağlantı (en uzak komşu) kümeleme yöntemi*

d*(u,v)w* = max {d*uw*, d*vw*}

3- *Gruplar İçi Ortalama Bağlantı kümeleme yöntemi*

T*u*, (*u*) kümesi içindeki bireylerin tüm mümkün çiftleri arasındaki benzerliklerin toplamı, N*u*' da (*u*) kümesi içindeki bireylerin sayısı olarak kabul edilecek olursa; kümeler birleşmeden önce T*u* =0 ve N*u* =1 olacaktır. (*u*) ve (*v*) birleştirildiğinde yeni oluşacak (*uv*) kümesi içindeki bireylerin tüm mümkün çiftleri arasındaki benzerliklerin toplamı,

T*(uv)* = T*u* + T*v* + d*uv* ile

(*uv*) kümesindeki birimlerin sayısı ise

N(*uv*) =N*u* + N*v*

şeklinde hesaplanır.

(*u*) ve (*v*) kümelerine ait çiftlerin benzerlik ölçülerinden ve birey sayılarından yararlanarak en benzer çiftin bulunmasında grup içi benzerlik ortalaması;

$$\bar{d}uv = \frac{T(uv)}{N(uv)\cdot\frac{(N(uv)-1)}{2}} = \frac{Tu+Tv+duv}{(Nu+Nv)\cdot\frac{(Nu+Nv-1)}{2}}$$

olarak hesaplanabilir.

Hesaplanan bu ortalamalar minimum olduğunda (*u*) ve (*v*) kümeleri birleştirilir. (*u*) ve (*v*)'nin satır ve sütunları D matrisinden çıkarılır. Bu işlemler tekrarlanmak suretiyle tüm bireylerin kümelerde birleştirilmesi sağlanıncaya kadar kümeleme işlemi devam eder (Anderberg, 1973; Tatlıdil, 1992; Mac Queen, 1967).

4- *Gruplar Arası Ortalama Bağlantı kümeleme yöntemi*

$$\bar{d}uv = \frac{duv}{Nu \cdot Nv}$$

olarak hesaplanır.

5- *Merkezi (Centroid) Bağlantı kümeleme yöntemi*

$$d(uv)w = \frac{Nu\cdot duw + Nv\cdot dvw}{N(uv)} - \frac{Nu\cdot Nv\cdot duv}{N(uv)^2}$$

olarak hesaplanır.

6. *Ortanca (Median) Bağlantı kümeleme yöntemi*

$$d(uv)w = \frac{1}{2}\cdot duw + \frac{1}{2}\cdot dvw - \frac{1}{4}\cdot duv$$

olarak hesaplanır.

7. *Ward kümeleme tekniği*

$$d(uv)w = \frac{\{(Nw+Nu)\cdot duw + (Nw+Nv)\cdot dvw - Nw\cdot duv\}}{Nw+N(uv)}$$

olarak hesaplanır.

8. *McQuitty Bağlantı kümeleme yöntemi*

$$d(uv)w = \frac{d(uw)+d(vw)}{2}$$

olarak hesaplanır.

9. *k-Ortalama Yöntemi*:

$X_1, X_2, X_3, \ldots, X_n$ bireyin her birinden elde edilen *p* karakterli gözlem vektörleri, çok boyutlu X uzayında birer nokta olarak düşünülecek olursa ve aynı uzayda, her grup birey için küme merkezleri olarak,

$a_{1n}, a_{2n}, \ldots, a_{kn}$; $j = 1,2,\ldots,k$ seçildiğinde;

$$W_n = \frac{1}{n}\sum_{i=1}^{n} \min_{1\le j\le k} \left|\left| X_i - a_{jn} \right|\right|^2$$

kuralına göre bireyler en yakın kümeye sınıflanmaktadır (Hawkins et al., 1982; Minitab, 1995; Blashfield, 1976).

Kümeleme analizinden önce değişken sayısının azaltılması

Çalışmamızda değişken sayısı kümelenecek birim sayısından fazla olduğundan (n=24)<(p=38), bazı çok değişkenli analizlerin yapılabilmesi için değişken sayısının azaltılması (n≥p) gerekmektedir.

Orijinal değişkenlerin ölçüm değerlerinin, değişim aralıklarının ve ölçü birimlerinin fazla olduğu durumlarda, değişken (p) sayısının birim (n) sayısından çok fazla olduğu (n<p) vb. durumlarda korelasyon yada kovaryans matrislerini nonsingular (tekil olmayan matris) hale getirmek için veri indirgemesi yapmak ve kümeleme analizi uygulamak için temel bileşen skorları hesaplayarak Temel Bileşenler analizinden yararlanılmaktadır (Özdamar, 1999).

Bu durumları göz önüne alarak aşağıda sıralanacak boyut indirgeme yöntemleri kullanıldı:

1. Temel Bileşenler analizi (Principal Component Analysis)

Aralarında korelasyon bulunan p sayıda orijinal değişkenin açıkladığı varyans yapısını daha az sayıda (p>k) orijinal değişkenlerin doğrusal bileşenleri olan değişkenlerle ifade etme yöntemine Temel Bileşenler analizi denilir. Temel bileşenler, bizzat kendileri bir sonuç olmaktan ziyade sonuç almayı sağlayıcı özelliğe sahiptir. Çünkü Temel bileşenler daha geniş incelemeler için bir ara adım özelliği taşımaktadırlar (Özdamar, 1999; Sharma, 1996).

İstatistik analizlerde değişkenler arasında önemli düzeyde yüksek birlikte değişim ya da korelâsyonların bulunması arzu edilmediğinden, veri setinin bir şekilde bu değişkenlerden arındırılarak kullanılması daha uygun olacaktır (Özdamar, 1999).

