T.C.
İNÖNÜ ÜNİVERSİTESİ
SOSYAL BİLİMLERİ ENSTİTÜSÜ

ÇOK DEĞİŞKENLİ GRAFİKLER ÜZERİNE BİR İNCELEME

YÜKSEK LİSANS TEZİ

DANIŞMAN HAZIRLAYAN

Prof. Dr. Mehmet GÜNGÖR **Sedat ALKAN**

MALATYA-2019

T.C.
İNÖNÜ ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI

YÜKSEK LİSANS TEZİ

ÇOK DEĞİŞKENLİ GRAFİKLER ÜZERİNE BİR İNCELEME

Sedat ALKAN

Danışman
Prof. Dr. Mehmet GÜNGÖR

MALATYA, 2019

T.C.

İNÖNÜ ÜNİVERSİTESİ

SOSYAL BİLİMLERİ ENSTİTÜSÜ

ÇOK DEĞİŞKENLİ GRAFİKLER ÜZERİNE BİR İNCELEME

YÜKSEK LİSANS TEZİ

DANIŞMAN	HAZIRLAYAN
Prof. Dr. Mehmet GÜNGÖR	**Sedat ALKAN**

Jürimiz .....tarihinde yapılan savunma sınavı sonucunda bu yüksek lisans tezini (oybirliği /oyçokluğu) ile başarılı bulunarak Ekonometri Anabilim dalında yüksek lisans tezi olarak kabul edilmiştir.

Jüri Üyelerinin Unvan Ad Soyadı	İmzası
1. Prof. Dr. Mehmet GÜNGÖR	
2. Doç. Dr. Yunus BULUT	
3. Dr. Öğr. Üyesi Fahrettin ÖZBEY	

İnönü Üniversitesi Sosyal Bilimler Enstitüsü Yönetim Kurulunun ..... tarih ve .....sayılı kararıyla bu tezin kabulü onaylanmıştır.

Prof. Dr. Mehmet KUBAT

Sosyal Bilimler Enstitüsü Müdürü

ONUR SÖZÜ

Prof. Dr. Mehmet Güngör'ün danışmanlığında yüksek lisans tezi olarak hazırladığım **"ÇOK DEĞİŞKENLİ GRAFİKLER ÜZERİNE BİR İNCELEME"** başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın tarafımdan yazıldığını ve yararlandığım bütün yapıtların hem metin içinde hem de kaynakçada yöntemine uygun biçimde gösterilenlerden oluştuğunu belirtir, bunu onurumla doğrularım.

Tarih:

Ad-Soyad: Sedat ALKAN

İmza:

TEŞEKKÜR

Tez çalışmamın her aşamasında ilgi ve önerilerini eksik etmeyen ve beni yönlendiren çok değerli danışman hocam Prof. Dr. Mehmet Güngör'e sonsuz teşekkürlerimi sunuyorum.

Tez çalışmamda eksik kaldığım noktalarda bana desteğini esirgemeyen Dr. Muhammed Hanifi Van'a teşekkürlerimi iletiyorum.

Lisans ve lisansüstü eğitim döneminde maddi ve manevi olarak desteğini hiçbir zaman eksik etmeyen abim Yusuf Alkan'a teşekkürlerimi borç bilirim ve hep minnettar olacağım.

Kaynaklar konusunda desteğini esirgemeyen lisans döneminden ve halen arkadaşım olan Mustafa Taşkın'a da sonsuz teşekkürlerimi borç bilirim.

Aynı zamanda eğitim hayatımda ve tez sürecinde her konuda desteklerini esirgemeyen aileme de sonsuz teşekkür ediyorum.

Sedat ALKAN

ÖZET

Bu çalışmada, günlük hayatta oldukça karmaşık gibi görünen çok değişkenli verilerin daha anlaşılır şekilde gösterimi için sıkça kullanılan grafiksel yöntemler hakkında detaylı bilgiler açık bir şekilde sunulmaya çalışılmıştır. Bunun yanı sıra çok değişkenli grafiksel yöntemlerde kullanılan tekniklerin çoklu veri kümesinde gizli kalmış yapıların ortaya çıkarılmasında son derece başarılı olduğu vurgulanmıştır.

Çalışmanın birinci bölümünde, konuya giriş ele alınacaktır. Gerçek yaşamda elde edilen veriler ile ilgili olarak kayıp değerlerin bulunması, verilerin aşırı değerler içermesi, veri içerisinde uyumsuzlukların bulunması gibi konulara değinilmiştir. Veri temizlemenin olduğu bu adımlarda; kayıp verilerin, aykırı değerlerin tespit edilmesi ve verilerdeki uyumsuzlukların ortadan kaldırılması gibi işlemler yapılmaktadır. İkinci bölümünde, çok değişkenli grafikler hakkında detaylı bilgiler verilmiştir. Son kısımda ise, çok değişkenli grafikler ile ilgili uygulamalara yer verilmiştir.

Anahtar Kelimeler: Çok Değişkenli Grafikler, Aşırı Değerler, Kayıp Değerler.

ABSTRACT

In this study, detailed information about graphical methods commonly used for more comprehensible representation of multivariate data which seems to be quite complex in daily life is tried to be presented clearly. In addition, it is emphasized that the techniques used in multivariate graphical methods are extremely successful in revealing hidden structures in multiple data sets.

In the first part of the study, introduction to the subject will be discussed. Issues such as finding missing values, excessive values of data, and discrepancies in the data were discussed. In these steps where data cleaning is performed; loss data, detecting outliers and eliminating any inconsistencies in the data. In the second part, detailed information about multivariate graphs is given. In the last part, applications related to multivariate graphs are given.

Key Words: Multivariate Graphs, Extreme Values, Lost Values.

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

GRAFİKLER

BÖLÜM 1

GİRİŞ

Günlük hayatta, çoğu alanlarda toplanan ya da elde edilen verilerin sonuçlarına ulaşmak için çeşitli analizler yaparız. Bu analizler yapılırken kimi zaman sonuçları etkileyen bir takım sıkıntılar ortaya çıkabilir. Araştırmacı analiz yaparken ya da analize geçmeden önce bazı hususlara dikkat etmelidir. Veri analizi yapılmadan önce yapmamız gereken çalışmanın amacını açık bir şekilde tanımlamaktır. Çalışma amacı, sorun üzerine odaklanmış ve açık bir dille ifade edilmiş olmalı, elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği tanımlanmalıdır. Sorun ile tam örtüşmeyen bir veri toplama çalışması, sorun çözmeye yetmeyeceği gibi sonuç kısmında başka problemlerin de ortaya çıkmasına neden olabilecektir. Amaç belirlendikten sonra, veri hazırlama işlemine geçilir. Amaca uygun veriler seçilir ve bu veriler analize hazır hale getirilir. Veri analizinde güvenirliliğin arttırılması için veri ön işleme adımı muhakkak yapılmalıdır. Aksi halde hatalı girdi verileri, bizi hatalı sonuçlara götürebilir.

Veri ön işleme; veriler üzerinde herhangi bir analiz türünün uygulanmasını engelleyecek veri problemlerinin çözümü için, verilerin doğasının anlaşılması, anlamlı veri analizinin başarılması için ve verilen bir veri kümesinden daha anlamlı bilginin çıkarılması için yapılmaktadır. Bu adımda uygulanacak her işlem, en son adımda verilecek olan kararı etkileyecektir. Analiz için toplanan veriler hatalı ya da aşırı değerler içerebilir. Veriler içerisinde uyumsuzluklar, hatta eksiklikler olabilir. Bu yüzden veri ön işleme aşamasında verideki kusurlar, eksiklikler ve hatalar giderilerek veriler analiz aşamasına hazırlanmış olur.

Modelin kurulması noktasında ortaya çıkacak problemler, bu adıma tekrar tekrar geri dönülmesine ve verilerin yeniden düzenlenmesine sebep olacaktır. Bu problem verilerin hazırlanması ve modelin kurulması adımları için sürecin gecikmesine sebep olacaktır. Veri ön işleme sayesinde elde edilen sonuçların kalitesi ve harcanacak zaman arttırılmış olur.

Özellikle değinmemiz gereken husus verilerin temizlenmesi yöntemidir. Gerçek yaşamda elde edilen veriler içerisinde mutlaka bazı problemler bulunur. En çok karşılaşılan sorunlar veri içerisinde kayıp değerlerin bulunması, verilerin aşırı değerler içermesi, veri içersinde uyumsuzlukların bulunmasıdır. Bu yüzden veri temizlemenin olduğu bu adımlarda; kayıp verilerin, aykırı değerlerin tespit edilmesi ve verilerdeki uyumsuzlukların ortadan kaldırılması gibi işlemler yapılmaktadır.

1.1. KAYIP DEĞERLER

Kayıp veriler, herhangi bir analiz yaparken karşımıza çıkabilecek muhtemel durumlardan bir tanesidir. Kayıp değerler çok sayıda sebepten dolayı karşımıza çıkabilir. Kayıp değerler, veri giriş hatalarından, eksik bilgi toplanmasından, bilginin alındığı birimin cevap vermekten kaçınmasından ya da birimin o bilgiye sahip olmamasından kaynaklanabilir. Analiz için cevap verenlerin dışında ya da cevaplayıcı kaynaklı olan ve kayıp veriye yol açan sürece kayıp veri süreci denilmektedir. Cevaplayıcı kaynaklı oluşan kayıp veri sürecinin önceden görülmesi imkansız olabilir. Böyle bir durumda, araştırmacının yapması gereken kayıp veri sürecini ortaya çıkartan bir durumun var olup olmadığının araştırılmasıdır. Araştırmayı yaparken araştırmacının iki önemli noktayı dikkate alması gerekir. İlki kayıp verilerin gözlemlere rasgele mi dağıldığı yoksa belirgin bir yapı mı oluşturduğudur. İkinicisi ise, kayıp verilerin ne kadar sıklıkla karşımıza çıktığıdır. Bunların araştırılması gerekir.

Kimi araştırmacılar kayıp veriye sebep olan değerlerleri veri grubundan çıkartmaktadırlar. Bu durumda kimi zaman gözlem sayısında büyük bir ölçüde eksilme olmaktadır. Bundan dolayı örneklem büyüklüğü olumsuz olarak etkilenmektedir. Hatta araştırmanın güvenilirliğini ve sonuçlarını ciddi düzeyde etkilemektedir. O zaman kayıp veri oluştuğunda veriye yeni gözlem değerleri eklenebilir, verideki eksik değerler çeşitli istatistiksel yaklaşımlarla giderilmeye çalışılır. Burada önemli olan kayıp verinin nasıl oluştuğu değil, kayıp veri sorununun uygun bir çözüm yönteminin bulunmasıdır. Uygun çözüm yönteminin bulunması için kayıp veri kavramının doğru bir şekilde bilinmesi gerekir ve kayıp veri sürecinin ayrıntılı olarak incelenmesi gerekir.

Kayıp verilerin araştırılmasıyla, kayıp verilerin hangi değişkenlerde olduğu ve ne kadar sayıda olduğu, bir değişken için kayıp olan verinin diğer değişkenleri ne kadar

etkilediği, eksik olan değişken silindiğinde gözlem sayısının ne kadar etkilendiği görülmektedir. Genellikle çok değişkenli araştırmalarda tam veriye ulaşmak bazen mümkün olmayabilir. Bu araştırmalar yapılırken öncelikle kayıp verinin derecesini belirlemek çok önemlidir.

1.1.1. Kayıp Veri Mekanizmaları

Kayıp veri sorununun giderilmesinde en önemli aşama, kayıp veri mekanizmalarının açıklanmasıdır. Kayıp veri sorununu halletmek için doğru atama yöntemi seçilmeden önce, bir kayıp veri mekanizması belirlenmelidir. Bunun sebebi hangi çözüm ve atama yönteminin doğru ve uygun olabileceği kayıp veri mekanizmaları sayesinde ortaya çıkmaktadır.

Kayıp verilerin analiz edilmesinde mekanizmanın rolü Rubin'in 1976'daki teorisinde formülize edildikten sonra araştırmacılar tarafından kullanımaya başlanmıştır. Rubin'in teorisinde kayıp veriler rassal değişkenler olarak nüfuz edilmiş ve bir dağılıma atanmıştır. Little ve Rubin oluşturduğu mekanizmalar 3 tanedir. Bunlar:

1.1.1.1. Tamamıyla rassal olarak kayıp mekanizması (MCAR)

Tamamen rasgele olarak kayıp veri mekanizması, rasgeleliğin en yüksek olduğu durumdur. Verilerin kayıp olma durumu gözlenen veya kayıp olan diğer verilerden bağımsızdır. Yani, X ve Y gibi iki değişken ele alındığında, Y değişkenine cevap alınamama olasılığı X değişkeni ile ilişkili değil ise, Y'deki eksik veri Y değişkeninin kendi yapısından kaynaklanıyorsa bu tür eksik veri mekanizmaları tamamıyla rassal olarak kayıp mekanizması (MCAR) olarak tanımlanır (Little ve Rubin, 1987).

1.1.1.2. Rassal olarak kayıp mekanizması (MAR)

Rasgele olarak kayıp veri mekanizması, MCAR'ın sınırlandırılmış farklı bir türüdür. Kayıp veri mekanizması kayıp verilere değil gözlenen verilere bağlıdır. Kayıp veri olasılığı araştırma konusuna bağlı ise, bu kayıp veri türü rassal olarak kayıp (MAR) olarak adlandırılır. Tek taraflı bir bağımsızlık söz konusudur. Yani, X ve Y gibi iki değişken alındığında, X değişkenindeki cevapsızlık olasılığı Y değişkenine bağlı iken, Y değişkenindeki cevapsızlık olasılığı X değişkenine bağlı değildir.

1.1.1.3. Rassal olarak kayıp değil mekanizması (NMAR)

Bu mekanizma kayıp verilere bağlıdır. Kayıp veri oluşumu araştırmacının

birimleri ya da olayları ölçmemesine dayanır. Bu durum veri analizine oldukça zarar verebilir. X ve Y gibi iki değişken ele alındığında cevap olasılığını her iki değişkene de bağlı olması mümkün ise eksik veri rassal olarak kayıp değil (NMAR) kabul edilir, diğer bir ifade ile eksik olma rassal değildir ve eksik veri diğer değişkenleri kullanarak tahmin edilemez.

1.1.2. Kayıp Veri Sorununun Giderilmesine Yönelik Yöntemler

Kayıp verinin rassal olup olmadığını öğrendikten sonra yapacağımız diğer adım, kayıp veri sorununun giderilmesine yönelik yaklaşımlarda bulunmaktır. Kayıp veri sorununun giderilmesine yönelik çözümlemeler, kayıp veri sürecinin rassallığı ve eksik veri tahmininde kullanılan yöntemler, genel anlamda silme ve atama yöntemleri olarak iki alt başlıkta incelenir.

1.1.2.1. Silme Yöntemleri

1.1.2.1.1. Liste Bazında Silme Yöntemi

Liste bazlı silme yönteminde sadece kayıp veri bulundurmayan gözlemler kullanılır. Bir veya birden fazla kayıp verisi olan gözlemler veri setinden silinerek veri tam veri durumuna dönüştürülür. Eğer cevap değişkeni kayıp veri içeriyorsa kullanılacak en mantıklı yöntem liste düzeyinde veri silmedir. Eğer veri seti tamamen rasgele olarak kayıp veri barındırıyorsa, bu yöntem uygulandıktan sonra istatistiksel güç daha düşük olacaktır. Eğer veri setinde rasgele olmayan kayıp veriler varsa, bu yöntem uygulandıktan sonra elde edilen veri seti taraflı sonuçların oluşmasına sebep olacaktır. Bu yöntemin dezavantajı, eksik gözlem içeren değişkenlerin çıkartılmasının sebep olduğu bilgi kaybıdır. Kayıp gözlemlerin bir değişken ya da değişken grubunda yoğunlaşmadığı ve birim bazında silme işleminin gerçekleştiği uygulamalarda örneklem büyüklüğü küçülecektir. Bundan dolayı, ilk etapta örneklem büyüklüğü yeterliyken silme sonucunda yetersiz olacaktır.

1.1.2.1.2. Çiftler Bazında Silme Yöntemi

Bu yöntem eldeki olası tüm verinin kullanılması yöntemi olarak da bilinmektedir. Bu yöntemde ilgili hesaplamalar veri seti içerisindeki kayıp gözlemleri dikkate almaz ve verinin mümkün olanı ile gerçekleştirilir. Çiftler bazında silme yönteminde, veri setinden gözlem biriminin veya değişkenlerin silinmesi söz konusu değildir. Bu yöntemde, sadece hesaplamalarda kayıp veriler göz önünde bulundurulmamaktadır.

Liste bazında silme yönteminde, bir birimin gözlemlenen değerleri bir değişkendeki gözlemlenemeyen değerinden dolayı çıkartıldığından araştırmacılar eldeki mümkün tüm verinin kullanılması yöntemine başvururken, bu yöntemde kayıp gözlemler atanmış değerlerle doldurulmamakta, eldeki mümkün tüm değerler yardımıyla dağılımın ortalama, standart sapma gibi tanımlayıcı istatistikleri, korelasyon ve kovaryans gibi ilişki ölçüleri hesaplanmaktadır. Bu yöntemde aşama aşama her bir değişkenin ortalama ve varyansları ayrı ayrı hesaplanır, ayrıca çift olarak ele alındıklarında mümkün olan tüm çiftler için kovaryans ve korelasyonlar hesaplanır.

1.1.2.2. Atama Yöntemleri

1.1.2.2.1. Ortalama Değerin Atanması

En sık kullanılan yöntemlerden biridir. Kayıp verinin ait olduğu değişkenin kayıp olmayan verilerinin ortalaması alınır. Ortalama alındıktan sonra elde edilen sonuç kayıp veriye atanır. Bu durum sonucunda, veri setinin ortalaması sabit bir şekilde dururken, varyansı küçülür. Varyansa gerekli önemi vermemesi, veri setinin korelasyon yapısına negatif yönde taraflılık kazandırması bu atama yönteminin tercih edilmesini olumsuz yönde etkilemektedir. Hesaplama kolaylığı bakımından sıklıkla tercih edilen bu yöntemin dezavantajı ise, parametre tahminlerindeki bozulmalardır. Bu yöntemde elde edilen ortalama, tam gözlemlerin ortalamasına eşit olmasına karşın standart sapma değeri küçüleceğinden değişkenler arasındaki korelasyon katsayıları da daha düşük elde edilecektir.

1.1.2.2.2. En Benzer Birime ya da Birimlere Benzetme

Benzer veriler kullanılarak yapılan bir atama türüdür. Veri setindeki tüm gözlemler benzer özelliklere göre gruplara ayrılırlar. Kayıp veri ataması yapılacak olan verinin yer aldığı gruptan rasgele bir gözlem seçilir. Rasgele seçilen bu değer kayıp veriye atanır. Çok sayıda alt gruplar oluşturmak, atama sonucu yapılan tahminlerin geçerliliğini arttırır, ama bu durumda alt grup örneklem hacimlerinin küçük olmamasına dikkat edilmelidir.

Bu atama yönteminde kayıp gözlemin bulunduğu birime, en benzer birimler tespit edilerek bu birimin o değişken için almış olduğu değer, kayıp gözlem tahmini olarak kabul edilir. Bu yüzden bu yöntemin kullanılabilmesi için kayıp gözlem içeren birime benzerliği hesaplanacak birimin, kayıp gözlemin bulunduğu değişken için değerinin

gözlemlenmiş olması gerekir. Eksik gözlem yerine, eksik gözlemin bulunduğu birime en benzer birimin gözlem değerinin kullanılması, aynı verinin birden çok kullanılması anlamına gelecek ve veride ortalama atama yöntemine benzer avantaj ve dezavantajları ortaya çıkaracaktır.

1.1.2.2.3. Dış Kaynaklı Atama

Dış kaynaktan atama yönteminde, genellikle önceden yapılmış benzer çalışmalardan elde edilen sabit bir değerin kayıp veriye atanması durumu söz konusudur. Önceki yöntemlerden tek farkı, atanacak değerin kaynağının farklı bir veri seti olmasıdır. Ortalama atama yöntemi ile benzer olumsuz özelliklere sahiptir. Araştırmacı dış kaynaklardan elde edilen değeri, eldeki veriler yardımıyla elde edilecek ortalama gibi bir değerden daha geçerli olabileceğinden emin olmalıdır.

1.1.2.2.4. Kayıp Gözlem ile Tam Gözlemin Yer Değiştirmesi

Örnekleme giren fakat genellikle tüm değerleri kayıp olan gözlem ile bu gözleme benzeyen ancak örnekleme hiç girmeyen başka bir gözlemin yer değiştirmesidir.

1.1.2.2.5. Regresyon Atama

Kayıp verilere tam veriler üzerinden elde edilen bir regresyon modeli ile değer atama yöntemidir. Bu durumda kayıp verinin bulunduğu değişken bağımlı değişkeni, tam verilerin olduğu değişkenler ise bağımsız değişkenleri oluşturmaktadır. Kayıp verinin türüne göre atama için kullanılacak regresyon modelleri; ikili veri tipleri için probit ya da logit modeller, kesikli tamsayı değerler için Poisson Regresyon ve diğer sürekli tip veriler için Klasik En Küçük Kareler Regresyon Modeli tercih edilebilir. Atama sonucunda veri setinde zaten var olan ilişki daha da kuvvetlenmiş olur. Bu atama yönteminin aynı veri seti üstünde kullanımı birden fazla sayıda olursa, elde edilen yeni veri seti kendine özgü bir yapıya ulaşır ve genellenebilirliği azalır.

1.1.2.2.6. EM (Expactation-Maximization) Yöntemi

Kayıp değerler ile veri analizi için gerekli olan modeldeki bilinmeyen parametreler arasındaki ilişkiye bağlı çalışan en çok benzerlik yöntemidir. Bu yöntem genellikle çok değişkenli normal model varsayımını kullanmaktadır. İki aşamadan oluşmaktadır. E adımında tam gözlemleri kullanarak eksik gözlemleri tahmin etmek için oluşturulan ortalama vektörü ve kovaryans matrisi ile stokastik regresyon modeli

oluşturulur ve model yardımı ile eksik gözlemler tahmin edilir. İkinci aşamada ise, atama gerçekleştirildikten sonra ortalama, standart sapma, korelasyon gibi değişkenlere ilişkin tahminlerde bulunulur. Bu iki aşama; elde edilen ardışık tahminler arasındaki fark önemli derece azalıncaya kadar tekrarlanır.

1.1.2.2.7. Çoklu Atama Yöntemi

Çoklu atama yönteminin temelini, kayıp veriye D ($D > 2$) defa atama yapma oluşturur. Kayıp verinin ait olduğu değişkene ait belirli bir modele bağlı dağılım belirlenir ve bu dağılımdan rasgele seçilen değerler ile atama yapılır. D tane tanımlanmış veri setinden elde edilen sonuçların birleştirilmesi sonucunda bir tahmin elde edilir. Bu yöntemin temeli eksik gözlemlerin iki veya ikiden fazla atama yöntemini birlikte kullanarak tahmin edilmeye çalışılmasıdır. Bu yüzden bu yaklaşım karma bir tahmin değeri bulmayı amaçlar. Bu tahmin değeri genellikle iki veya ikiden fazla yöntemle bulunmuş tahmin değerlerinin ortalamasıdır. Çoklu atama yöntemi, çoğu araştırmacı tarafından tek bir yöntemle elde edilen tahmin yöntemlerine göre daha güvenilir görülmektedir. Çoklu atama yönteminin en önemli avantajlarından bir tanesi anlaşılır olmasıdır. Aynı zamanda analiz değeri alan değişkenlerin normalliğinin ihmal edildiği durumlarda da kuvvetli sonuçlar vermektedir. Liste bazında veri silme, çiftler bazında veri silme ve yerine ortalamayı koyma yöntemlerinden pek çok durumda üstün olmaktadır. Dezavantajı ise, atama işleminin uzun sürmesidir.

1.1.2.2.8. K En Yakın Komşu (KNN) ile Kayıp Değer Atama Yöntemi

K en yakın komşu (KNN) ile kayıp değer atama yönteminin temelini KNN algoritması oluşturmaktadır. Bu algoritma, bir örneklemde yer alan gözlemlerin her birinin belirli bir gözlem değerine göre uzaklıklarının hesaplanması ve elde edilen en küçük k tane gözlemin seçilmesi ile elde edilir. KNN algoritmasında uzaklık hesaplamalarında kullanılmak üzere birbirinden farklı fonksiyonlar belirlenmiştir. Bunlara örnek olarak Öklid (Euclidean), Manhattan ve Minkowski Uzaklık Fonksiyonları gösterilebilir. Bu uzaklık fonksiyonları içerisinde kullanımı en yaygın olan Öklid Uzaklık Fonksiyonudur. Bu fonksiyon ile p boyutlu bir uzayda i ve j noktaları arasındaki uzaklık şu şekilde elde edilir:

$$d(i,j)=\sqrt{\sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^2} \qquad (1.1)$$

Veri setinin ikiden fazla sayıda değişken içerdiği durumlarda Standardize Edilmiş Öklid Uzaklık Fonksiyonu kullanılır. Her değişken kendi içerisinde z dönüşümü uygulanarak standardize edilir ve Eşitlik 1.l'e yerleştirilir. Böylece değişkenler arasındaki ölçüm farklılıktan ortadan kalkmış olur.

1.1.3. Kayıp Verilerin Tamamlanması (SPSS)

Kayıp verilerle karşılaştığımız durumlarda bu verileri veri grubundan çıkartmadan nasıl analize dahil edebileceğimizi bir de SPPS programı ile görelim. SPSS'te Transform-»Replace Missing Values menüsüne girilir.

Method kısmından istediğimiz herhangi biri işaretlenir. Burada:

a) Series Mean: Serinin ortalamasını alarak, o değişkendeki eksik gözlemlere ilgili değişkenin var olan değerlerinin aritmetik ortalaması atanır.

b) Mean of Nearby Points: Ortalama hesaplamasında kullanılan eksik değerlerin altındaki ve üstündeki tam gözlem değerlerinin aritmetik ortalaması alınarak, eksik gözlem yerine bu değer konur. Ancak yakın noktaların mesafesi (span of nearby points) seçeneğiyle programda var olan ‘‘2’’ değeri seçilerek eksik veriye iki altındaki ve iki üstündeki gözlemlerin ortalaması atanabilir.

c) Median of Nearby Points: Medyan hesaplamasında kullanılan eksik değerlerin altındaki ve üstündeki tam gözlem değerlerinden yararlanılarak bir medyan değeri hesaplanır ve eksik gözlem yerine bu değer yazılır. Ancak yakın noktaların mesafesi (span of nearby points) seçeneğiyle programda var olan ‘‘2’’ değeri seçilerek eksik veriye iki altındaki ve iki üstündeki gözlemlerin ortancası atanabilir.

d) Linear Interpolation: Eksik değerden önceki son tam gözlem değeri ve eksik değerden sonraki ilk tam gözlem değeri eksik olan yerlere yerleştirmede kullanılır. Eğer serideki ilk gözlem ve son gözlem eksik değer ise kayıp değer yerleştirilmez.

Sonuç olarak, veri setini kayıp değerlerden arındırmak için yapılabilecek bazı işlemler şunlardır (Bilen, 2004; Oğuzlar, 2004):

1. Eğer maliyetli olmayacaksa ve bu durum bize zaman kaybettirmeyecekse, kayıp değerin ait olduğu birime başvurmak.

2. Kayıp değerin bulunduğu değer ya da değişkenler veri setinden çıkartılabilir. Bu yöntem veri kaybına sebep olduğu için genellikle tercih edilmez.

3. Eğer kayıp değerin yeri doldurulabiliyorsa, kayıp değerin tamamlanması. Örneğin, görüştüğümüz birey Türk vatandaşı ise, bu kısımda boş ise, uyruğu alanına T.C. yazılarak tamamlanması.

4. Değişkenin genel eğiliminin gösterdiği değeri, kayıp değerin yerine atamak. Örneğin, değişkenin ortalaması, mod ya da medyanının kayıp değer yerine atanması.

5. Var olan verilere dayalı olarak kayıp değerin tahmin edilmesi.

6. Kayıp değer yerine "Bilinmeyen" gibi global bir sabitin atanması.

1.2. AYKIRI DEĞERLER

Gerçek hayatta elde edilen verilerin, bilimsel yöntemler kullanılarak değerlendirilebilmesi için bazı ön işlemlerden geçmesi gerekebilir. Veri ön işleme aşamalarından olan veri temizleme ve verinin eksik olması veya tutarsız olması gibi problemlere ek olarak, aşırı veya uç değerler içermesi özellikle istatistiksel analizlerin yapılması için dikkat edilmesi gereken bir başka problemdir. Veri setinin normal dağılması gerektiği durumda eğer veri seti normal dağılmıyorsa buna sebep olan nedenlerden bir tanesi de aykırı değerlerin olmasıdır. Aykırı değerler, veri setinin ortalamasının çok uzağına düşen değerler olarak ifade edilir. En genel tanımıyla aykırı değer "evren ya da örneklem normlarının çok dışarısında bulunan veri noktaları" olarak tanımlanmaktadır.

Aykırı değerler, bir tane olabileceği gibi birden fazla da olabilir. Bu değerler, verilerin standart sapmasını arttırıken, dağılımın şeklini de değiştirebilir. Dolayısıyla istatistiksel olarak, karar süreci sonrasında yanlış kararlar verilmesine sebep olabilirler. Veri setindeki diğer değerlerle mukayese edildiğinde, veri setine uygun olmadığı anlaşılan aşırı değerlere aykırı değer denir. Aşırı değerlerin oluşmasına birçok durum sebep olabilir; hatalı veri girişi, yanlış kodlama, her zaman ortaya çıkmayan bir olayın görülmesi, vb. sebeplerden dolayı ortaya çıkabilir. Hatalı veri girişi ya da yanlış

kodlama, verinin temizlenmesi aşamasında düzeltilmelidir. Bu aykırı değerler hatalı veri girişi sebebiyle olabilirken, eğer elde edilecek veri tartım aleti ile yapılacak ise, ölçüm aletinin hatalı olmasından veya tamamen deneme materyalindeki farklılıktan oluşmuş olabilir. Aşırı değerlerin çok fazla olması veri setinin normal dağılmayacağını ve yapacağımız istatistiksel analizlerin etkilenmesine neden olabilir. Bundan dolayı veri setinin normal dağılması sonuçların daha güvenilir olmasını sağlayacaktır.

Gözlem değerleri çok fazla olan bir veri setinin istatistiksel olarak analizi yapılırken aykırı değerler üzerinde çok fazla düşünülmeden analizden saf dışı edilebilir. Tam tersi düşünülürse gözlem değerleri küçük olduğunda tek bir gözlemin bile analiz sonuçlarına etkisi çok kıymetli olacaktır. Bundan dolayı aykırı değerlerin doğru tespit edilmesi ve giderilmesi küçük gözlem değerleri için büyük değer taşımaktadır. Bunun yanı sıra, aykırı değer tespiti mikrodizin verileri ve klinik biyokimya verileri gibi büyük veri setlerinin kalite kontrolü ve ilaç endüstrisi için ayrı bir öneme sahiptir.

Aykırı değer tespit yöntemleri tanımlayıcı istatistik ve teste dayalı yöntemler olmak üzere ikiye ayrılmaktadır. Standart sapma yöntemi ve box-plot grafik yöntemi tanımlayıcı istatistiklerle aykırı değer tespit yöntemlerine girerken, test yöntemleri verinin dağılım şekline göre parametrik ve parametrik olmayan test yöntemleri olarak iki şekilde incelenmektedir. Eğer veri normal dağılıma uyumlu ise, aykırı olduğundan şüphelenilen değer, Dixon, Grubbs t, Rosner, Discordance ile değerlendirilir. Normal dağılıma uyumlu olmayan veri setinde ise Walsh testi ile aykırı değer tespiti yapılmaktadır.

a) Standart Sapma Yöntemi

Anakütledeki değişkenler normal dağılım gösteriyorsa, analizi yapılan konunun hassasiyetine göre ± 2 ya da ± 3 standart sapmanın altında ve üstünde kalan değerler aykırı değer olarak belirlenir. Bu yöntemin kullanılabilmesi için gözlem değerlerinin 120 ve üzeri olmalısı gerekir.

b) Box-Plot Yöntemi

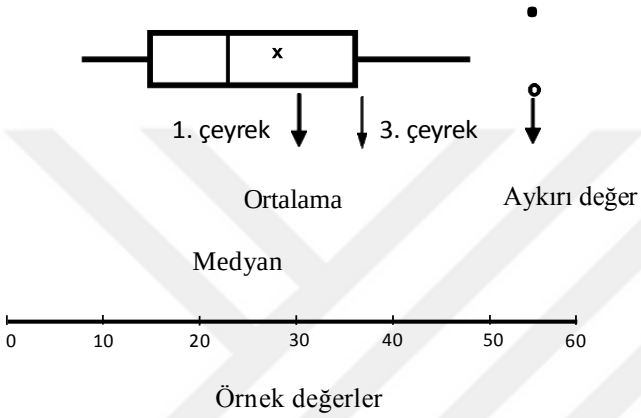
Bu yöntemde değerlerin % 25'inin başladığı sınır 1. kartil ve % 75'inin başladığı sınır 3. kartil olmak üzere, kartiller arası genişlik değeri ise, IQR olarak hesaplanır. Bu değer kullanılarak alt ve üst sınırlar belirlenir.

IQR = 3. kartil - 1. kartil

Üstsınır = 3. kartil + 3/2 IQR

Altsınır = 1. Kartil - 3/2 IQR

Bu sınırların dışındaki değerler aykırı değer olarak belirlenir. Box-plot grafiği ile tespit edilen aykırı değerler Şekil 1.1'deki gibi görünür.