Temel Bileşenler analizi ile en etkili yani varyansın büyük bir bölümünü açıklayan Temel Bileşene ait skorlar belirlenerek, kümeleme analizinde boyutu indirgenmiş değişkenler olarak kullanıldı.

2. Korelasyon analizi

Değişkenler arasındaki ikili korelasyon katsayıları hesaplanarak, aralarında yüksek derecede ilişki bulunan değişkenler veri setinden çıkarıldı ve kalan değişkenlerin yardımı ile birimler (Fıstık tipleri) sözü edilen kümeleme yöntemleri ile sınıflandırıldı.

3. Adımlı Ayırım (Stepwise Diskriminant) analizi

Veri setlerine diskriminant analizinin uygulana bilmesi için, veri setlerinin MANOVA (çok değişkenli varyans analizi) uygulaması için gerekli varsayımları taşıması gerekmektedir (Özdamar, 1999). Bu varsayımlar;

a- Veri matrisi çok değişkenli normal dağılım göstermeli

b- Değişkenlerin varyans ve kovaryansları homojen olmalı

c- Değişkenlerin ortalamaları ve varyansları arasında korelasyon bulunmamalı

d- Veri seti grupların birbirinden ayrılmasında rol oynamayacak gereksiz değişkenleri içermemelidir.

Söz konusu varsayımlar sağlandığında; birimlerin belirli kümelere atanmasında en etkili olan değişkenler aşamalı olarak *Adımlı Ayırım (*Stepwise Diskriminant) analizi ile veri setine dâhil edilir. Böylece kümelemede en az etkisi olan değişkenler veri setinden elemine edilmiş olur.

ARAŞTIRMA BULGULARI ve TARTIŞMA

Bu çalışmada, Acar, (1997) tarafından Şanlıurfa Ceylanpınar Tarım İşletmesinde (TİGEM) yapılan bir araştırma sonucunda elde edilen veriler kullanıldı. Ölçümler 24 farklı fıstık tipine ait erkek fıstık ağaçları (Pistacia vera L.) üzerinde gerçekleştirildi. Fıstık ağaçlarını türlerine ayırmak amacıyla 38 adet farklı karakter için ölçüm alınmıştır. Bu karakterler (değişkenler) morfolojik ve fenolojik gözlemlere bağlı olarak Acar, (1997) tarafından geleneksel metotlara dayanarak teşhis edilmiştir. Fıstık tipleri erkenci ve geççi olmak üzere temel iki gruba ayrılmaktadır.

Bununla birlikte erkenci ve geççi tipler arasında yer alan ve bazen erkenci tiplerle, bazen de geççi tiplerle kümelenen bir kısım melez tipler orta geççi tipler olmak üzere üçüncü bir grup oluşturmuştur. Sözü edilen fıstık tipleri Çizelge 1'de verilmiştir.

Fıstık tiplerine ilişkin değişkenlerin bir kısmı aralık ölçeğinde (sürekli ve kesikli) değişken özelliği gösterirken, orantı şeklinde özelliklerde vardır. Bunun yanı sıra bazı karakterler adlandırma (Nominal), bazıları da sıralama (ordinal) ölçeğinde değer alan özellikler olup, bu tip değişkenlere puanlar verildi. Kümeleme analizinde kullanılmak üzere her bir karakter (C1,C2,…,C38 olarak) yeniden tanımlanıp, bu karakterlere ilişkin değişken tiplerini Çizelge 2'de görmek mümkündür.

Daha öncede belirtildiği gibi veriler, farklı ölçüm ölçeğine sahip ise en uygun uzaklık ölçüsü Öklid (Euclid) veya Öklid uzaklığının karesi (karesel Öklid) dir.

Çizelge 1. Kümeleme Analizinde Kullanılan Fıstık Tipleri ve Kod Numaraları

Erkenci Fıstık Tipleri	Orta Geççi Fıstık Tipleri	Geççi Fıstık Tipleri
3	1	7
4	2	13
5	8	14
6	9	17
10	12	18
11	16	19
15	22	20
24	23	21

Çizelge 2. Karakterlere İlişkin Tanımlama ve Değişken Tipleri

Tanımlama	Karakterler	Değişken Tipi
C1	Büyüme Biçimi	Adlandırma
C2	Sürgün Uzunluğu	Sürekli (Oran ölçeği)
C3	Dal Sayısı	Kesikli
C4	Sürgün Rengi	Nominal (Sıralama)
C5	Dal Kalınlığı	Sürekli (Oran ölçeği)
C6	Yaprak Uzunluğu	Sürekli (Oran ölçeği)
C7	Yaprak Genişliği	Sürekli (Oran ölçeği)
C8	Yaprakçık Sayısı	Kesikli
C9	Uç Yaprak Uzunluğu	Sürekli (Oran ölçeği)
C10	Uç Yaprak Genişliği	Sürekli (Oran ölçeği)
C11	Uç Yaprak İriliği	Sıralama
C12	Uç Yaprak Şekli	Adlandırma
C13	Uç Yaprak Tepesi	Adlandırma
C14	Uç Yaprak Tabanı	Adlandırma
C15	Uç Yaprak Kenarı	Adlandırma
C16	Yaprak Rengi	Adlandırma
C17	Karagöz Ağırlığı	Sürekli (Oran ölçeği)
C18	Karagöz Şekli	Adlandırma
C19	Karagöz Rengi	Adlandırma
C20	Sürgündeki Karagöz Sayısı	Kesikli
C21	Karagöz Uzunluğu	Sürekli (Oran ölçeği)
C22	Karagöz Genişliği	Sürekli (Oran ölçeği)
C23	Gözlerin Kabarması	(Aralık ölçeği.)
C24	Gözlerin Patlaması	(Aralık ölçeği)
C25	Çiçeklenme Başlangıcı	(Aralık ölçeği)
C26	Tam Çiçeklenme	(Aralık ölçeği)
C27	Çiçeklenme Sonu	(Aralık ölçeği)
C28	Çiçeklenme Süresi	Kesikli
C29	Salkım Sayısı	Kesikli
C30	Salkım Uzunluğu	Sürekli (Oran ölçeği)
C31	Bir Salkımdaki Çiçek Sayısı	Kesikli
C32	Verim Potansiyeli	Sürekli (Oran ölçeği)
C33	Çiçek Tozu Üretimi	Sürekli (Oran ölçeği)
C34	Salkım Ağırlığı	Sürekli (Oran ölçeği)
C35	Bir Çiçekteki Anter Sayısı	Kesikli
C36	Bir Anterdeki Çiçek Tozu Sayısı	Kesikli
C37	Çiçek Tozu Canlılık Oranı	Orantı
C38	% 20 Sakkarozda Çiçek Tozu Çimlenme Oranı	Orantı