Şekil 1.1. Aykırı Değerin Box-Plot Grafiğiyle Tespit Edilmesi

Teste dayalı yöntemler için hipotez;

H_0:Veri setinde aykırı değer yoktur

H_1:Veri setinde en az bir aykırı değer vardır şeklinde oluşturulur.

c) Dixon Testi

Bu test bir veri setinde ortalamadan uzakta yer alan bir gözlemin aykırı değer olup olmadığını bulmak için kullanılan bir testtir. Örnek büyüklüğü, $3 \leq n \leq 25$ arasında olduğu ve bir tek aykırı gözlem olduğunun düşünüldüğü durumlarda kullanılır. Ayrıca bu test, aykırı olarak düşünülen gözlemin dışındaki gözlem değerlerinin normal dağıldığını varsayar. Veriler artan sıraya göre dizildiğinde, aykırı gözlem olarak düşünülen gözlem veri setindeki gözlemlerin en küçüğü ya da en büyüğü olacaktır. Bu test parametrik bir testtir.

d) Rosner Testi

Rosner testi, bir veri setinde 25 ve üzeri gözlem sayısı olduğunda ve veriler normal bir dağılım gösterdiğinde, ayrıca $r_0 \leq 10$ 'a kadar aykırı değer hesaplayabilen bir

testtir. Bu testte hem küçük hem de büyük aykırı değerler tanımlanabilir. Bu yüzden, test her zaman çift yönlüdür. Rosner testinin aykırı değer sayısı arttıkça hesaplama işlemi zorlaşır.

Rosner testinde, en uzak gözlemden başlayarak i=0,1,2,…,10 sıralama değerleri verilir. 10, ortalamaya en yakın aykırı değerin sıralamasını, 0 ise, ortalamaya en uzak aykırı değerin sırasını ifade eder. Rosner tesinde kaç aykırı gözlemden şüpheleniliyorsa, o sıra belirlenip (i), en yakın değer red edilenceye kadar işlem 0'a kadar yürütülür. En yakın değerin ilk kez red edildiği durumda o ve o değerin altındaki gözlemlerin aykırı değerler olduğuna karar verilir (West, 1999a). Parametrik bir testtir.

e) Discordance Testi

Bu test, küçükten büyüğe doğru sıralanmış bir veri setinin, en solunda ya da en sağında bulunan bir tek aykırı değeri tespit edebilen bir yöntemdir. Örnek genişliğinin 3 ile 50 arasında olması gereklidir. Bu testin uygulanabilmesi için gözlem değerlerin normal dağılıma sahip olması gerekir. Parametrik bir testtir.

f) Grubbs Testi

Frank Grubbs tarafından aykırı değerlerin tespiti için geliştirilen bu test verilerin normal dağılımdan geldiğini varsayar. Bundan dolayı, bu testin uygulanabilmesi için aykırı olarak düşünülen gözlemin dışındaki değerlerin normal dağılımlı olması gerekir. Eğer veriler normal dağılım gösteriyorsa, bu veriler artan sıraya göre dizilirler, ortalama ve standart sapma değerleri hesaplanır. Normal dağılıma sahip olan, 3 ve 100 arasında gözlem değeri içeren veri setlerinde, aykırı değer tespiti için kullanılan bir testtir. Tek seferde en fazla iki değeri test edebilir. İkiden fazla değer olması durumunda ise testin tekrarlanması gerekir. Test işleminden önce veri seti küçükten büyüğe dizilir. Aykırı değer olup olmadığı test edilen her bir değer için bir T değeri hesaplanır. Hesaplanan değerler, tablo kritik değerini aşıyorsa veri aykırı değer olarak kabul edilir. Parametrik bir testtir.

g) Walsh Testi

Bu test, veri setinde çok sayıda olan aykırı değerleri test etmek için kullanılan parametrik olmayan bir testtir. Bu testteki veriler için herhangi bir normallik varsayımına gerek yoktur. Veri setindeki gözlem değerleri küçükten büyüğe doğru

sıralanır. Bu test örnek büyüklüğünün 60'tan küçük olduğu durumlarda uygulanmaz. Eğer örnek büyüklüğü $60 < n \leq 220$ aralığında ise $\alpha = 0.10$, eğer örnek büyüklüğü $n > 220$ den fazla ise $\alpha = 0.05$ olarak kabul edilir.

1.2.1. Aykırı Değerlerin Etkileri

Aykırı değerler yani, uyumsuz olan bu veriler, istatistiksel analiz çıktılarında aşırı bir etkiye sebep olabilmektedir. Aykırı değerler belirlenmeden yapılan veri analizinde, aykırı değer içeren analiz, ortalamanın sola ya da sağa çarpılmasına, korelasyonların daha düşük veya yüksek olmasına ve regresyon katsayılarının yanlı hale gelmesine sebep olabilir. Bundan dolayı, bazı etkili gözlemler tanımlanmalı ve veri analizindeki rollerine karar verilmelidir. Aykırı değerlerin, hata varyansını artırıcı etkileri vardır ve istatistiksel testin gücünü azaltırlar.

İkinci olarak uç değerler, eğer puanlar rasgele dağılmışsa, normalliği etkilerler. Üçüncü olarak, yanlılığa neden olabilirler.

1.2.2. Çok Değişkenli Aykırı Değer Belirleme Yöntemleri

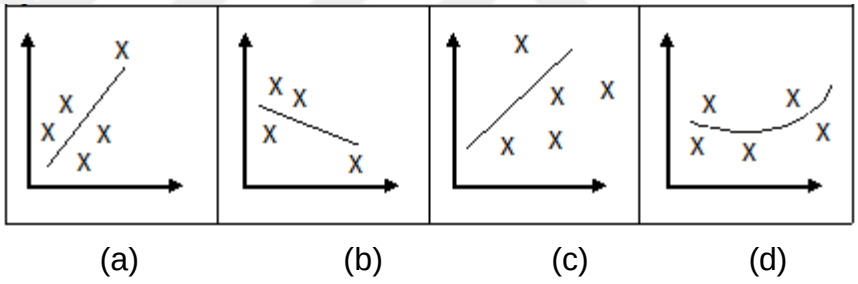
1.2.2.1. Mahalanobis Uzaklığı

Bu yöntem, tek değişkenli uzaklık yöntemiyle yakından ilişkili olsa da, çoklu normal verilerde ve büyük örneklemlerde, değişken sayısına bağlı olarak elde edilen serbestlik derecesi ile ki-kare (χ^2) dağılımına uyar ve örneklem sayısı arttıkça da bu kestirim daha iyi hale gelir (Johnson ve Wichern, 2002).

Mahalanobis uzaklığı değerinin, bağımsız değişken sayısını serbestlik derecesi alan (χ^2) tablo değeriyle mukayese edilmesi gerekir. Hesaplanan Mahalanobis uzaklığı değeri, bulunan tablo değerinden küçük ise, çok değişkenli normallik varsayımının karşılandığı söylenebilir. (χ^2) tablo değerinin üzerinde bir Mahalanobis değerine sahip olan gözlemler, aşırı değerler olarak belirlenmektedir ve bu değerler verilerden çıkartılabilmektedir. Bu işlemler için 0.01 ya da daha tutucu olmak isteniyorsa 0.001 anlamlılık düzeyinin dikkate alınması tavsiye edilmektedir (Büyüköztürk, 2003).

1.2.2.2. Cook Uzaklığı

1977'de Cook tarafından sunulan Cook Uzaklığı asıl olarak, belirli bir modele göre bir gözlemin etkisini göstermek üzere geliştirilmiş yöntemdir. Cook Uzaklığı aynı zamanda kutu grafiği(Boxplot) yönteminde olduğu gibi, bir tek gözlemin aşırı değer olup olmadığı ile ilgili bir yöntemdir. Bu istatistik tüm parametre vektörü içindeki birinci gözlemin etkisini özetler (Akt. Yalaz ve Kaya, 2009). Cook Uzaklığı, bir gözlemin tahmini regresyon katsayısına olan etkisini belirlemek için de kullanılır. Cook Uzaklığı ile uç değer olarak bulunan gözlemin hemen silinmesi yerine, buna neden olan durumlar belirlenerek modellerin regresyon eşitlikleri ve katsayıları belirlenmeli ve buna göre önlem alınmalıdır (Akt. Walfish, 2006). Her bir gözlemin tek tek parametre kestirimleri üzerinde önemli etkisi olabilmektedir. Bu tür bir gözlemin modelden çıkartılması sonuçları değiştirebilir. Bunlar model biçimleri üzerinde de etkili olabilirler.



Şekil 1.2. Tek bir gözlemin regresyon eğrisi üzerindeki etkisi

Şekil 1.2'deki grafikler incelendiğinde, tek bir noktanın bile (grafikler üzerinde işaretlenmiş gözlemler) doğruyu değiştirmede yeterli olduğu görülmektedir (Çetin, 2009). Christensen'e (1987) göre, Cook uzaklığı'nun 1'den büyük (CU >1) değer aldığı gözlemler, uç değer içeren gözlem olarak tanımlanmaktadır. Bu durumda bu gözlem veri kümesinden ya atılır, ya model yeniden ele alınır ya da modele yeni değişkenler eklenir (Akt. Altunkaynak, 2003). Fakat çalışmalarda genellikle 4/n (n: gözlem sayısı) değeri ile karşılaştırılarak karar verilir. 4/n değerinden büyük değerlerin uç değer olduğu ifade edilir (Vural, 2007).

1.2.2.3. Kaldıraç Nokta

Veri noktasının merkezinden oldukça uzakta olan nokta, regresyon doğrusunu büyük bir güçle çektiğinden dolayı Kaldıraç Nokta olarak adlandırılırlar (Vural, 2007).

Kaldıraç Noktası, çok değişkenli aykırı değerleri belirlemek için kullanılan istatistiksel bir ölçüdür. Kaldıraç Noktası, Mahalanobis Uzaklığı ile bağlantılıdır fakat, farklı bir ölçekte ölçüldüğü için ki-kare dağılımına dayalı anlamlılık testleri uygulanamaz. Kaldıraç Noktası, 0-1 aralığında değişen bir değer almaktadır. Kaldıraç Noktası sıfıra eşitse, gözlemler üzerinde hiçbir etkisi olmadığı anlamına gelir. Kaldıraç Noktası 1 değerine eşitse gözleme tahmin edilenden çok daha fazla etki ettiği anlamına gelir. Yani, Kaldıraç Noktası 1'e yaklaştıkça, gözlemin aykırı değer içerme olasılığı artar (Field, 2009). Kaldıraç gücü 2p/n'den (p: değişken sayısı, n: gözlem sayısı) büyük olduğunda, Belsley'e (1980) göre yüksek kaldıraç aykırı değerdir. Küçük örneklemler için Vellman ve Welsch (1981) 3p/n (p: değisken sayısı, n: kişi sayısı) ölçütünü önerir (Akt. Vural, 2007).

1.2.3. Aykırı Değerlere Veri Dönüşümü Uygulanması

Veri setinde bulunan aykırı değerlerin istatistiksel analiz sonuçları üzerinde anlamlı bir etkisi olduğu tespit edildiğinde, aykırı değerleri silmeye alternatif olarak, bu verilere dönüşüm uygulanması da uygulanacak adımlar arasındadır. Osborne'a göre veri dönüşümü, bir değişkenin değerini matematiksel olarak değiştiren bir uygulamadır. Veri dönüşümü, dağılımın çarpıklığını normalleştirerek veya varyansların homojenliğini sağlayarak aykırı değerlerin etkisini minimize etmektedir. En çok doğrusallık varsayımının karşılanamadığı durumlarda, veri dönüşümünün uygulanması tavsiye edilmektedir (Howell, 1997). Dönüşümleri kullanarak, aşırı değerler ve göreceli olarak en büyük değerler veri setinde saklanabilir, dahası değişkene ait hata varyansı ve eğim düşürülebilir (Hamilton, 1992). Tabachnick ve Fidell (2001), veri dönüşüm yöntemlerinin kullanılmasının çok değişkenli istatistiksel tekniklerin varsayımlarını karşılamak amacıyla istatistiksel gücü arttırdığını ve yanlılıkları azalttığı için kullanılması gerektiğini belirtmişlerdir. Buna rağmen, dönüşümler test edilen modele uygun olmayabilir ya da istenmeyen bir şekilde yorumlamalara sebep olabilir. Değişkenlerin ortalama ve varyansları birbirinden önemli ölçüde farklı olduğu durumlarda büyük ortalama ve varyansa sahip değişkenlerin diğerlerin üzerindeki baskısı daha fazla olur ve onların rollerini önemli ölçüde azaltır. Ayrıca veri setinde farklı ölçü birimleri kullanılarak elde edilen değişken değerlerinin birimlerinden arındırılması, aşırı değerlerin etkisinin azaltılması, nitel değişkenlerin nicel değişkenlere dönüştürülmesi gibi nedenlerle de veri dönüştürme işlemi kullanılır. Veri dönüştürme

ile analize dahil edilecek değişkenlerin, yapılacak analizlerin varsayımları sağlaması sağlanabilir.

İlgili literatürde dört farklı veri dönüşümü tekniğinden söz edilmektedir (Büyüköztürk, 2009; Hair vd., 2009; Hinton, Brownlow, McMurray ve Cozens, 2005; Osborne, 2002; Tavsancıl, 2008):

1.2.3.1. Karekök Dönüşümü (Square-Root Transformation)

Dağılımın sola çarpık (pozitif çarpık) olduğu durumlarda kullanılmaktadır. Varyans, aritmetik ortalama ile orantılı ise karekök dönüşümü yapılmaktadır.

1.2.3.2. Ters Dönüşüm (Inverse Transformation)

Bu dönüşümde, her bir ölçümün 1 değerine göre tersi alınmaktadır (1/x). Çok büyük değerleri küçük, çok küçük değerleri ise büyük hale getirmektedir. Aşırı derecede çarpık dağılımlarda daha iyi sonuçlar vermektedir.

1.2.3.3. Logaritmik Dönüşüm (Logarithmic Transformation)

Bu dönüşüm pozitif değerler için uygulanır, negatif veriler için uygulanamaz. Tam verilere sabit sayı eklenip negatiflikten kurtarılarak yapılabilmektedir. Logaritmik dönüşüm verileri "sıkıştırarak" bir araya getirir ve aşırı sağa çarpık (aşırı pozitif) dağılımlara uygulanmaktadır. Modeli doğrusallaştırmak (varyansı eşitlemek, normalleştirmek vb.) için yapılmaktadır. Bu yöntem standart sapması çok yüksek olan bir gruba uygulanmamalıdır. Bunun sebebi, bu dönüşüm bireylerin birbirlerinden olan farklılıklarını artıracağından standart sapma değerini de yükseltecektir. Puanların 1'den küçük olması durumunda her bir değere 1 eklenerek dönüşüm yapılmaktadır.

1.2.3.4. Arc Sinüs Dönüşümü (Arc-sinus Transformation)

Bu dönüşüm, veriler oran şeklinde elde edilmişse kullanılır ve doğrusal olmayan iki değişkenli bağlantılarda, bir değişkenin karesini alarak doğrusal olmama problemini etkili bir şekilde azaltmaktadır.

Yapılan analizler sonucunda veri setinde aykırı gözlem değeri var ise, bu durumda ilk olarak aşırı derecede aykırı değerler için sapmanın sebebi tanımlamaya çalışılır ve en önemlisi veri girişi ve kayıtların doğru bir şekilde aktarılıp aktarılmadığıdır. Eğer buradan bir sonuç çıkmazsa, aykırı gözlemleri etkileyecek iç ve dış etkenler

araştırılmalıdır. Eğer buradan da sağlam bir sonuç alınamazsa, aşağıdaki yöntemlerden biri kullanılabilir (Thode, 2001; Alpar, 1997; West, 1999b).

- Aykırı değer veri seti aralığında ise interpolasyon(ara değer bulma) yöntemi kullanılabilir. Eğer gözlem değeri ilk veya son gözlem değeri ise o zaman ekstrapolasyon(dış değer bulma) yöntemi kullanılarak yeni bir değer belirlenebilir (Türkbal, 1981). Aykırı gözlemlerin yerine gözlem değerinin belirlenmesi için kullanılacak en iyi yöntemdir. Eğer bu yöntem yapılamaz ise aşağıdaki yöntemlerden herhangi biri kullanılabilir.
- Örnek sayısı yeterli büyüklükteyse ve aykırı değerler çok değil ise (bir ya da birkaç tane) örnekten çıkartılabilirler.
- Örnek sayısı yeterli büyüklükte değil ise, aykırı değer kendisine en yakın aykırı olmayan değere yuvarlanır.
- Örnek setinde aykırı değer sayısı fazla ise ve bu büyük bir sorun teşkil ediyorsa, uygun bir dönüşüm tekniği kullanılmalıdır.
- Veri setinde aykırı değer varsa ve bu çok büyük bir sorun teşkil etmiyorsa herhangi bir değişiklik yapılmadan analizler yapılır. Çalışmada bu aykırı değerlerin olduğu dipnot olarak düşülür.
- Veriler dönüşüm yöntemi ile normalliğe dönüşmüyorsa, verilerin yapısına uygun olan parametrik olmayan testlere başvurulmalıdır.
- Veriler küçükten büyüğe sıralanır. Sıralanmış veri bölmelere ayrılarak aşırı değerler bulunabilir.
- Veri seti kümeleme analizi ile kümelere ayrılır. Benzer değerler aynı grup veya küme içinde yer alırken, aykırı değerler kümelerin dışında yer alırlar.
- Regresyon yöntemiyle veri setindeki verilere bir fonksiyon uydurularak aykırı değerler bulunabilir. Uydurulan bu fonksiyona uymayan değerler aykırı değerlerdir.
- Değişkenlere ait kutu diyagramları çizilir. Kutu diyagramlarından aşırı değerler gözlemlenebilir.
- Değişkenlerin grafikleri aracılığıyla aşırı değerler bulunabilir.
- Temel bileşenler analizinde elde edilen ilk iki temel bileşenin serpilme diyagramı incelenerek aşırı değerler bulunabilir.

1.3. VERİLERİN STANDARTLAŞTIRILMASI

Çok değişkenli analizde, genellikle birimleri değişik olan değişkenlerle ilgilenilir. Değişkenler farklı değil de aynı birimde oldu mu sonuçlar daha farklı bir şekilde çıkmaktadır. Birimleri farklı olan değişkenlerden daha iyi sonuçlar elde edebilmek için ilgili değişken değerleri standartlaştırılarak aynı birime dönüştürülür. Genellikle kümeleme analizinde standartlaştırılma yoluna gidilir. Benzer şekilde çok boyutlu ölçekleme, regresyon vb. birçok konuda standartlaştırma yaklaşımlarından yararlanılır.

Özellikle uzaklık ölçümleri, çoğu farklı ölçeklere veya değişkenler arasındaki büyüklüklere oldukça duyarlıdır. Genellikle büyük dağılım gösteren yani standart sapması büyük olan değişkenler, benzerlik değeri sonuçlarını daha fazla etkilemektedir.

Değişken sayısı arttıkça, değişkenlerin ölçüldüğü ölçekler de birbirinden farklılık gösterebilmektedir. Bundan dolayı, verileri analize dahil etmeden önce standartlaştırılması gerekir. Ağırlıkları farklı ölçü birimleriyle ölçülen değişkenlerin birlikte analize atanması yanlıştır ve sonuçların hatalı çıkmasına neden olacaktır. Bundan dolayı analizdeki tüm değişkenleri aynı değerle ifade etmek gerekir.

1.4. VERİLERİN NORMALLEŞTİRİLMESİ

Veri analizlerinin uygun bir hale getirilmesi için yapılan dönüştürme işlemlerinden en sık kullanılanı verilerin normalleştirilmesidir. Veriler normalleştirilme işleminden geçirilerek, 0 ile 1 ya da -1 ile 1 aralıklarına indirgenmiş olurlar. Verilerin normalleştirilmesinde kullanılan bazı dönüşümler aşağıdaki gibidir:

1.4.1. Z Skorlarına Dönüştürme

Bu yöntem, oransal ve aralık ölçekli veriler söz konusu olduğunda verilerin çok değişkenli normal dağılım gösterdiği varsayımıyla verilere uygulanan bir yöntemdir. Değişkenlerin standartlaştırılması ile ilgili olarak birçok yöntem olmakla birlikte, en çok kullanılan standartlaştırma formu, değişkenlerin her birini "Z skorları" olarak da bilinen standart değerlere dönüştürmektir. Bunun için $z= (x_i-\mu) /\sigma$ formülünü kullanmak gerekir. Bu formüle göre bütün veriler, aritmetik ortalaması ‘‘0’’ ve standart sapması “1’’ olan bir dağılım haline dönüştürülür; böylece farklı ölçekteki veriler aynı esasa getirilerek standartlaştırılmış olur. Bu yeni skora, *z* değerleri ya da standart değerler

denir. Bu işlemler sonucunda ortalama ($\overline{Z} = 0$) olduğu için, verilen bir puanın ortalamanın altında ya da üstünde olduğu hemen söylenebilir; çünkü ortalamanın üzerindeki değerler pozitif, altındakiler negatif işaretlidir. Ayrıca standart sapma 1 olduğu için standart değerin sayısal büyüklüğü herhangi bir gözlemin aritmetik ortalamadan kaç standart sapma uzakta olduğunu belirtir. Günümüzde artık bu işlemler bilgisayar programlarınca yapılmaktadır. Gelişmiş bilgisayar programları ile standart olmayan birçok değişkenin ve gözlemin işleme alındığı kümeleme analizleri yapılabilmektedir.

İstatistiksel yazılımlarda z standartlaştırması yapan modüller vardır. Örneğin SPSS'de Analyze>Descriptive, Statistics>Descriptives… yolu ile açılan pencerede istenilen değişkenlerin sağ kutuya geçirilmesi sonrasında, ''Save standardized variables as variables'' kutucuğunun işaretlenmesiyle z ile standartlaştırılmış değişkenler veri dosyasına eklenir. Standartlaştırılmış yeni değişkenin adı z ile başlar. Örneğin adı global ise standartlaştırılmış değişkenin adı zglobal olur.

$$Z_i = \frac{x_i - \overline{x}}{s} \qquad (1.2)$$

1.4.2. Diğer Standartlaştırma Yaklaşımları

Z standartlaştırması dışında başvurulan standartlaştırma yaklaşımlarından 7 tanesi aşağıda verilmiştir.

1.4.2.1. Dağılım Aralığı 1 Olacak Şekilde Standartlaştırma

Her bir değer (x_i) dağılım aralığına (R) bölünerek elde edilir. Eğer aralık sıfır (0) ise bütün değerler sıfırdır.

Dağılım Aralığı (Range) = Bir örneklem yığınındaki en yüksek ve en düşük değerler arasındaki fark.

$$x_i' = \frac{x_i}{R} \qquad (1.3)$$

1.4.2.2. $-1 \leq x \leq 1$ Aralığına İndirgeme

Eğer veri seti heterojen bir yapıda ve aşırı değerler söz konusu ise bu yöntem tercih edilir. Daha çok aşırı değerlerin olduğu durumlarda kullanılır. Dönüştürülecek değişkenlerde + veya – değerler var olduğunda ve $|x_{\min}| \leq x_{\max}$ olduğu durumlarda

uygulanır. Dizideki en büyük değer $x_{\max}$ ise, indirgeme formülü aşağıdaki gibidir:

$$x_i' = \frac{x_i - \left(\frac{max + min}{2}\right)}{\frac{R}{2}} \qquad (1.4)$$

1.4.2.3. $0 \leq x \leq 1$ Aralığına İndirgeme

Her bir değerden en küçük değer çıkartılıp, elde edilen değerin dağılım aralığına bölünmesi ile elde edilir. Bu yaklaşım da aşırı değerlerin olduğu durumlarda tercih edilen yaklaşımlardan biridir. Dağılımdaki değerler arasındaki eksi işaretli değerler artı işaretli konuma gelirler. Veri seti heterojen bir yapıda ve aşırı değerler söz konusu ise, değişkenlerin değerleri 0 ile 1 aralığına dönüştürülebilir. Dizideki en büyük değer x_{max}, dizideki en küçük değer x_{min} ve rank $R = x_{max} - x_{min}$ olmak üzere indirgeme şöyle yapılır:

$$x_i' = \frac{x_i - x_{\min}}{R} \qquad (1.5)$$

1.4.2.4. Ortalama 1 Olacak Biçimde İndirgeme

Oluşturulacak olan indirgenmiş değişkenin ortalamasının pozitif ve 1 olması gerektiğinde uygulanan bir yöntemdir. Dönüştürme işlemi şu formülle yapılır:

$$x_i' = \frac{x_i}{\overline{x}} \qquad (1.6)$$

Eğer ortalama sıfıra eşit ise formül şöyle olur:

$$x_i' = \frac{x_i + 1}{\overline{x} + 1} \qquad (1.7)$$

1.4.2.5. Standart Sapma 1 Olacak Şekilde İndirgeme

Eğer indirgenmiş değişkenin standart sapmasının 1 olması isteniyorsa, bu yöntem tercih edilir. İndirgeme formülü şöyledir:

$$x_i' = \frac{x_i}{s_x} \qquad (1.8)$$

Orijinal değişkenin standart sapması 0 ise, bu dönüşüm uygulanamaz. Dönüşümün şart olduğu düşünülürse, diğer yöntemlerden uygun olan bir yöntem seçilir.

1.4.2.6. Maksimum Değer Bir Olacak Şekilde İndirgeme

Her bir gözlemin dağılımdaki en büyük değere bölünmesi ile elde edilir. İndirgenmiş değişkenin maksimum değerinin 1 olması isteniyorsa, bu yöntem tercih edilir.

İndirgeme formülü şöyledir:

$$x_i' = \frac{x_i}{x_{max}} \qquad (1.9)$$

Eğer dizide maksimum değer sıfır ise, formül şöyledir:

$$x_i' = \frac{x_i}{|x_{min}|} + 1 \qquad (1.10)$$

1.4.2.7. Ortalama 50, Standart Sapma 10 Olacak Şekilde Standartlaştırma (T standartlaştırması)

Bu standartlaştırma T standartlaştırması olarak da bilinir. T skorları, çoğunlukla 20 ile 80 arasında değer alır.

$$T=10z+50 \qquad (1.11)$$

BÖLÜM 2

ÇOK DEĞİŞKENLİ GRAFİKLER

Günlük hayatta karşılaşılan olaylar; bazen tek bir değişkenin değil, çok sayıda değişkenin etkisi altında olmaktadır. Bu sebepten dolayı eğer tanımlanacak yapı birden fazla değişkenin etkisi altındaysa, bu yapının gerçeğe uygun en iyi şekilde yorumlanabilmesi için çok değişkenli istatistiksel metotların kullanımı önemlidir. İstatistiksel grafiklerin amacı nicel veya nitel bilginin görsel sunumlarını sağlamaktır. Nicel teknikler için örnekler; varyans, nokta tahmini, güven aralıkları, en küçük kareler regresyonu ve hipotezleri test gibi analizlerdir. Diğer bir teknik de grafik tekniğidir. Bu teknikler de; saçılım grafiği, histogram, kutu grafiği, biplot grafikleri vs.

Metodolojik bir araç olarak istatistiksel grafikler, araştırmacılara incelenen verilerle ilgili önemli bilgiler veren ve araştırma sürecinin sonraki aşamalarını yönlendiren bir takım stratejiler ve teknikler içermektedir. Bu yüzden de istatistiksel çalışmalarda verilerin görselleştirilmesi için genellikle grafiklere başvurulur. Chambers vd. (1983) çalışmalarında, herhangi bir istatistiksel analiz için grafiksel gösterimlerin öneminin olduğunu dile getirmiş ve iyi seçilen bir grafiğin önemini vurgulayarak, bir grafiğin seçiminin doğru olması ve bunun seçimi kadar güçlü herhangi bir istatistiksel araç olmadığını belirtmişlerdir. Bu tanım, özellikle çok boyutlu veri yapısının yorumlanmasında kullanılan çok değişkenli istatistiksel analizler için doğrudur. Çok değişkenli veri kümeleri için çizilecek grafikler bazı durumlarda yazılı bilgilerden daha anlaşılır ve daha açıklayıcı olabilmektedir. Veri görselleştirme, insanın algılama yetenekleri ve insanlar arası yorumlama farklarını dikkate alarak analiz gerçekleştirmeye olanak sağlamaktadır. Veri görselleştirme teknikleri ile etkili bir biçimde verinin portresinin çıkarılması sağlanabilir ve veri hakkında genel bir sonuç çıkarılabilir. Örneğin, veri görselleştirilmesi sayesinde değişkenlerin dağılımları, değişken grupları arasındaki kümelenmeler, korelasyonlar gibi ilişkiler ortaya çıkabilmektedir. Hatta bazı yazarlar ‘‘hiçbir istatistiksel araç, iyi seçilmiş bir grafik kadar güçlü değildir’’ cümlesini sık sık dile getirmektedir. Tablo sunumları ile grafiksel bir sunumu karşılaştırdığımız zaman grafikler okuyucunun ilgisini daha çok çekmektedir. Bunlarla birlikte, grafik sunumlardan yararlanmanın birçok nedeni ve avantajı söz konusudur. İyi oluşturulmuş istatistiksel grafik, verilerdeki güvenlik

mesajını açık ve hızlı bir şekilde iletebilir.

Birçok veri analizinde, verilerin birbiriyle olan ilişkilerinin iyi anlaşılması oldukça büyük önem taşımaktadır. Veri analizi sırasında, insan algı sistemiyle bilgisayar sistemleri arasında bir bağlantı kurar ve bu bağlantıyı kurmanın en iyi yolu verinin görselleştirilmesidir. Veri analizi esnasında, analizciler toplanan veriler arasında yapılar, örüntüler, ilişkiler aramaktadırlar. Verilerin grafiksel bir şekilde gösterilmesi, analizcinin veri yapılarını anlamasını kolaylaştırmaktadır. Ancak çoğunlukla analistçiler çok boyutlu verilerle uğraşırlar. İnsanların özellikle araştırmacıların algılama sistemleri yalnızca üç boyutla sınırlı olduğu için daha fazla boyut içeren veriler insan algı sisteminin dışına çıkmaktadır. Bu sebepten dolayı, veri görselleştirme teknikleri çok boyutlu veriyi iki veya üç boyuta indirgeyerek görselleştirip, diğer taraftan da veriler arasındaki ilişkiyi korumalıdır. Bu indirgeme sırasında bir miktar bilgi kaybı söz konusudur. Görselleştirmedeki temel amaçlardan bir tanesi de bu kaybı minimize etmektir.

Bir görselleştirme türü olan çok değişkenli veri görselleştirme, bilim toplulukları ve mühendislik tasarımından, endüstri ve finansal piyasalara kadar çeşitli alanlarda sayısız uygulama içeren, birden fazla özellik arasındaki ilişkinin hayati bir önem taşıdığı etkin bir araştırma alanıdır. Çok değişkenli veri görselleştirme, bilgi görselleştirme ile aynı zorlukları çoğu zaman yaşamaktadır. Bu yüzden de bir sorunun iyi görsel sunumlarını bulmak zor olabilir ve bu durum bazen bizi kararsız kılabilir. Aynı zamanda, çok değişkenli veriler tek bir görsel ekranda bazen özniteliklerini kodlarken sorunlar ortaya çıkarmaktadır. Çok değişkenli veri görselleştirmesinin asıl amacı, verileri incelerken, farklı nitelikler arasındaki muhtemel ilişkiyi göstermektir. Çoğu zaman, görsel görüntüye bakmadan önce bazı korelasyonlar keşfedilmemiş olur. Bu yüzden tam olarak görselleştirmeden sonra elde etmek istediğimiz özellikler bunları ortaya çıkarmaktır. Verilerde ne kadar değerli bilginin bulunduğunu bilmiyoruz, bu nedenle onu görselleştirerek fikir sahibi olmak istiyoruz. Bununla birlikte, veri gösteriminde gösterilecek kalıp veya ilişki hakkında hiçbir şey bilmediğimiz takdirde, belirli bir görselleştirme tekniklerinin etkililiğini asla değerlendiremeyiz.

Gözlenen her bir birim ikiden fazla bilgi sağladığında bu bizim çok değişkenli verilere sahip olduğumuzu gösterir. İstatistikte bu verilere sık sık rastlarız. Örnek verilecek olursa gözlemlenen bir vakanın önemli özelliklerini tanımlamak için çoğunlukla ikiden fazla veri gerekmektedir. Değişkenler arasındaki ilişkiler doğrudan ilgi konusu olduğunda, iki değişkenli veya çok değişkenli verilerin açıklanması için kimi zaman sunum zorunlu hale gelmektedir. Bu nedenle, çok değişkenli grafiksel yöntemler için bilginin ilk kaynağı, değişkenler arasındaki ilişkileri bulmak ve göstermektir. İkinci bir bilgi kaynağı, benzer gözlemsel grupların belirlenmesinde çeşitli değişkenlerle ilgili bilginin yararlı olabileceği gerçeğidir.

Çok değişkenli grafikler çizilirken dağılım aralıklarının çok farklı olduğu durumlarda değişkenler için standartlaştırılmaya gidilmesi iyi bir yaklaşım olur. Böylece, değişkenlerin ortalamaları ve yaygınlıkları birimden bağımsız bir konuma getirilmiş olur.

Grafik yöntemlerin bazı avantajları Schmid (1954) tarafından listelenmiştir:

1. Diğer sunum türleriyle karşılaştırıldığında, iyi tasarlanmış grafikler, ilgi yaratmada ve okuyucunun ilgisini çekmek için daha etkili olur.

2. Grafiklerle tasvir edilen görsel ilişkiler daha kolay anlaşılır ve daha kolay hatırlanır.

3. Grafiklerin kullanımı zaman kazandırır; çünkü istatistiksel verilerin büyük ölçülerinin önemli anlamı bir bakışta görselleştirilebilir.

4. Grafikler, tablo biçiminde veya metinsel sunum biçimlerinden elde edilenden daha eksiksiz ve daha dengeli bir anlayışa neden olan bir sorunun kapsamlı resmini sunar.

5. Grafikler, gizli gerçekleri ve ilişkileri ortaya çıkarabilir ve aynı zamanda analitik düşünmeye ve soruşturmaya teşvik edebilir.

6. Grafikler, büyük ve karmaşık veri setleri için yararlı özetler sunar.

7. Grafikler, verilerin ayrıntılarını göz ardı etmenin ve önemli özelliklerin vurgulanmasının etkin bir yolunu sunar.

8. Grafiksel analiz, araştırmacı ve veriler arasındaki etkileşimi daha fazla kolaylaştırır.

9. Grafikler, genellikle boyutlarına bakılmaksızın veri setleri içindeki desenleri, eğilimleri ve göreli miktarları açığa çıkarmak için faydalı bir araçtır.