Veri setinde değişken sayısı birim (denek) sayısından fazla olduğundan ($n<p$), bu tip bir veri matrisine kümeleme analizine ilişkin bazı çok değişkenli test istatistiklerinin uygulana bilmesi için değişken sayısı Temel Bileşenler, Adımlı Ayırım (Stepwise Diskriminant) ve Korelasyon analizleri ile ($n\geq p$) azaltıldı.

Değişkenler (p) arası korelasyon analizi sonuçlarından yararlanarak, aralarında yüksek derecede (**: $P<0.01$) ilişki bulunan değişkenler tespit edilerek veri setinden çıkarıldı. Bu suretle hem fıstık ağaçlarının sınıflandırılmasında birbirinden bağımsız olan karakterler (değişkenler) tespit edilmiş, hem de veri matrisinin boyutu ($n\geq p$) indirgenmiştir (Ancak yapılan analizler de korelasyonla elemine edilen değişkenlerle elde edilen kümeleme dağılımları olumlu sonuçlar vermediğinden, detayının verilmesine gerek görülmedi. Bununla birlikte Stepwise Diskriminant analizi ile yapılan boyut indirgemede de olumlu sonuçlar alınmadığından detayı verilmedi).

Üçüncü yöntem olarak, Temel Bileşenler analizinden yararlanarak $n<p$ boyutunda olan veri matrisinin boyutu indirgenerek değişken sayısı azaltıldı. Bireylerin kümelenmesinde en etkili ve en fazla varyasyonu açıklayan Temel Bileşenler ise Çizelge 3 deki gibi elde edildi.

Çizelge 3. Temel Bileşenler (Principal Component Analysis) Analizi Sonuçları

Bileşkeler	Özdeğer	Varyans	Eklemeli Varyans (Cumulative)
1	***9.7703***	***0.257***	***0.257***
2	***7.0284***	***0.185***	***0.442***
3	***4.6561***	***0.123***	***0.565***
4	***2.9303***	***0.077***	***0.642***
5	***2.1037***	***0.055***	***0.697***
6	***1.8353***	***0.048***	***0.745***
7	***0.0440***	***0.044***	***0.789***
8	***0.0380***	***0.038***	***0.828***
9	***0.0370***	***0.037***	***0.864***
10	***0.0270***	***0.027***	***0.892***
11	0.0230	0.023	0.914
12	0.0200	0.020	0.935
13	0.6365	0.017	0.952
14	0.5900	0.016	0.967
15	0.4076	0.011	0.978
16	0.2923	0.008	0.986
17	0.1746	0.005	0.990
18	0.1576	0.004	0.994
19	0.0890	0.002	0.997
20	0.0595	0.002	0.998
21	0.0428	0.001	0.999
22	0.0272	0.001	1.000
23	0.0000	0.000	1.000
24	0.0000	0.000	1.000
.	.	.	.
.	.	.	.
38	0.0000	0.000	1.000

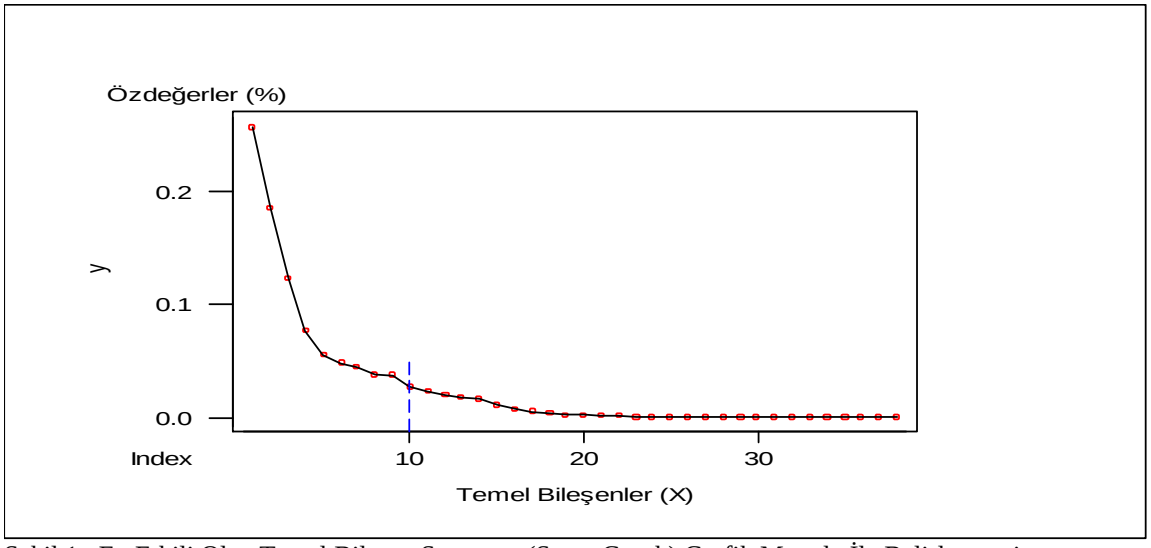
Analizde kullanılacak en etkili Temel Bileşenlerin sayısını belirlemek amacı ile özdeğerlerin, varyansı açıklama oranlarından yararlanarak Temel Bileşenlere (Ana Bileşenlere) ilişkin grafik (Scree Graph) çizdirildi. Böylece, Temel Bileşenlerin sayısı Şekil 1'de verilen grafik vasıtasıyla belirleme imkânı doğmuştur.