İstatistiksel grafikler, verilerin özel etkilerini araştırır, aykırı değerleri gösterir, kalıpları belirler, modelleri teşhis eder ve genellikle yeni ve belki de beklenmedik olayları araştırır. İstatistiksel grafikler, bir veri kümesinin içeriğini keşfetmek için kullanışlıdır. Bir analizdeki değişkenler hakkındaki sorulara cevap bulmak için kullanılabilir (dağılım şekiller, aralıklar, tipik değerler veya aşırı değerler). İstatistiksel modellerde, varsayımların kontrolü için kullanılabilir. İstatistiksel bir modelin doğrudan görsel sunumları ve modelin kalıntıları varsayımların incelenmesini büyük ölçüde kolaylaştırır. Verideki özel etkileri yakalamak, beklenmeyen durumları göstermek için kullanılır. Grafik gösterimler, veriler hakkında bir hikâye anlatmayı ve karşılaştırma yapmak için gereken bilişsel çabayı azaltmayı amaçlamalıdır.

2.1. Çok Değişkenli Saçılım Grafikleri

Çok değişkenli olan bir verinin değişkenlerinin, teker teker incelenmesinde tek değişkenli grafiksel yöntemlerden yararlanılabilir. Bu tek değişkenli grafiklerden ilki histogram, çeşitli aralıklara düşen veri noktalarının sayılarını gösteren özet bir grafiktir. Verinin frekans dağılımı hakkında kaba bir yaklaşımdır. İkincisi kutu grafikler; çoğunlukla verinin ortalaması, çeyreklikleri ve aralığı bilgilerini görüntüleyen grafiktir. Sonuncusu dal-yaprak grafikleri John Tukey tarafından geliştirilmiş histogram tarzı bir veri cetvelidir. Veri grubundaki sayıların kendileri ile bir diyagram oluşturmak için kullanılır. Aynı zamanda bu grafikler dağılımların şekillerini, yani verilerin simetrikliği ve çarpıklığı hakkında bize bilgi vermektedir. Bu grafiklerin bir kusuru bir veri grubundaki en az kesinlikteki noktalar olan dağılım kuyruklarını vurgulama eğilimi göstermeleridir. Ayrıca dağılım hakkındaki pek çok ayrıntıyı saklarlar. Yani çok değişkenli bir veri kümesinin değişken değişken incelenmesi değişkenler arasındaki ilişkileri açığa çıkarmadığından dolayı, pek de kullanılan bir yaklaşım değildir. Bundan dolayı, çok değişkenli veriyi özetlemek gerekirse, veri kümesini en iyi şekilde resmedecek grafiklerden faydalanılır.

Saçılım grafikleri, iki boyutlu bir grafik üzerinde veri noktaları görüntüleyerek iki değişken arasındaki ilişkiyi gösterir. Açıklayıcı olarak düşünülen değişken x ekseni

üzerine ve bağımlı değişken de y ekseni üzerine çizilir. Saçılım grafikleri, özellikle çok sayıda veri noktası olduğunda faydalıdır. İki değişken arasındaki ilişki hakkında aşağıdaki bilgileri sağlarlar:

- Kuvvet
- Şekil (doğrusal, eğrisel vb.)
- Yön (artı veya eksi)
- Uç değerlerin varlığı

Bir saçılım grafiği ile iki değişken arasındaki ilişki gösterildiği zaman mutlaka bir sebep-sonuç ilişkisi bulunması gerekmez. Her iki değişken de bunların değişimini açıklayan başka bir üçüncü değişkene veya tamamen başka bir sebebe bağlı olabilir. Alternatif olarak, göze çarpan bir ilişki basit bir tesadüfün eseri de olabilir

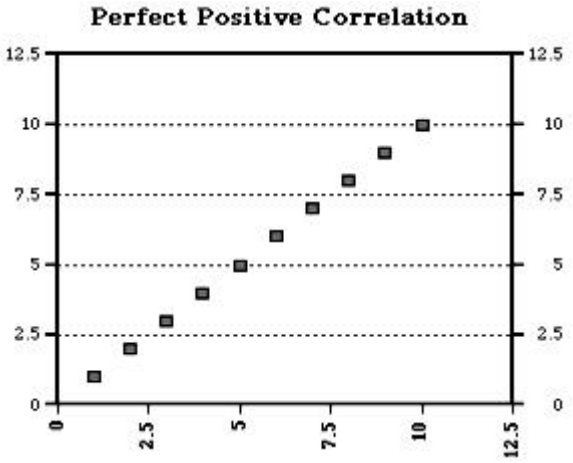
Çok değişkenli veriler için istatistikçiler tarafından sevilen, en bilinen, en basit grafik türü saçılım grafikleridir. Bir dağılım grafiği, genellikle doğrusal bir korelasyon katsayısı çalışmadan veya bir regresyon çizgisine uymadan önce çizilen iki değişkenli verilerin yararlı bir özetidir. İki değişken arasındaki ilişkinin iyi bir görsel resmini verir ve korelasyon katsayısı veya regresyon modelinin yorumlanmasına yardımcı olur. Çok değişkenli istatistikte değişkenler arasındaki ilişkilerin doğrusal olup olmadığını aralarında ilişkinin olup olmadığını görmek için saçılım grafiklerinden yararlanılır. Yani iki sayısal değişkenin beraber değişiminin analiz edildiği grafik türüdür.

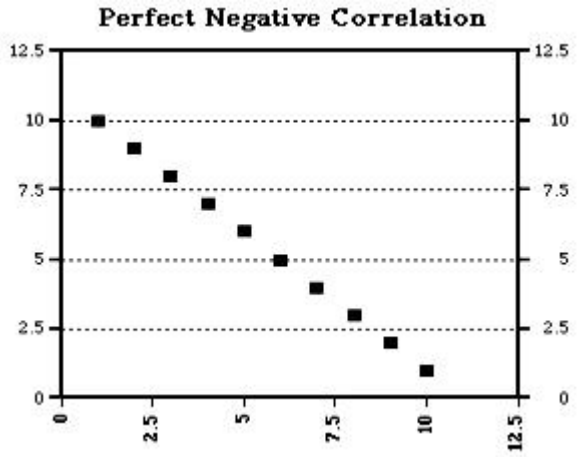
İki veya daha fazla değişken arasındaki korelasyon veya neden-sonuç ilişkisini değerlendirmek için saçılım grafiği matrisi kullanılır. İki veya daha fazla gösterge arasındaki nedensellik bağlantılarını araştırırken, korelasyonları hızlı bir şekilde tanımlamak için saçılım grafiği matrisini kullanabilirsiniz. Aşağı doğru eğimli bir dağılım, yatay eksendeki değişkeni artırdığımızda dikey eksendeki değişkenin azaldığını gösterir. Yukarı eğimli saçılımlar için benzer bir açıklama yapılabilir. İki sayısal (sürekli ya da kesikli) değişken arasındaki ilişkilerin grafikle incelenmesi Pearson korelasyon katsayısının doğru kullanımı ve ilişkinin modellenmesi açısından son derece önemlidir. Ancak, korelasyonun nedensellik olmadığını ve fark edilmeyen başka bir değişkenin sonuçları etkileyebileceğini unutmayalım.

İkiden fazla değişkenin bulunduğu çok değişkenli veri kümelerinde, her değişken çiftine ilişkin saçılım grafiğinin tek bir grafik üzerine yerleştirilmesi ile elde edilen p×p boyutlu matris şeklindeki saçılım grafiklerine de sıklıkla başvurulur. Böylece, tüm ikişerli saçılım grafiklerinin aynı anda incelenmesi sağlanır. Saçılım matrisleri, iki ve daha fazla değişkenin aynı anda birbirileriyle ikili ilişkilerini (karşılıklı ilişki) gösteren bir grafiktir. Saçılım matrislerinde n tane değişkenin (boyut) $n(n-1)/2$ tane ikili ilişki grafiği bulunur. Saçılım matrisleri, ikili değişkenler arasındaki korelasyonların tespiti açısından sağladığı kolaylıklar sayesinde en sık kullanılan çok boyutlu görselleştirme tekniğidir. Ancak veri boyutumuzun büyük olması nedeniyle saçılım matrislerinden bilgi elde edilmesi zorlaşmaktadır.

Dağılım grafiği, değerler arasındaki potansiyel ilişkileri bulmamıza ve veri kümelerindeki aykırı değerleri bulmamıza yardımcı olur. Dağılım grafiği, aynı anda iki veya daha fazla önlemin korelasyonunu görselleştirmek için harika bir yoldur. Dağılım grafiğinin deneyimsiz bir kullanıcı için anlaşılması zor olabilir, çünkü her iki eksende de ölçüm değeri vardır ve üçüncü, isteğe bağlı olan ölçüm, yorumlamaya karmaşıklık katar. Açıklayıcı etiketlerin kullanılması, görselleştirmenin yorumlanmasını kolaylaştırmak için iyi bir yoldur. Değerler üst üste yerleştirilebilir ve siz yakınlaştırıncaya kadar görünmez.

Zamanı dikkate almaksızın, çok fazla sayıda veri noktasını karşılaştırmak istediğimizde saçılım grafiklerini kullanırız. Saçılım grafiğine ne kadar çok veri eklersek, o kadar iyi karşılaştırmalar yapabiliriz. Saçılım grafikleri, veri noktası kümelerinin ve değerlerinin dağılımlarının işlenmesi için idealdir. Veri kümesi çok sayıda veri noktası içeriyorsa kullanabileceğiniz en iyi grafik türü budur.



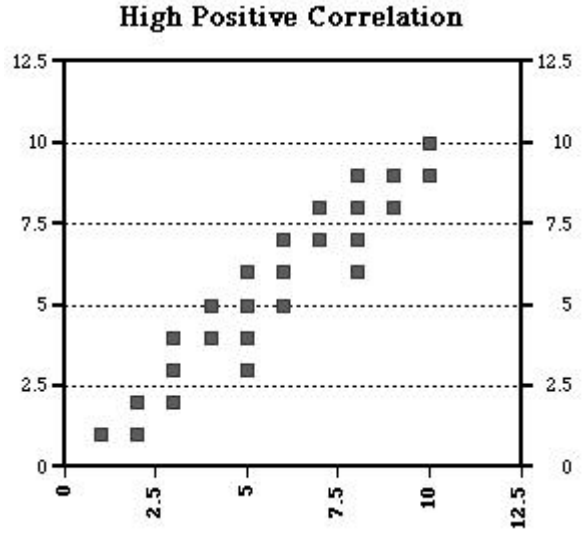


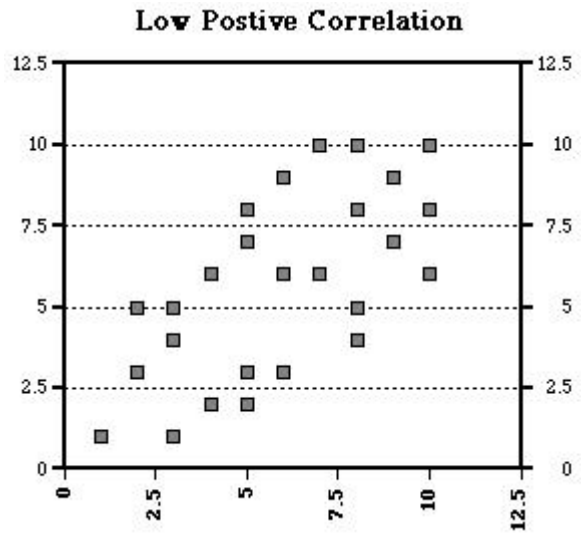
Saçılım diyagramı, her bir eksende bir değişken olan sayısal verilerin çiftlerini, aralarındaki ilişkiyi bulmak için kullanılan grafiklerdir. Değişkenler korelasyonluysa, noktalar bir çizgi veya eğri boyunca düşecektir. Korelasyon ne kadar iyi olursa, noktalar da o denli sıkılaşacaktır.

Mükemmel bir pozitif korelasyon oluştuğunda ‘‘1’’ değeri verilir. Mükemmel bir negatif korelasyon için ise ‘‘-1’’ değeri verilir. Hiçbir korelasyon mevcut değilse, verilen değer ‘‘0’’olur. Sayı ‘‘1’’ veya ‘‘-1’’e yaklaştığında, korelasyon daha güçlü veya değişkenler arasındaki ilişki daha güçlü olacaktır. Sayı ‘‘0’’a yaklaştığında korelasyon zayıf olacaktır.

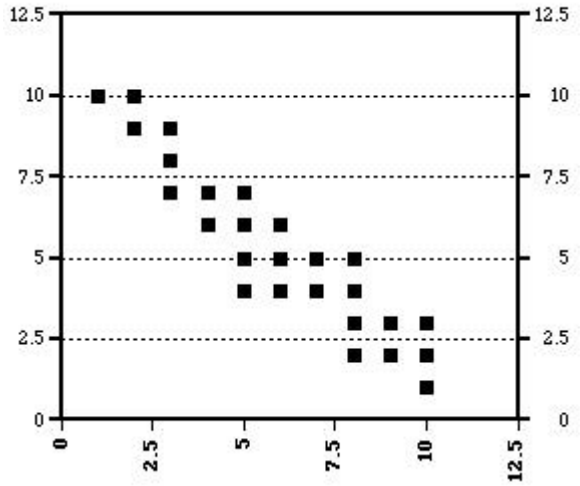
Bir dağılım şeklinde, iki koordinat vardır: İlki, çift içindeki ilk veriye karşılık gelir. Bu X koordinatıdır, sola veya sağa gittiğimiz değişkenlerin yer aldığı koordinattır. İkinci koordinat, çift içindeki ikinci veriye karşılık gelir. Bu da Y koordinatıdır, yukarı veya aşağı gittiğimiz değişkenlerdir. Veriler soldan sağa

doğru ilerlerken yokuş yukarı bir desen gösteriyorsa, X ve Y arasında pozitif bir ilişki olduğunu gösterir. X değerleri arttıkça (sağa ilerledikçe), Y değerleri artmaya yani, yukarı taşınma eğilimi gösterir. Soldan sağa doğru hareket ederken veriler yokuş aşağı bir desen gösteriyorsa, X ve Y arasında negatif bir ilişki olduğunu gösterir. X değerleri arttıkça (sağa ilerledikçe) Y değerleri azalmaya yani, aşağıya inme eğilimi gösterir. Veriler herhangi bir desene benzemiyor gibi görünüyorsa, X ve Y arasında hiç bir ilişki yoktur.

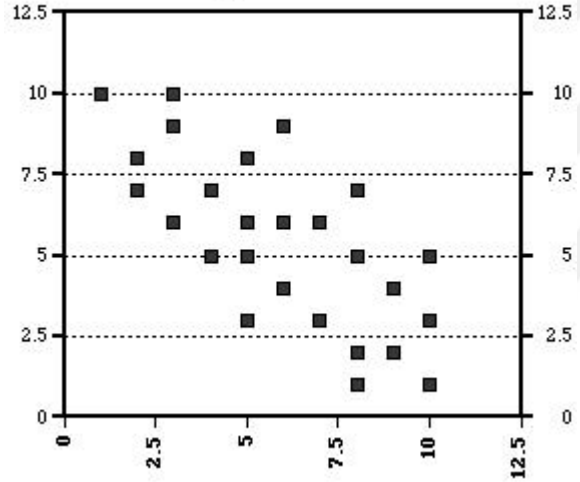




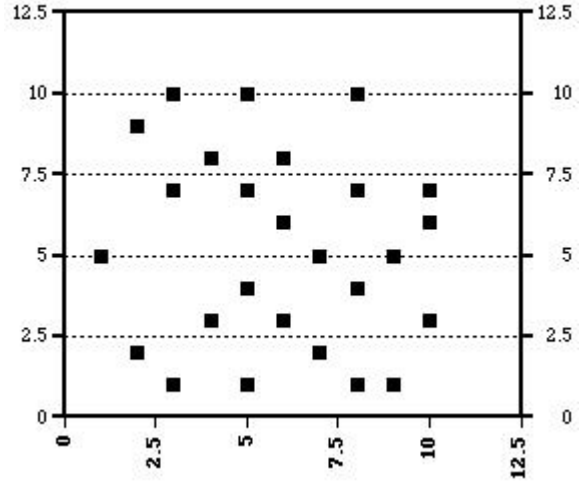
High Negative Correlation



Low Negative Correlation



No Correlation



2.2. Kabarcık Grafikleri

Kabarcık grafikleri, bir veri noktasını kabarcık boyutuna göre iki değer arasındaki farkı görüntülemek için kullanılır. Kabarcık ne kadar büyükse, iki değer arasındaki fark da o kadar fazla olur. Bu grafikler iki ve üç boyutlu saçılım grafiklerine benzer; ancak üçüncü boyut kabarcıklarla ifade edilir. Bu yaklaşımda, kabarcığın merkezi iki değişkenin kesim noktasında yer alırken kabarcığın büyüklüğü (çapı) üçüncü değişkenin değerlerine göre değişiklik gösterir. Bu nedenle, bu grafik temel olarak klasik saçılım grafiğinin üç değişken için geliştirilmiş şeklidir. Kabarcık grafikler yardımıyla, saçılım grafiğindeki gibi araştırmacının değişkenler arasındaki ilişkileri görsel olarak incelenmesi ve olası aykırı değerleri saptaması söz konusu olabilmektedir. Kabarcık grafikleri, veri kümesinde dört değişken olması durumunda kolaylıkla kullanılabilmektedir. Bu tür bir veri kümesinde, verilerin çok değişkenli saçılımlarını incelemek üzere x, y ve z koordinatlarına ek olarak kabarcığın büyüklüğünün de kullanıldığı üç boyutlu grafiklerden yararlanılmaktadır.

2.3. Kontur çizgisi

Bir kontur çizimi, iki boyuttaki üç sayısal değişken arasındaki ilişkilerin grafiksel bir temsilidir. İki değişken X ve Y eksenleridir ve üçüncü değişken Z, kontur seviyeleri içindir. Kontur seviyeleri eğri çizilir; Eğriler arasındaki alan, enterpolasyonlu değerleri göstermek için renkle kodlanabilir. Kontur grafiğindeki gözlemleri değiştirebilir, tanımlayabilir ve etiketleyebilir, grafiğin yönünü kontrol edebilir ve eksenlerde gösterilen bilgileri kontrol edebilirsiniz.

Diğer bir taraftan saçılım grafiklerinin farklı bir sunumu olup, merkez çevresinde istenen yüzde sınırları içerisinde kalan ve kalmayan gözlemlerin belirlenebilmesi açısından çizilen özel bir grafik türüdür. Bir kontur grafiği, büyük veri kümelerinde aşırı lekelenmeyi önlemek için özellikle yararlı olan tahmini nokta bulutlarının yoğunluğunu gösterir.

Kontur grafiği, üç değişken arasındaki potansiyel ilişkiyi keşfetmek için kullanabileceğimiz bir grafiktir. Yani herhangi bir verinin, bir diğer 2 bağımsız değişkendeki değişikliklere dayanarak bir değişkenin değerindeki değişikliği görebilmek için oluşturulan grafik türüdür. Örneğin; bir kontur grafiği, boylam, enlem ve yükseklik yerine x, y ve z değerlerinin çizildiği topografik bir harita gibidir. Kontur

grafiği, 2 boyutlu bir formatta sabit *z* dilimleri çizerek, 3 boyutlu bir yüzeyi temsil eden grafiksel bir tekniktir. Yani, *z* için bir değer verildiğinde, bu *z* değerinin oluştuğu yerde x, y koordinatlarını bağlamak için çizgiler çizilir.

Örn: Sıcaklık ve Basınç, bölgesel rüzgarın gücünü nasıl etkiler?

2.4. Paralel Koordinatlar

Bu teknik ilk olarak 1985 yılında Inselberg tarafından bulunmuş olup, Wegman tarafından 1986 yılında çok boyutlu verileri analiz etmek ve görselleştirmek için geliştirilmiştir (Martinez ve Martinez 2005). Paralel Koordinatlar, k-boyutlu veri setini 2 boyutlu uzaya indirgeyen görselleştirme tekniği olup, birçok değişkeni birlikte karşılaştırmak ve aralarındaki ilişkileri görmek için kullanılan grafiksel bir yöntemdir. Paralel koordinatlarda, her bir değişkene kendi ekseni verilir ve tüm eksenler birbirine paralel olarak yerleştirilir. Her bir eksen farklı bir ölçü biriminde çalıştığından her eksen farklı bir ölçek alabilir veya tüm ölçekleri eşit tutmak için tüm eksenler normalleştirilebilir. Her eksen üzerindeki değer işaretlendikten sonra bu değerler düz çizgiler ile birleştirilir. Bu tekniğin en büyük dezavantajı bir kaç bin adetten daha fazla nesne içeren veri setleri için uygun olmamasıdır. Nesne sayısı arttıkça üst üste binen çok sayıda çizgi görüntüyü yorumlanabilir olmaktan çıkarmaktadır. Değerler, tüm eksenlere bağlanan bir dizi çizgi olarak çizilir. Bu, her bir çizginin her biri birbirine bağlanmış olan, her eksene yerleştirilmiş olan noktaların bir koleksiyonu olduğu anlamına gelir.

Eksenlerin düzenlenme sırası, okuyucunun verileri anlama şeklini etkileyebilir. Bunun bir nedeni, bitişik değişkenler arasındaki ilişkilerin sonradan bitişik olmayan değişkenlere göre algılanmasının daha kolay olmasıdır. Bu yüzden eksenleri yeniden sıralamak değişkenler arasındaki kalıpları veya korelasyonları görmeye yardımcı olabilir. Paralel koordinatlar kümeleri, sapan değerleri, değişken çiftleri arasındaki korelasyonları bulmada kullanılabilirler. Kategorik değişkenler, paralel eksenlerden bir tanesi olmak üzere grupları ya da sınıfları belirtmekte kullanılabilir. Burada her kategori için değişik renkler kullanılarak grupları veya sınıfları ayrıştırabiliriz.

2.5. İkon Grafikleri

Çok değişkenli veriler için, geometrik gösterimlere alternatif olarak ikon grafikleri kullanılmaktadır. İkon grafikleri, bir tablodaki sayıları simgelere dönüştürmek için kullanılan en etkili grafiksel yöntemlerdendir. Bu simgeleme yöntemi sonucunda elde edilen veri görüntüsünün anlamını kolaylıkla anlayabilir, aynı zamanda görüntülenebilecek maksimum boyut ve veri öğelerinin sayısını belli bir sınırlamaya bağlı kalmadan sembolize edebiliriz. İkon grafikleri, her bağımsız değişkeni bağımlı değişkenleri temsil etmek üzere değişebilen ve bunları çeşitli niteliklere göre temsil etmede kullanır. Bu grafiklerde bağımsız gözlem birimlerinin boyutları, bu değişkenlerin alacağı ya da almış olduğu değerlere göre oluşturulan grafiksel sembollerle gösterilir. Bu tekniklerde, her veri öğesi bir simge ile temsil edilir ve özelliklerin değerleri simgenin rengine, şekline ve yönüne atanır. Ardından, simgeler grafik olarak çizilir ve analist veri nesnelerini bir bütün olarak görüntüleyerek bilgi alabilir. Grafik olarak çok boyutlu ikon grafikleri; yapıları, renk, uzunluk, genişlik, yönlendirme, şekil, boyut, desen ve doku gibi çeşitli görsel değişkenleri göz önünde bulundurur ve böylece geleneksel piksel görüntülerinden daha fazla bilgi taşıyabilir.

İkonografik veya ikon temelli teknikler, her çok boyutlu veri öğesini bir simgeyle veya daha özel olarak bir ikonla eşler. Görsel özellikler, veri niteliği değerlerine bağlı olarak değişir. Çeşitli grafiksel parametreler genellikle bir simgede bulunur ve bu da çok boyutlu verilerin işlenmesini mümkün kılar. Ayrıca, grafiksel özelliklerin gözlemleri insan tarafından memnuniyetle karşılanmaktadır. Bununla birlikte, tüm boyutları eşit olarak işleyen geometrik tekniklerin aksine, ikonlardaki bazı özellikler diğerlerinden daha belirgindir, bitişik öğelerin ilişkilendirilmesi daha kolaydır ve farklı grafiksel özelliklerin algılanma doğruluğu insanlar arasında büyük farklılıklar gösterir. Böylece sonucun yorumlanmasında önyargılar ortaya çıkar.

İkon grafiklerinin bir avantajı, belirli bir veri setinin tüm veri değerlerini açık ve anlaşılır bir şekilde birleştirmeleri ve böylece kullanıcıların çok değişkenli korelasyonları tanımaları amaçlanmaktadır. Burada ikonların boyutları, yapılan gözlemlere bağlı olarak değerlerin bir fonksiyonu olacak şekilde hazırlanır. Bu şekilde, çok değişkenli her bir gözlem adına ilgilenilen değişken değerlerine göre eşsiz olan ve araştırmacılar tarafından da kolayca ayırt edilebilen bir resim oluşturulmuş olur.

İkon grafiklerinin çok değişkenli veri analizlerinde kullanılması oldukça etkilidir. Bu grafikler çok değişkenli veri kümeleri hakkında genel bir bilgi edinilmesinde, özellikle aşırı değerlerin saptanmasında ve ilgilenilen değişkenlerin belirli bir gözlem üzerindeki ağırlıklarının belirlenmesinde oldukça etkilidir. İkon grafikleri özellikle kümeleme analizi sonrasında sonuçların görsel olarak doğrulanması amacıyla da sıklıkla kullanılan bir grafik türüdür. İkon grafikleri kullanmış olduğumuz görsel ikonun türüne göre farklı isimler almaktadır. Dairesel ikon grafikleri grubunda en sık kullanılan grafik türleri yıldız, poligon ve güneş ışığı grafikleridir. Her bir gözlem için değişken değerlerinin çubuk ya da çizgi grafiklerle gösterilmesi ise adımsal ikon grafikleri grubu içinde yer almaktadır. Diğer taraftan oldukça yaygın olarak kullanılan ve tercih edilen bir ikon grafiği türü olan Chernoff yüzleri ise kendi başına bir kategori olarak kabul edilebilir.

2.5.1. Chernoff Yüzleri

1973'te Herman Chernoff tarafından bulunan bu yaklaşım, çok değişkenli bir veri kümesindeki aşırı değerleri belirlemek için kullanıldığı gibi, çok değişkenli bir veride kendi içlerinde benzer bireyler barındıran farklı kümeleri belirlemekte kullanılan kümeleme çözümlemesinde de sıklıkla kullanılan bir yaklaşımdır. Aynı zamanda Chernoff, çok değişkenli verilerdeki örüntüleri, kümeleri, korelasyonları, trendleri bulmak ve bunları sunmak için The Journal of the American Statistical Association (Amerikan İstatistik Derneği Dergisi) dergisinde "The Use of Faces to Represent Points in K-Dimensional Space Graphically (K Boyutlu Uzaydaki Noktaları Sunmak İçin Yüzlerin Kullanımı)" adlı makalede ilk defa Chernoff yüzlerini görselleştirme yöntemi olarak dile getirmiştir. Çok boyutlu verilerle uğraşan Herman zorluklar yaşadığını dile getirmekteydi. N-boyutlu veri kümesini görselleştirmenin zorluğundan dolayı bunun yerine bu verilerin 2 boyutlu bir şekilde temsil edilebileceği yüzleri çizmeye başlamıştır. Bu görselleştirmeyle insanlar yüzlerdeki çok küçük farklılıkları bile fark edecekti. İnsanlar, yüz ifadelerini yorumlamada ne kadar iyi olurlarsa, bu beceriyi de böyle temsil edilen grafikleri yorumlama konusunda benimseyebileceklerini düşünüyordu. Herman Chernoff çok değişkenli karışık verileri insanların kolayca algılayabileceği yüz şekillerine yani karikatürlere benzetmeye çalışmıştır. Böylece, insanların kolayca yüz farklılıklarını algılaması özelliğinden faydalanılarak verinin temelleri hakkında görsel bilgi elde edilmesi amaçlanmıştır (Spinelli ve Zhou).

Herman Chernoff'un ortaya attığı bu yöntemde 18 ve daha az değişkenli veriler, her bir gözlem değeri karikatürize edilmiş bir yüzde temsil edilirler. Chernoff yüzleri yaklaşımı, okuyucunun daha kolay yorumlayabilmesini sağlayabilmek için farklı yüz özelliklerinin (burun eğriliği, kaş uzunluğu, kaş kalınlığı, kulak yarıçapı, gözün dış merkezliği, burun genişliği, yüz genişliği, gözler arası açıklık, gözlerin büyüklüğü) gibi değişkenlerin aldığı değerlerle doğru orantılı olacak şekilde yüzler çizilmesi yaklaşımıdır. Diğer boyutların aldığı değerler ile orantılı olarak insan yüzünün burun, ağız, kulak, göz ve yüz şekli değiştirilir. Değişkenlerin yüzlerle ifade edilmesi, yüzdeki herhangi bir bölge ile ilişkilendirilebilir. Örneğin değişkenlerden biri kaş uzunluğu ile ilişkilendirilirken, bir diğeri yüz genişliği, bir diğeri göz bebeği büyüklüğü, bir diğeri kaş uzunluğu vb. ile ilişkilendirilebilir. Sonuç olarak, bir insan yüzündeki çeşitli farklı özelliklerden faydalanarak her bir gözlemi çok sayıda değişken açısından görsel bir şekille yani, insan yüzüyle temsil edebilebiliriz. Diğer taraftan, yüzler saçlı bir şekilde çizilebildiği gibi saçsız da çizilebilir. Saçlı olarak yapılan çizimlerde saçların koyuluğu ve açıklığı ya da yana yatmışlığı da bir özellik olarak dikkate alınmalıdır.

Chernoff yüzlerinin en önemli avantajlarından biri, okuyucuya yabancı olmadıkları bir görsel sembol sağlamasıdır. Okuyucu, yüz ifadelerini kolay ve doğru bir şekilde tanıyabileceği için bu tür görsel bir sunumu da kolayca anlayabilecektir. Verideki değişken sayısı arttıkça yüzlerdeki şekillenmeler daha çarpıcı konuma gelebilmektedir. Bu tekniğin bazı dezavantajları da söz konusudur. Chernoff yüzlerinin en büyük dezavantajı değişkenlerin nicel görselleştirmelerine olan yoksunluğudur. Çünkü biz Chernoff yüzlerinde değişkenleri, örüntüleri, kümeleri, korelasyonları ve trendleri anlamak için sayısal bir inceleme yaparız (Martinez ve Martinez, 2005). Chernoff yüzlerinin diğer bir dezavantajı ise insan yüzündeki bazı organların diğerlerine göre daha fazla dikkat çekmesidir. Örneğin gözler kulaklardan daha dikkatli algılandığı için karşılaştırma yanılgıları oluşabilir (Bilgin ve Çamurcu, 2007).

Chernoff Yüzleri genellikle yıllar boyunca eleştirilere maruz kalmıştır. Örneğin, belirli bir yüz türünü nasıl yorumlayacağına dair kişisel algılama son derece doğrusal olmayabilir. Farklı izleyiciler, farklı algılama durumlardan dolayı aynı yüzü farklı şekillerde yorumlayabilir. Üstelik yüzlerin nasıl yorumlanacağı, kişisel algılaması izleyiciden izleyiciye farklı olabilir. Özellikler iyi seçilmişse, kümeleri, aykırı değerleri

ve diğer gizli durumları görmek çok daha kolaylaşacaktır, ancak seçim başarısız olursa diğer taraftan bu durumlar çok daha zorlaşacaktır.

2.5.1.1. Chernoff Yüzlerinin Yapısı

Orijinal yüz 1973 yılında Herman Chernoff tarafından tasarlandı ve uygun aralıklarda (0-1) 18 değişken; $x_1, x_2, \ldots, x_{18}$ kullanıldı. Farklı yüz özellikleri aşağıdaki gibidir:

x_1 : Chernoff 1973 yarıçapı yüzünün köşesine, | OP |

x_2 : OP'nin açısı yatay

x_3 : Dikey yüz büyüklüğü, | OU |

x_4 : Üst yüzün dış merkezliliği

x_5 : Alt yüzün dış merkezliliği

x_6 : Burun uzunluğu

x_7 : Ağız dikey pozisyon

x_8 : Ağız eğimi

x_9 : Ağız genişliği

x_{10} : Gözlerin dikey pozisyonu

x_{11} : Gözlerin ayrılması

x_{12} : Gözlerin eğimi

x_{13} : Gözlerin dış merkezliliği

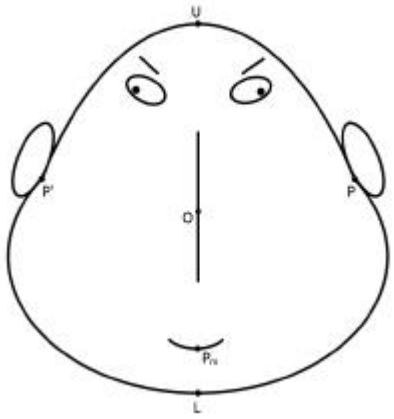
x_{14} : Göz boyutu

x_{15} : Göz bebeğinin konumu

x_{16} : Dikey pozisyonda kaşlar

x_{17} : Kaşların eğimi

x_{18} : Kaş büyüklüğü



Şekil 2.1. Chernoff Yüzü Yapısı Örneği

Yukarıdaki şekilde, yüzün taslağı iki elipsden oluşur. Merkez noktadan O bir ışın bir köşe noktasına P çekilir. Karşı tarafta ikinci bir köşe noktası P' alınır, böylece OP ve OP' nin her ikisi de aynı uzunlukta ve dikey olarak simetrik olur. Sırasıyla U ve L ile temsil edilen yüzün üstü ve alt kısmı OU ve OL'nun dikey olması ve eşit uzunlukta olması için seçilmektedir. Yüzün üst kısmı U, P, P' ve x_4 ile belirlenen bir elipsdir. Alt kısım benzer şekilde L, P, P' ve x_5 ile tanımlanır. Burun burun delikli O' dur; ağız dairesel bir yay şeklindedir. Gözler elips şeklinde gösterilir. Kaşlar gözlerin üstünde eğik çizgi parçalarıdır. Her iki gözü ve kaşları dikey olarak O'dan geçen dikey çizgiye göre simetriktir. Yüzün genişliğinin ve yüksekliğinin iki santimetreye kadar normalleştirilmesi, daha sonra x_1 ve x_3'ün etkilerini azaltmak için uygulanır. Üç dışmerkezlik ve ağız eğriliği hariç tüm diğer parametreler daha sonra 0-1 aralığında lineer olarak normalize edilir.

2.5.1.2. Chernoff Yüzleri'nin Geliştirilmesi

Herman Chernoff, 1973'te Chernoff Yüzleri ile bilgi görselleştirme fikrini ortaya koyduktan sonraki zamanlarda, araştırmacılar bu yüzleri benimsemiş ve geliştirmişlerdir. Bu araştırmacılar, orijinal Chernoff Yüzleri'nin bazı eksikliklerine dikkat çekerek ve geliştirilmiş yüzleriyle bu sorunları gidermeye çalışmıştır.