Grafik (Scree graph) vasıtasıyla belirlenen ve varyansın % 89'unu açıklayan ilk on temel bileşene ait skorlar kümeleme analizinde boyutu indirgenmiş değişkenler olarak kullanıldı (Çizelge 4).

Küme yapılarını ve küme sayılarını belirlemeden önce ilk iki temel bileşen vasıtasıyla da birimlerin dağılımını ve kaç kümeye ayrılabileceklerini Şekil 2'de çizilmiş olan grafik vasıtasıyla tahmin etmek mümkündür.

Temel Bileşenler, Adımlı Ayırım (Stepwise Diskriminant) ve Korelasyon analizi ile boyutu indirgenmiş olan değişkenlere yukarıda sözü edilen kümeleme analizi metotları ayrı ayrı uygulandı. Elde edilen küme yapıları incelendiğinde bu dağılıma en çok uygunluk gösteren yöntemin, Temel Bileşenler analizi ile boyutu indirgenmiş değişkenlerin Ward metodu ile birlikte kullanımı sonucu olduğu saptandı (Şekil 3). Nedenine gelince, uygulanan tüm kümeleme yöntemlerine ait dendogramlar incelendiğinde erkenci ve geççi fıstık tiplerini birbirinden en iyi ayıran yöntemin Ward yöntemi olduğu bunu Tam Bağlantı ve k-ortalama yöntemi izlemiştir.

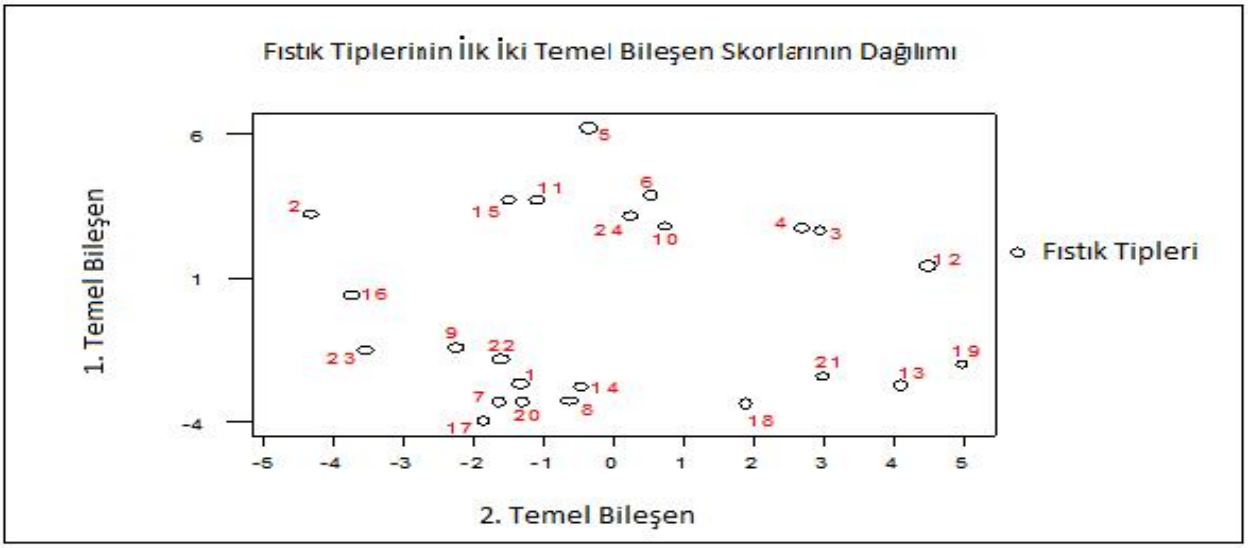
Temel Bileşenler analizi ile boyutu indirgenen Ward ve Tam bağlantı yöntemi Şekil 3 ve Şekil 4'de karşılıklı incelendiğinde erkenci ve geççi çeşitlerin yaklaşık benzer bir şekilde sınıflandırıldığı gözlenmektedir.