Orijinal Chernoff Yüzleri'nde yaygın olarak bilinen bir sorun, eğer belirli parametreler aşırı değerlere yakınsa, diğer tüm parametreleri de değiştiriyor olmasıdır. Diğer parametreler hala aynı olabilir, ancak Chernoff Yüzleri şeklindeki grafik

gösterimleri yine de değişecektir. Bu problem burun boyu için aşırı değerlerle en kolay şekilde gösterilebilir. Sadece burun uzunluğu için değer değiştirilirken, Chernoff Yüzleri'nin diğer parametreleri de bu değerden etkilenir. Burun uzunluğundaki bu örnek, bu eksikliğin en belirgin etkisine sahip örnek olabilir, ancak sorun diğer parametrelerde de mevcuttur.

Chernoff Yüzleri hakkındaki eleştirilerden biri de yüzlerin çizgi film(karikatür) benzeri görünümler olmasıydı. Herman Chernoff tarafından icat edilen yüzler, insan yüzlerine pek benzemiyordu. Bu yüzden bazı araştırmacılar, yüzlere daha insancıl bir görünüm vermeye çalıştılar [Flury ve Riedwyl, 1981] [Kabulov ve Tashpulatova, 2010]. Diğer bir geliştirme durumu ise bazı bilim adamlarının gelişmiş Chernoff Yüzleri'nde daha fazla parametre olmasını istemesiydi. Chernoff Yüzleri'nin sağ ve sol kısımları simetriktir. Bu nedenle, bir gözlemci tüm parametreleri görmek için yüzün sadece bir yarısına bakmalıdır. Bu gerçeğe bağlı olarak, toplam alanın yarısı boşa harcanır, çünkü bir yarısı diğer yarının yansıtılmış bir kopyasıdır. Flury ve Riedwyl, "Asimetrik Yüzler Aracılığıyla Çok Değişkenli Verilerin Grafiksel Gösterimi" başlıklı makalelerinde "Yeni Yüzler" i icat etti [Flury ve Riedwyl, 1981]. Orijinal simetrik Chernoff Yüzleri'nde çifte değer probleminden kaçınmak için asimetrik yüzler kullandılar. Yüzlerin asimetrik versiyonu, sonuç yüzüne dahil edilen 36 farklı veri noktasına izin verir.

2.6. Dairesel İkon Grafikleri

Açıklayıcı veri analizinin potansiyel olarak güçlü tekniklerinden biri çok boyutlu ikon grafikleridir. İkon çizimlerinin temel fikri, bireysel gözlem birimlerini, değişkenlerin değerlerinin, nesnelerin belirli özelliklerine veya boyutlarına atandığı belirli grafiksel nesneler olarak göstermektir. Atama, nesnelerin genel görünümünün, değerlerin yapılandırılmasının bir fonksiyonu olarak değişeceği şekildedir. Böylece nesnelere, değerlerin konfigürasyonları için benzersiz olan ve gözlemci tarafından tanımlanabilecek görsel "kimlikler" verilir. Bu tür ikonların incelenmesi, hem basit ilişkilerin hem de değişkenler arasındaki etkileşimlerin belirli kümelerini keşfetmeye yardımcı olabilir.

Dairesel ikon grafikleri her ne kadar isminde olduğu gibi daireselmiş gibi algılansa da sadece merkezleri bir olup çıkan kenar uzunlukları kimi zaman eşit olmayabilmektedir. Kimi zaman bir tekerlek gibi kimi zaman da çoklu yolların şehrin merkezi olan bir bulvarda kesişmesi gibi yol uzunlukları farklılık göstermektedir. Bu grafiklerde değişkenlerin değerleri, ikonun merkezi ile kenarları arasındaki uzaklıkla ifade edilmektedir.

Bu grafikler, değişkenler arasındaki etkileşimli ilişkileri belirlemede oldukça yararlı olabilirler. Yararlı olabileceği nokta, ilgilenilen değişkenlerin değer bileşimlerine göre çok farklı ve belirli bir genel ikon şekli verebilmesidir. Bu yüzden bu grafik çeşitleri, aşırı değerleri bulmada ve kümeleme analizi sonuçlarını görsel olarak doğrulamada genellikle kullanılmaktadır. İkonların ilgili değişkenlerin değer aralıkları arasındaki genel farklılıkları değişken birimlerinden bağımsız olarak yansıtabilmesi ve ikon içinde değişkenlerin değer aralığının birbirleriyle uyumlu olabilmesi için değişkenleri standartlaştırılması uygun olacaktır. Veride değer aralığı diğerlerine göre büyük olan bir değişken bulunması durumunda, değer aralığı küçük olan değişken veya değişkenlere ilişkin gösterimler küçülecek ve bazen bu tür değişkenler grafikte görülemeyecektir. Değişkenlerin nasıl standartlaştırılacağı ve aşırı değerler nasıl saptanıp ve bununla ilgili neler yapılacağı hakkında genel bilgiler için Bölüm 1'e bakınız.

İkon grafiklerinin en büyük dezavantajlarından birisi, gözlem sayısının çok fazla olması halinde yorumlamada güçlüklerin yaşanabilmesidir. Bu yüzden, buna benzer grafikler gözlem sayısının çok fazla olmaması yani orta büyüklüklerde olması durumunda tercih edilmektedir. Yani değişken sayısının 20'den az olmamasıyla birlikte en fazla 100 değişkene kadar çıkabilir. Bu değerden fazla olanlar için doğruların üst üste gelme durumu ve renklerdeki yakınlık derecesine göre karışma durumu oluşacaktır. Bu da hem yorumlamanın güçlüğüne hem de karışıklığa sebep olacaktır. En fazla kullanılan üç çeşit dairesel ikon grafiği vardır. Bunlar yıldız grafikleri, poligon grafikleri ve güneş ışığı grafikleridir.

2.6.1. Yıldız Grafikler

Yıldız Grafiği, çok değişkenli verilerin görselleştirilmesi için kullanılan popüler yöntemlerden biridir. Basit ve nispeten sezgisel bir geometrik şekil olan yıldız grafikleri, köşeleri tümü aynı kökene sahip olan bir eksenler topluluğuyla tanımlanan yıldız biçimli bir çokgenle temsil edilir. İlgilenilen değişkenlerin göreli büyüklükleri her bir gözlem için merkezden yıldız köşelerine olan göreli uzunluk ile temsil edilir. Başka bir deyişle, her bir doğrunun uzunluğu, ilgili değişkenin büyüklüğüyle doğru orantılı çizilir. Her bir köşeye karşılık gelen noktaların birleştirilmesiyle, şekli küresel olarak veri kümesinin örneğini tanımlayan geometrik bir şekil elde edilir. Profil ve Andrews çizimlerinde olduğu gibi, her küme için değişkenleri temsil eden değerler standartlaştırılmış değişkenlerin ortalaması olacaktır. Daha sonra çizgilerin bitiş noktaları birbiriyle birleştirilir. Eğer veride beş değişken var ise, her bir gözlem beş köşeli bir yıldız ile tanımlanır, veride 8 değişken var ise, her bir gözlem 8 köşeli bir yıldız ile tanımlanır; ancak verinin yapısına bağlı olarak yıldız şeklindeki bir sembole ulaşılamayabilir. Bir yıldızdaki değişkenler, 12'den başlamak üzere sırasıyla saat yönünde yer alır. Bu grafik yardımıyla, bireylerin genel durumlarında daha etkili olan değişkenler daha kolay bir şekilde saptanabilmektedir. Bu yöntem, negatif olmayan ölçümler için uygundur. Boyutların sayısı ve veri kümesinin boyutu arttıkça, görselleştirmesi kısa sürede çok fazla dağınık hale gelir, çünkü birçok ışının küçük dairesel alan içinde kalması gerekir ve ayrı ayrı simgeler de çok küçük olur.

Yıldız grafikleri çok değişkenli gözlemlerin keyfi sayıda değişken ile görüntülenebilmesi için yararlı bir yöntemdir. Her gözlem, her değişken için bir ışın bulunan yıldız şeklinde bir şekil olarak temsil edilir. Belirli bir gözlem için, her ışının uzunluğu bu değişkene göre orantılı olarak yapılır. Yıldız grafiklerindeki her bir doğru arasındaki açı $360°/k$'dir, burada k çizilen eksen sayısıdır. Gözlemler arasındaki en büyük görsel ayırımı sağlamak için ilgili özellikleri ölçen ışınlar birlikte gruplandırılmalıdır. Bu sayede aşırı gözlemler, diğer verilerden oldukça farklı görünen bir yıldız olarak göze çarpacaktır.

Yıldız grafikleri şu durumları göstermektedir: Belirli bir gözlem için hangi değişkenin baskın olduğunu, hangi gözlemler en çok benzerdir yani gözlem kümeleri var mıdır? Aşırı değerler var mıdır? Bu grafik sayesinde bu durumlar gözlemlenebilir.

Yıldız grafiğinin temel zayıflığı, analizin en fazla 100 gözlem değeriyle sınırlı olmasıdır. Yıldız grafiklerine alternatifler çubuk grafikler ve paralel koordinat çizelgeleridir.

2.6.2. Poligon İkon Grafikleri

Şekil olarak yıldız grafiğine benzeyen bu grafik türünde, her bir gözlem bir poligon (çokgen) şekliyle gösterilmektedir. Analiz edilecek olan değişkenlerin relatif büyüklükleri her bir gözlem için merkezden poligon köşelerine olan relatif uzaklık ile ifade edilir. Poligon grafiklerinin yıldız grafiklerinden tek bir farkı bulunmaktadır. Bu fark, poligonun içerisindeki çizgilerin olmayışıdır. Bu tür bir simge grafiğinde, her durum için ayrı bir çokgen simgesi çizilir; her bir durum için seçilen değişkenlerin göreceli değerleri, simgenin merkezinden poligonun ardışık köşelerine kadar olan mesafe ile temsil edilir (saat yönünde, saat 12: 00'de başlar).

2.6.3. Güneş Işığı Grafikleri

Güneş ışığı grafiklerinde, her bir gözlem için güneşe benzer semboller seçilir. Merkezden çıkan her doğru bir değişkeni temsil etmektedir. Her bir gözleme ilişkin değişken değerlerinin konumu güneşin merkezinden başlamak üzere belirlenir. Güneş ışığı grafiğinin, yıldız ikon grafiğinden tek farkı, yıldızların belirlenen ortak bir güneş ışığı şablonu üzerine yerleştirilmesidir. Bu tür bir simge grafiğinde, her durum için ayrı bir güneş benzeri simge çizilir. Her bir ışın seçilen değişkenlerden birini temsil eder ve ışınının uzunluğu 4 standart sapmayı temsil eder. Her durum için değişkenlerin veri değerleri bir satıra bağlanır. (saat yönünde, 12: 00'da başlar)

2.7. Adımsal İkon Grafikleri

Adımsal ikon grafikleri, dairesel ikon grafiklerinden farklı olarak bir daire üzerinde çizilmezler. Adımsal ikon grafikleri adı altında sıklıkla kullanılan grafik türü profil grafikleridir. Bu yaklaşımda, tüm değişkenler dikkate alınarak her bir gözlemin profilini gösteren doğrular çizilir. Bir profil grafiği, çok değişkenli bir veri setindeki tüm değişkenlerin nispi davranışlarını incelemek için kullanılan bir tekniktir. Profil grafikleri, her değişken için, tek bir grup için veya birden fazla grup boyunca örnekleme araçları çizilerek oluşturulabilir. Bir profil grafiği oluşturmanın temel amaçlarından biri, profillerin paralel olup olmadığını değerlendirmek için keşiftir. Bu sayede, her bir

değişken yatay eksende yer alırken, her bir gözleme ilişkin değişken değerleri ise y ekseninde yer alır. Sonuç olarak, her bir gözlem için bir profil çizilmiş olacak. Verilerdeki değişkenlerin net ve anlamlı bir grafik özeti için, değişkenlerin hepsinin çizimden önce aynı ölçüm birimlerine sahip olması gerekir. Örneğin, ağırlıkla ilgili değişkenler farklı birimlerde (gram, kilogram ve pound) ölçülürse, aynı ölçüm skalasına yerleştirilmeli veya standart bir skora (Z skorları) dönüştürülmelidirler. Burada değişkenlerin birimleri farklı olduğunda verilerin standartlaştırılması daha uygun olacaktır. Birbirlerine çok yakın olan çizgiler, benzer olan (standartlaştırılmış profil oluşturma değişkenlerinin ortalama değerlerine göre) kümeleri gösterir ve birbirinden uzak olan çizgiler, benzer olmayan kümeleri gösterir.

Bu tür grafiklerin olumsuz yönlerinden biri, gözlem sayısının fazla olması durumunda yorumlamada karşılaşılan güçlüklerdir. Profil grafikleri, satır sayısı (kümelenme sayısı) arttıkça ve değişken sayısı arttıkça takip edilmesi daha zorlaşır, ancak orta sayıda değişken ve küme için yararlı bir grafik gösterimi sağlayabilirler. Aynı veriye ilişkin seçenek bir sunum ise doğrular yerine çubukların kullanılması ile çizilen çubuk-profil grafikleridir.

2.8. Andrews Grafikleri

Andrews eğrileri, çok boyutlu verileri görselleştirmek için her bir gözlemi bir fonksiyon üzerine eşleyerek kullanılan bir yöntemdir. Andrews eğrileri, her bir profillemenin değişkenlerini Fourier serileri kullanarak, her bir kümenin 2 boyutlu bir eğri olarak gösterilmesini sağlayan bir şekilde birleştirir. Fourier serisi, matematikçi Joseph Fourier (1768 - 1830) tarafından araştırılan, bilim ve mühendislik boyunca geniş bir uygulama yelpazesine sahip olan sonsuz trigonometrik serinin bir türüdür. Sonsuz Fourier serileri sinüs ve kosinüs fonksiyonları ile katsayılar çarpılarak elde edilir. Profil grafiklerinde olduğu gibi, her bir küme için değişkenleri temsil eden değerler, standartlaştırılmış değişkenlerin ortalaması olacaktır. Her küme için bir Andrews eğrisi oluşturulur ve farklı kümeleri temsil eden aynı grafikte renkli eğriler çizilir. Birbirine çok yakın olan eğriler, benzer olan kümeleri gösterir ve birbirinden uzaktaki eğriler, benzer olmayan kümeleri gösterir. Profil grafikleri olarak yorumlanmaları kolay olmamasına rağmen, profilleme için kullanılan değişkenlere göre benzer olan bölümleri belirlemede yararlı olabilirler. Andrews grafiği, yatay eksen boyunca t değerleri için

dikey eksende çizilen fonksiyon f(t) değerlerine sahip bir grafiktir.

Profil grafiklerinde olduğu gibi, satır sayısı (kümelerin sayısı) arttıkça Andrews eğrilerinin takip edilmesi güçleşir. Buna ek olarak, eğrilerin kendileri, bir değişkenin büyüklüğünün, bu değişkeni temsil eden dikey çizgideki yükseklikleriyle kümeler arasında karşılaştırılabildiği profiller olarak yorumlanmaları kadar kolay değildir. Andrews eğrilerinin araçları, mesafeyi ve varyansları belirli bir sabite kadar koruyabildiği gösterilmiştir. Bu, birbirine yakın fonksiyonlarla temsil edilen Andrews eğrilerinin, karşılık gelen veri noktalarının da birbirine yakın olacağı anlamına gelir.

Andrews eğrileri çok boyutlu verileri görselleştirmek için, bir yöntem olarak 1972 yılında Andrews tarafından geliştirilmiştir. Adrews eğrilerinde gözlem değerleri aşağıdaki Eşitlik 2.1'deki fonksiyon kalıbı kullanılarak dönüştürülürler. Dönüşen bu değerlerin daha sonar çizgi grafikleri çizilerek Andrews eğrilerine ulaşılır.

$$f_x(t) = x_1/\sqrt{2} + x_2\sin(t) + x_3\cos(t) + x_4\sin(2t) + x_5\cos(2t) + \ldots,$$
$$-\pi < t < \pi \qquad (2.1)$$

Burada; $x_1, x_2, \ldots, x_n$ verilerimizin değişkenleridir.

Andrews eğrileri, orijinal veri setinin ortalamasını ve varyansını içerisinde barındırırlar. Andrews eğrilerinin kullandığı fonksiyon kalıplarından elde edilen değerlerin birbirine yakın olması gözlem değerlerinin birbirine yakın olduğunu, birbirine uzak olması gözlem değerlerinin de birbirine uzak olduğunu gösterir. Buradan hareketle Andrews eğrileri verilerin küme yapılarının anlaşılmasında da, aşırı değerlerin tespitinde de kullanılabilirler (Martinez ve Martinez, 2005).