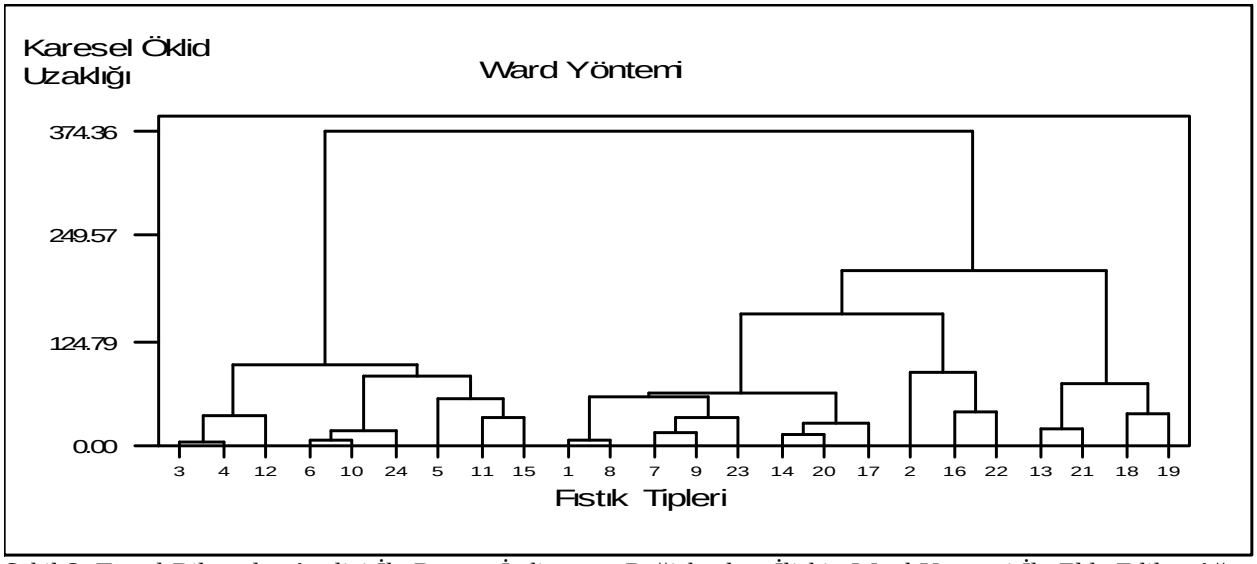
Şekil 1. En Etkili Olan Temel Bileşen Sayısının (Scree Graph) Grafik Metodu İle Belirlenmesi

Çizelge 4. Temel Bileşenler Analizi İle Elde Edilen Yeni Değişkenlere İlişkin Scorlar

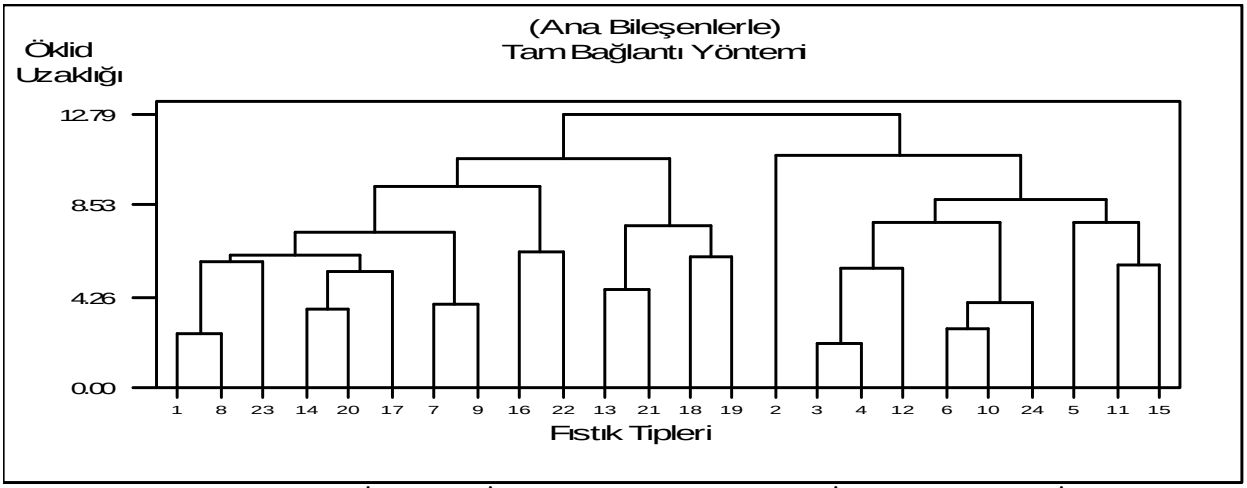
Fıstık Tipleri	Değişken Scorları									
	1	D2	D3	D4	D5	D6	D7	D8	D9	D10
1	-2,644	-1,323	-3,851	2,313	0,182	-0,801	0,019	0,627	0,308	0,408
2	3,267	-4,355	3,286	3,647	-0,416	1,094	-0,028	2,418	-0,314	1,265
3	2,679	2,944	-0,218	-1,818	-1,070	-1,115	0,158	0,268	1,584	0,970
4	2,764	2,684	0,311	-1,545	-0,020	-1,646	-0,908	0,727	1,156	0,011
5	6,254	-0,356	-1,688	-0,543	2,606	1,377	0,735	-1,921	0,325	2,310
6	3,898	0,548	-3,331	-0,093	-1,253	0,013	-0,420	0,511	-1,166	-1,058
7	-3,288	-1,644	-1,489	-1,327	0,042	1,620	-0,026	2,295	1,730	-0,290
8	-3,229	-0,641	-3,921	1,029	0,259	-0,108	-0,307	-0,456	-0,918	1,135
9	-1,410	-2,259	-1,226	-2,717	0,371	-1,011	0,760	1,816	0,572	-0,011
10	2,855	0,729	-2,309	1,665	-2,190	0,875	-1,212	-0,034	-1,152	-0,947
11	3,751	-1,105	1,301	-0,560	-0,697	-2,126	-0,105	0,182	-0,603	-1,640
12	1,450	4,513	0,652	1,617	1,358	-0,294	-2,294	-0,877	1,339	-0,118
13	-2,719	4,122	-0,872	1,634	-0,272	0,953	2,986	0,294	0,568	-0,313
14	-2,761	-0,471	1,485	0,952	0,854	-0,519	-0,505	0,480	0,745	-0,238
15	3,739	-1,500	1,275	-0,187	3,957	0,956	0,777	0,353	0,462	-1,977
16	0,405	-3,719	3,862	-0,561	-1,546	-0,072	-0,234	-0,960	-0,559	0,765
17	-3,967	-1,859	-0,031	1,297	0,835	0,830	-2,446	-1,842	0,388	-1,255
18	-3,343	1,897	1,751	-1,269	-0,133	-0,526	-2,333	0,772	-0,592	1,638
19	-1,953	5,001	1,874	-1,771	1,304	1,825	0,174	1,008	-3,511	-0,082
20	-3,267	-1,293	1,014	-0,161	0,380	-2,554	0,988	-1,184	-0,598	-0,261
21	-2,365	3,013	2,968	2,340	-0,756	-0,536	2,141	-0,949	0,730	-0,057
22	-1,804	-1,611	0,808	-2,812	-2,438	3,127	0,096	-1,667	1,247	-0,482
23	-1,509	-3,552	-1,207	-1,322	1,273	-1,218	0,995	-1,199	-1,368	0,201
24	3,201	0,237	-0,444	0,191	-2,631	-0,142	0,987	-0,662	-0,372	0,025



Şekil 2. İlk İki Temel Bileşen Vasıtasıyla Elde Edilen Grupların (Birimlerin) Dağılımı



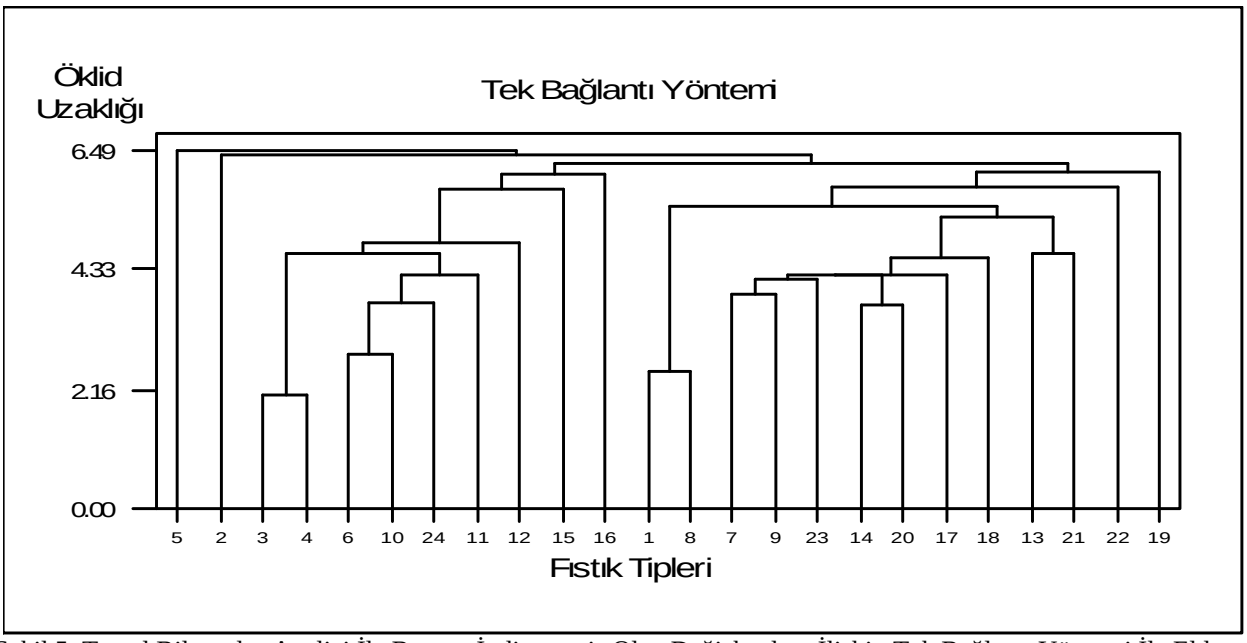
Şekil 3. Temel Bileşenler Analizi İle Boyutu İndirgenen Değişkenlere İlişkin Ward Yöntemi İle Elde Edilen Ağaç Grafiği



Şekil 4. Temel Bileşenler Analizi İle Boyutu İndirgenmiş Olan Değişkenlere İlişkin Fıstık Tipleri İle Tam Bağlantı Yöntemine Göre Elde Edilen Ağaç Grafiği

Ancak Şekil 5'de Tek Bağlantı yöntemine göre elde edilen dendogram Çizelge 1'de verilen fıstık tipleri ile karşılaştırıldığında erkenci ve geççi tiplerin birbiri ile karıştırıldığı görülmektedir. Diğer kümeleme yöntemleri de benzer bir durum sergilediğinden ayrıntılı dendogramları burada vermeye gerek görülmedi.

Temel Bileşenler analizi ile elemine edilen değişkenler, Çizelge 5'te verilen Hotelling Lawley İz ölçütüne göre değerlendirildiğinde; fıstık tiplerinin ($p<0.001$) önem düzeyinde 2, 3, 4, 5 kümeye ayrılabileceğini, ancak 2 kümeye ayrılması durumunda F değerinin daha büyük olduğunu görmekteyiz. Şekil 3 incelendiğinde fıstık tiplerinin erkenci ve geççi olmak üzere temel 2 kümeye ayrıldıkları görülmektedir. Bir alt kümede (3, 4 ve 5' inci kümelerde) ise fıstık tiplerinin birbirinden farklı özelliklerde alt türlere ayrıldıkları görülmektedir.



Şekil 5. Temel Bileşenler Analizi İle Boyutu İndirgenmiş Olan Değişkenlere İlişkin Tek Bağlantı Yöntemi İle Elde Edilen Ağaç Grafiği

Çizelge 5. Temel Bileşenler Analizi İle Boyutu İndirgenmiş Değişkenlere İlişkin Farklı Küme Sayılarının Hotelling Lawley Trace (İz) Ölçütü İle Önemlilik Düzeylerinin Belirlenmesi

Kümeleme Yöntemi	Küme Sayısı	Hotelling Lawley İz Ölçütü	Test İstatistiği F	Önemlilik Düzeyi P
k-Ortalama	2	15.56241	**20.231**	0.000<0.001
	3	15.59810	8.579	0.000<0.001
	4	23.41867	7.546	0.000<0.001
	5	73.53192	15.626	0.000<0.001
Ward Yöntemi	2	33.56928	**43.640**	0.000<0.001
	3	37.27117	20.499	0.000<0.001
	4	55.35712	17.837	0.000<0.001
	5	68.05547	14.462	0.000<0.001
Tam Bağlantı	2	22.56820	**29.339**	0.000<0.001
	3	40.64567	22.355	0.000<0.001
	4	44.54142	14.352	0.000<0.001
	5	61.49244	13.067	0.000<0.001

Temel Bileşenleri esas alarak kümeleme yöntemleri kendi aralarında önemlilik seviyelerine göre değerlendirecek olursa; fıstık tipleri 2 ve 4 kümeye ayrıldığında en iyi küme yapısını Ward yönteminin verdiğini Çizelge 6'da görmemiz mümkündür. Ancak, Fıstık tipleri 3 kümeye ayrıldığında Tam Bağlantı yönteminin en anlamlı küme yapısını verdiğini yine aynı cetvelden görebiliriz. Fıstık tipleri 5 kümeye ayrıldığında ise k-Ortalama yöntemine ait F değerinin daha büyük olduğu ve diğer yöntemlerden daha anlamlı küme yapısını verdiğini görmekteyiz.

Farklı boyut indirgeme metotlarına göre elemine edilen değişkenlerle Fıstık tipleri kümelendikten sonra, elde edilen kümelere bazı çok değişkenli test

istatistikleri uygulanarak Çizelge 7'de ki sonuçlar elde edildi. Daha öncede sözü edildiği gibi Temel Bileşenler analizi ile boyutu indirgenen değişkenler en iyi sonucu verdiği tespit edildi. Çizelge 7 incelendiğinde C_{max}, Wilk's Lambda ve Hotelling Lawley İz istatistikleri Fıstık tiplerinin iki kümeye ayrılması gerektiği görülmektedir. Bu durumu Şekil 2 ve Şekil 3'ün incelenmesinde görülebilmektedir.

Çizelge 6. Temel Bileşenler Analizi İle Boyutu İndirgenmiş Değişkenlere İlişkin Farklı Küme Sayılarının Wilk's Lambda Test İstatistiği İle Önemlilik Düzeylerinin Belirlenmesi

Kümeleme Yöntemi	Küme Sayısı	Wilk's Lambda Ölçütü	Test İstatistiği F	Önemlilik Düzeyi P
K-Ortalama	2	0.06038	**20.231**	0.000<0.001
	3	0.01975	7.339	0.000<0.001
	4	0.00272	7.121	0.000<0.001
	5	0.00009	10.695	0.000<0.001
Ward Yöntemi	2	0.02893	**43.640**	0.000<0.001
	3	0.00627	13.956	0.000<0.001
	4	0.00083	11.209	0.000<0.001
	5	0.00013	9.629	0.000<0.001
Tam Bağlantı	2	0.04243	**29.339**	0.000<0.001
	3	0.00576	14.612	0.000<0.001
	4	0.00120	9.777	0.000<0.001
	5	0.00022	8.220	0.000<0.001

Çizelge 7. Kümeleme Kriterleri İle Küme Sayısının Ward Yöntemine Göre Belirlenmesi

Değişken Tipleri	Küme Sayıları	C_{max}	M_{min}	Wilk's Lambda		Hotelling Lawley İz	
				F	p	F	P
Korelasyon analizi ile elemine edilen değişkenler	2	5.84	20.76E+30	**14.035**	0.000	**14.035**	0.000<0.001
	3	**17.39**	28.41E+29	6.338	0.000	6.887	0.001<0.01
	4	12.54	21.37E+28	7.845	0.000	7.286	0.000<0.001
	5	9.13	**12.73E+27**	6.918	0.000	5.697	0.000<0.001
Temel Bileşenler analizi ile elemine edilen değişkenler	2	**6.98**	20.76E+30	**43.640**	0.000	**43.640**	0.000<0.001
	3	6.31	28.41E+29	13.956	0.000	20.499	0.000<0.001
	4	6.08	21.37E+28	11.209	0.000	17.837	0.000<0.001
	5	5.56	**12.73E+27**	9.629	0.000	14.462	0.000<0.001
Stepwise Diskriminant analizi ile elemine edilen değişkenler	2	1.99	34416674657	**57.40**	0.000	**57.40**	0.000<0.001
	3	3.34	5,70502E+12	12.85	0.000	12.54	0.000<0.001
	4	**19.70**	**2715327701**	13.00	0.000	17.61	0.000<0.001
	5	14.46	9846561289	20.53	0.000	20.24	0.000<0.001

SONUÇ ve ÖNERİLER

Sonuç olarak fıstık tiplerinin kümelenmesinde en önemli rol oynayan karakterlerin çiçeklenme zamanı ile ilgili değişkenler olduğu söylenebilir. Çünkü fıstık tipleri erkenci ve geççi olarak iki guruba ayrıldığını görmekteyiz. Orta geççi fıstık tipleri ise büyük bir ihtimalle erkenci tiplerle geççi tiplerin birer melezi olduklarından, fenolojik olarak daha çok benzerlik gösterdikleri fıstık tipleri ile kümelendikleri Şekil 2.'den görülebilmektedir.

Değişken sayısının kümelenecek birey sayısından fazla olması (n<p) durumunda, değişken sayısını azaltmak için en iyi yöntemin Temel Bileşenler olduğu tespit edilmiştir. Özellikle küme sayısı hakkında önceden herhangi bir ön bilgi olmaması durumunda temel bileşen skorları ile tatmin edici sonuçlar elde edilebilmektedir.

Küme yapısının önceden bilinmesi durumunda, diskriminant analizi ile bireylerin doğru kümelere atanıp atanmadıkları konusunda önemli bilgilerin elde edilmesi mümkündür. Ancak söz konusu veri setimize ait küme yapısının ve değişkenlerin etkinliği hakkında herhangi bir bilgi olmaması durumunda, Stepwise Diskriminant analizi ile etkili değişkenleri belirleyerek boyut indirgemek pek mümkün olamamaktadır.

Blashfield, (1976); Monte Carlo teknikleri ile karışık dağılıma sahip veriler türeterek, oluşturulan örneklere farklı kümeleme yöntemlerini uygulamıştır. Sözü edilen karışık dağılımlı modeli en iyi belirleyen yöntemin Ward yöntemi olduğunu belirlemiş olup, bu durum çalışmamızda ki sonuçları desteklemektedir.

Ancak en iyi kümeleme sonuçlarını veren yöntemin, her zaman Ward yöntemi olduğunu söylemek pek doğru değildir. Çünkü birimlerin kümelenmesinde, kümeleme yöntemlerinin gösterdikleri başarı veri yapısına bağlı olarak değişmektedir. Yani Verilerin dağılış şekli değiştikçe, farklı bir kümeleme yöntemi daha iyi sonuçlar verebilmektedir. Bununla ilgili bir kısım

çalışmalar yapılmış olup, Fisher and Van Ness, (1971) ve Mainly (1994)'in de açıklamaları bu doğrultudadır.

Bu çalışma sürecinde verileri tavsiye edilen herhangi bir kümeleme yöntemi ile değerlendirmenin çok zor olduğu tespit edilmiştir. Veriler hakkında geniş bir bilgi birikimi olmadan veya konunun uzmanı olmadan sadece bazı kümeleme yöntemlerine ve çok değişkenli istatistik kriterlerine bağlı kalarak sonuçları değerlendirip yorumlamanın doğru olmadığı kanısına varmış bulunmaktayız. Çünkü küme oluşumları veri setindeki değişkenlerin etkinliğine göre farklılık arz etmektedir. Bu durum verilerin çok boyutlu uzaydaki dağılımının ve değişken yapısı ile değişken seçiminin önemini ortaya koymaktadır.

TEŞEKKÜR

Doktora çalışmam sürecinde katkılarını ve yardımlarını esirgemeyen Prof. Dr. Yüksel BEK'e en içten teşekkürlerimi sunarım.

KAYNAKLAR

Acar, İ., 1997. Ceylanpınar Tarım İşletmesinde Seçilmiş Bazı Erkek Antep Fıstığı Tiplerinin Morfolojik ve Biyolojik Özellikleri Üzerinde Bir Araştırma. Harran Üniv. Fen Bil. Enst. Y. Lisans Tezi (Basılmamış), 92 s.

Anderberg, M.R., 1973. Cluster Analysis For Applications. Academic Press, New York 359 S.

Blashfield, R.K., 1976. Mixture Model Test of Cluster Analysis: Accuracy of Four Agglomerative Hierarchical Methods. Psychological Bull., 83: 377-388.

Bliss F.A.,1981. Utilization of Vegetable Germplasm. Hortscience 16(2): 129-132.

Fisher, L., Van Ness, J.W., 1971. Admissible Clustering Procedures. Biometrika, 58: (1), 91-104.

Hawkins, D. M., Müller, M.W., Krooden, J.A., 1982. Topics in Applied Multivariate Analysis. Cambridge University Press, USA.

Huang Tung C., 2001. The Effects of Linkage Between Business and Human Resource Managament Strategies, Personal Review, 30(2):132-151.

Jain, S.K., 1975. Population Structure and the Effects of Breeding System. in Crop Genetic Resources for Today and Tomorrow, Ed. Frankel, O.H. and Hawkes, J.G., Cambridge Uni. Press.

Johnson, A.R., Wichern, D.W., 1992. Applied Multivariate Statistical Analysis. Prentice Hall International Inc. New Jersey.

Mac Queen, J., 1967. Some Methods for Classifications and Stopping Rules: An Evaluation. Comp. J., 20: 359-363.

Mainly, B.F.J., 1994. Multivariate Statistical Methods, Second Edition, Londra: Chapman– Hall. New York,

Minitab Inc, 1995. Minitab Release, V: 10,51.

Özdamar, K.,1988. Hastalık Olgularının İncelenmesinde Kümeleme Çözümlemesinin Kullanılması. T.C. Anadolu Üniversitesi Yayınları No:295, Tıp Fakültesi Yayınları No:25, Eskişehir.

Özdamar, K., 1999. Paket Programlar ile İstatistiksel Veri Analizi. Kaan Kitabevi, Eskişehir, 214s.

Sharma S., 1996. Applied Multivariate Techniques. John Wiley & Sons.Inc. New York.

Sneath, P. 1957. The Application of Computers to Taxonomy. Journal of General Microbiology, 17, 201-226.

Sokal R.R., Michener C.D. 1958. A Statistical Method for Evaluating Systematic Relationships. The University of Kansas Scientific Bulletin 38: 1409-1438.

Tan, A., 1983.Sayısal Taksonomik Yöntemlerle Varyasyonun Saptanması. Ege Bölge Zirai Araş. Enst. Yay. No: 30, Menemen- İzmir.

Tatlıdil, H., 1992. Uygulamalı Çok Değişkenli İstatistik Analiz. Hacettepe Üniversitesi, Ankara.