2.8.1. Andrews Eğrilerinin Özellikleri:

Andrews eğrilerinin çeşitli kullanışlı özellikleri bulunmaktadır. Andrews eğrilerinin bazı özellikleri aşağıdaki gibidir: (Garcia-Osorio ve Fyfe., 2005)

1. Andrews eğrileri veri setimizin ortalamasını içinde barındırmaktadır. Eğer $\bar{x}$ vektörü n çok değişkenli gözlemlerin ortalamasını temsil ediyorsa,

$$f_{\bar{x}}(t) = \frac{1}{N}\sum_{i=1}^{N} f_{x_i}(t)$$

2. Andrews eğrilerinde , $f_x(t)$ve $f_y(t)$, iki fonksiyon arasındaki uzaklık

$$\|f_x(t) - f_y(t)\|L_2 = \int_{-\pi}^{\pi} [f_x(t) - f_y(t)\,]^2 dt$$

olmak üzere; x_i, y_i noktaları arasındaki uzaklık,

$$\|f_x(t) - f_y(t)\|L_2 = \pi \sum_{i=1}^{d} (x_i - y_i)^2 = \pi\|\mathbf{x} - \mathbf{y}\|^2$$

karşılık gelen noktalar arasındaki Öklid Uzaklığı ile orantılıdır.

3. Doğrusal ilişki özelliğidir. Eğer bir y noktası, x ile z' yi birleştiren bir çizgi arasında kalıyorsa bütün t' ler için, $f_y(t)$, $f_x(t)$ ile $f_z(t)$ fonksiyonları arasında kalır.

Andrews eğrilerine getirilen en büyük eleştirilerden bir tanesi de çizilen şeklin biçiminin değişkenlerin sıralarına olan bağımlılığıdır. Yani, değişken sıraları değiştikçe Andrews eğrilerinin şekilleri değişecektir. Andrews eğrilerinde dönüştürme işlemi için kullanılan Eşitlik 2.1'deki seride ilk sıraya yerleşen değişken, grafiğimizin üzerinde en büyük ağırlığa sahip olan değişkendir. Yani değişkenlerin sıraları değiştikçe, ilk sırada olan değişkenin, grafiğin ağırlığına olan etkisi fazla olacak şekilde grafiğimizin biçimi değişecektir. Sonuç olarak, değişken sayısının permütasyonu kadar farklı sayıda grafik çizilecektir. Temel bileşenler analiziyle boyut indirgemesi yapılarak Andrews eğrilerinin bu dezavantajının üstesinden gelinebilir. Bu sayede en büyük varyans açıklayıcılık oranına sahip değişken, şeklimizde en büyük etkiye sahip olacaktır. Aykırı değerler, verilerin analiz aşamasında regresyon, kümeleme analizi gibi uygulamalarda sorunlara neden olurlar. Bu nedenle aykırı değerlerin veri setinden arındırılması gerekmektedir.

Veri setinde gerek değişken sayısının fazla olması, gerekse veri birimlerinin fazla olması durumunda aykırı değerleri ayıklamak oldukça zor bir iştir. Bunun için görselleştirme teknikleri kullanılarak aykırı değerler görsel bir şekilde tespit edilip, veri setinden ayıklanabilir. Unutulmamalıdır ki; veri görselleştirme teknikleri insan algılama yeteneklerini ve insanlar arası yorumlama farklılıklarını dikkate alarak analiz

gerçekleştirilmesine olanak sağlar. Görselleştirme teknikleri ile diğer yöntemlerle fark edilmesi daha zor olan bilgiye erişilmesi ve bilginin yorumlanması kolaylaşmaktadır. Ancak grafiksel tekniklerin matematiksel sonuçlar vermemesi gibi bir dezavantajı da bulunmaktadır (Kümeleme Analizi ve Andrews Analizi). Andrews eğrileri oluşturulurken değişkenlerin benzer birimlerde olması istenir. Eğer değişkenler farklı birimlerde ise, verilerin standartlaştırılması gerekecektir. Ayrıca elde edilen grafiklerin yorumlanması, değişkenlerin fonksiyona giriş sırasından etkilenmektedir. Bazı değişkenlerin diğerlerinden daha önemli olduğu düşünüldüğünde, en önemli değişkenin x_1, bir sonrakinin x_2 ve benzeri şekilde alınması daha uygundur.

2.8.2. Andrews Eğrileri İçin Varyasyonlar

Andrews'in kendisi de dahil olmak üzere birçok yazar, Andrews eğrileri için kullanılan formülün vektörlerine varyasyonlar önermiştir. Bu çalışmada, Andrews eğrileri için üç farklı alternatif kullanılmıştır. Bu alternatifler aşağıdaki gibi sıralanmıştır. Denklem (i) orijinal Andrew formülüdür ve diğer 3 formül de alternatif olarak sunulmuştur.

(i) $f_y^{(1)}(t) = y_1 sin(n_1 t) + y_2 cos(n_2 t) + y_3 sin(n_3 t) + y_4 cos(n_4 t) + \cdots,$
$-\pi \leq t \leq \pi$ (Andrews, 1972),

burada n_i birbirinden farklı tamsayılardır,

(ii) $f_y^{(2)}(t) = y_1 sin(2t) + y_2 cos(2t) + y_3 sin(4t) + y_4 cos(4t) + \cdots,$
$-\pi \leq t \leq$ (Andrews, 1972),

(iii) $f_y^{(3)}(t) = y_1 cos(t) + y_2 cos\left(\sqrt{2t}\right) + y_3 cos\left(\sqrt{3t}\right) + y_4 cos\left(\sqrt{4t}\right) + \cdots,$
$0 \leq t \leq \pi$ (Attributed toTukey by Gnanadesikan, 1997),

(iv) $f_y^{(4)}(t) = y_1(\sin(t) + y_2 \cos(t) + y_3 sin(2t) + y_4 cos(2t) + \cdots,$

$-\pi \leq t \leq \pi$ (Kulkarni and Paranjape, 1984).

2.8.3. Andrews Eğrilerine Bir Alternatif

Bir önceki başlıkta Andrews eğrileri için varyasyonlar listelenmişti. Daha açıklayıcı ve bilgilendirici hale getirme çabasıyla, burada uygun şekilde değiştirerek,

Andrews eğrilerine bir alternatif sunulmuştur. Andrews fonksiyonlarında değişkenler y_j, trigonometrik fonksiyonların katsayıları olarak kullanıldığından, verilerin istatistiksel varyasyonu, sinüs ve kosinüs dalgalarının periyodik varyasyonu ile karıştırılarak, yorumlanmaları bazen grafikleri zorlaştırır. Bu durum, çok değişkenli veriler için değişkenlerin ölçeklendirilmesi (veya eksikliği) ile ilgili olanlar, ilişkili değişkenler, eşit olmayan varyanslar gibi birkaç ek karmaşık sorun olabileceğinden, bu çok önemli bir konudur. Trigonometrik fonksiyonlar doğal bir seçim oluşturduğundan, tamamen atılamazlar. Böylece, her y_j 'nin bir sinüs terimine ve aynı zamanda bir kosinüs terimine bir katsayı olarak atanması için işlevler aranabilir. Böyle bir işlevin, Andrews eğrilerine atfedilen tüm istenen özellikleri muhafaza etmesi arzu edilir. Bu yüzden alternatif olarak aşağıdaki fonkiyon verilmiştir:

$$g_y(t) = \{y_1 + y_2(sin(t) + cos(t)) + y_3(sin(t) - cos(t)) + y_4(sin(2t) + cos(2t)) + \cdots\},$$
$$-\pi \leq t \leq \pi$$

2.9. Biplot

İki değişkenin olduğu durumlarda, *n* gözlemden oluşan bir veri kümesi için, iki boyutlu bir nokta saçılım grafiğinde bu durum gayet rahat bir şekilde yorumlanabilir; fakat ikiden fazla değişken olduğu durumlarda, çok değişkenli bir verinin görselleştirilmesi oldukça güç olacaktır. Temel bileşenler yardımıyla nxp matrisi iki boyuta indirgenebiliyorsa, elde edilecek faktör skorları yardımıyla çizilecek bir saçılım grafiği iki yeni değişkene ilişkin bilgiyi görsel olarak verebilecektir. Biplotlar, istatistiklerde kullanılan, basit iki değişkenli saçılım grafiğinin genelleştirmesi olan bir keşif grafiği türüdür. Biplot çizimlerinde bu sürece değişkenler, vektörler olarak katılır. Biplot, çok değişkenli verinin daha az boyutta (iki veya üç) yorumlanması için kullanılan grafiksel bir tekniktir. Ancak boyut indirgemesi yapılırken genellikle bilgi kaybı oluşmaktadır. Biplotların asıl amacı, bilgideki bu kaybın minimizasyonu sağlayıp bazı kriterleri optimize etmektir. Optimize edilen kriterlere bağlı olarak, biplotların çeşitli tipleri bulunabilir. Farklı kriterlerin oluşmasının sebebi, genellikle farklı uzaklık ölçülerine bağlıdır. Değişkenler, normal saçılım grafiklerinde olduğu gibi ölçeklerle doğrusal eksenle temsil edilir. İşaretleyiciler, değişkenleri (sütunlar) için işaretlerini

eksenlere dik olarak yansıtarak ve ölçek üzerindeki değeri okuyarak bulunur. Bu yansıtılan ölçek değerleri, gerçek düzlemde ikiden fazla değişkeni temsil etmesi genellikle mümkün olmadığından, gerçek değerlerin yaklaşık değerleridir.

Biplot, veri kümesinde değişkenler arasındaki bağlantıların, değişkenlere göre gözlemler arasındaki benzerliklerin veya farklılıkların görsel olarak yorumlanabilmesinde ve gözlemlerin sınıflandırılmasında kullanılan grafiksel bir tekniktir. Bu grafikler hem gözlemleri hem de değişkenleri dikkate aldığından dolayı biplot grafikler olarak isimlendirilir. Önceden belirttiğimiz gibi sadece iki değişken olduğunda bir saçılım grafiği hem gözlemlerdeki hem de değişkenlerdeki bilgiyi göstermekteydi. Yani, bir gözlemin grafik üzerindeki yerleşim yeri diğerine göre görülebilmektedir. Ayrıca, her bir gözlemin grafik üzerindeki yerine göre iki değişkenin her birinin göreceli önemi de belirlenebilmektedir. Bu teknik görsel bir grafik yardımıyla veriyi tüm ayrıntılarıyla incelemeye olanak sağlamaktadır.

Saçılım grafiklerine zıt olarak, eksenler dikey değildir; çünkü n-boyutlu değişkenler bir gösterimin bir yüzey üzerinde minimum bilgi kaybına sahip gösterimini gerçekmiş gibi sümile ederler. Bu nedenle, vektörler dik olduğunda yani 90 derece olduğunda kosinüs sıfır(0) olacaktır ve bu durumda değişkenler bağımsızdır. Fakat eğer çok yakınlarsa yani iki vektör arasındaki açı küçük(dar) ise iki değişkenin yüksek korelasyonlu olduğunu göstermektedir ya da iki değişken arasındaki açının 90 derecenin üstünde olması(geniş açılı olması) değişkenlerin negatif ilişkili(düşük korelasyonlu) olduğunu belirtir. Biplot grafiklerinde, vektörler arasındaki açının kosinüsü korelasyon büyüklüğünü ifade etmektedir Dahası aradaki açılar 90 dereceyi geçmemek üzere birbirine yakın olan iki vektör uzak olana göre daha yüksek ilişkilidir. r=-1'e yaklaştığında açı 180 dereceye yaklaşır. r=1'e yaklaştıkça ise açı sıfır(0) dereceye yaklaşır. Yani doğrular üst üste çakışacaktır. r sıfıra yaklaştıkça açı 90 dereceye yaklaşır.

Bir Biplot gösterimini yorumlarken bir diğer önemli husus, eksenlerin gösterimi ile ilgilidir. Normalde, bunlar tüm değişkenlerin araçlarının işareti olan ağırlık merkezinde buluşur. Ayrıca, değişkenlerin standart sapmalarının yaklaşık değerini gösterdiği için vektörlerin (değişkenlerin) uzunluğu önemlidir. Biplot grafiğinde verilerin farklı birimlerde olması halinde, sonucun grafiği ciddi şekilde etkileyeceği söz

konusu olacak. Bu yüzden verilerin çizim öncesinde standartlaştırılması uygun olacaktır. Birimlerin aynı olduğu, standart sapmaların ise farklı olduğu durumlarda, bunun etkileri vektörlere de yansımakta, küçük varyanslı değişkenin vektör boyu büyük varyanslı değişkene göre daha kısa olmaktadır.

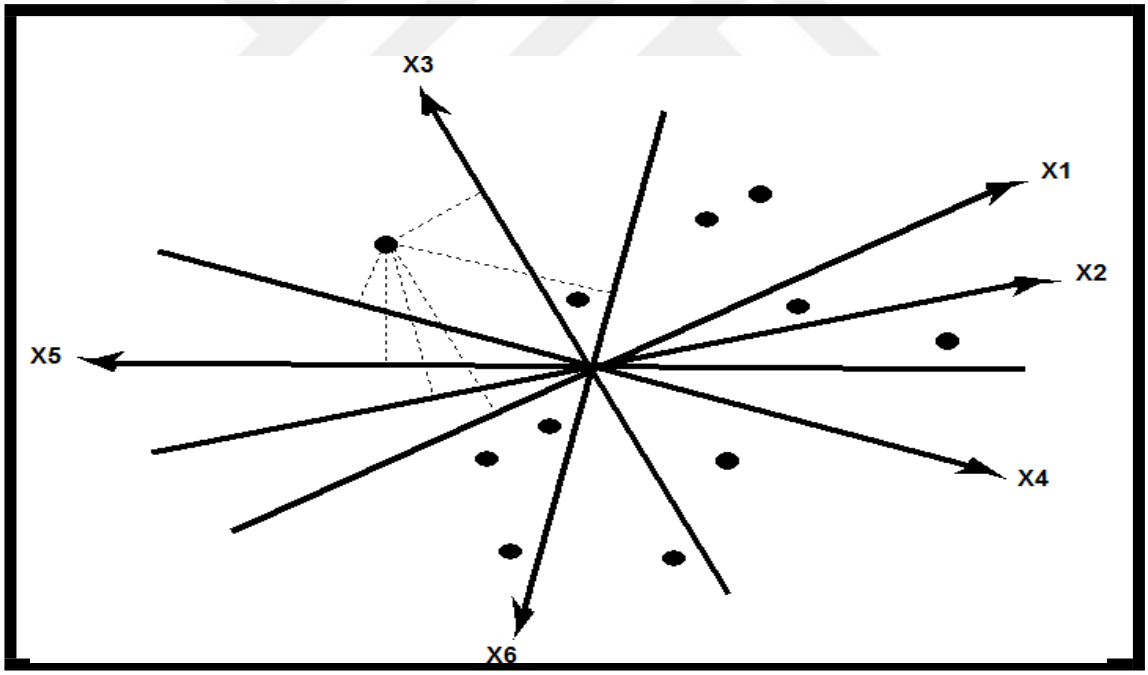
Kısaca, Biplot analizi değişkenleri ve vakaları birleştiren çok değişkenli verilerin grafiksel bir gösterimidir, ilk önce Gabriel (1971) tarafından önerilmiş ve birçok farklı tür ve türünde çok farklı bilimsel alanlarda test edilmiştir: Tıp (Gabriel, 1990), Genetik (Wouters ve diğerleri, 2003), Tarım (Yan ve diğerleri, 2000), Kütüphane Bilimi (Veiga de Cabo ve Martín-Rodero, 2011), Ekonomi ve İşletme (Galindo, Vaz & Nijkamp, 2011), Turizm (Pan, Chon & Song, 2008) veya Siyaset Bilimi (Alcántara ve Rivas, 2007). Kaynakça alanında, bu metodoloji ilk olarak küçük bir İspanyol üniversitesinin Sağlık Bilimlerindeki bilimsel aktiviteyi analiz etmek için Biplot analizinin uygulandığı konferans raporunda tanıtılmıştır (Arias Díaz-Faes ve diğerleri, 2011).

Biplot, iki değişkenli verilerin analizinde kullanılan nokta saçılım grafiklerinin çok değişkenli eşleri olarak göz önüne alınabilir. Biplot tekniği, çok değişkenli bir veri kümesinin ayrıntılı bir şekilde özetlenmesi yanında değişkenler arasındaki ilişkileri belirleme ve birimlerin sınıflandırılmasında oldukça başarılı bir tekniktir. Bu teknik görsel bir grafik yardımıyla veriyi tüm yönleri ile incelemeye imkan sağlamaktadır. Biplot sadece grafiksel bir teknik değil, aynı zamanda ayrıntılı istatistiksel analizler için gerekli bir yöntem ile desteklenmektedir. Biplot da ki "Bi" ifadesi grafik boyutunu değil birimlerin ve değişkenlerin aynı grafikte gösterileceğini ifade etmektedir. Bu teknik, tekil değer ayrıştırması prensibine dayanmaktadır.

Biplotlar simetrik ve asimetrik olmak üzere iki ana türde incelenebilir. Simetrik biplotlar, iki yönlü bir tablonun satırları ve sütunlarıyla ilgili bilgi veren yaklaşımlar olarak tanımlanırken, asimetrik biplotlar ise bir veri matrisinin değişkenleri ve gözlem birimleri üzerine bilgi veren yaklaşımlar olarak tanımlanır. Simetrik biplotlarda, satırlar ve sütunlar bilgi kaybı olmaksızın yer değiştirebilirken, asimetrik biplotlarda obje türü farklı olduğundan böyle bir yer değişikliği mümkün değildir. Simetrik biplotlarda, hem gözlemler hem de değişkenler noktalar ile temsil edilirken asimetrik biplotlarda, gözlemler noktalar ile değişkenler ise vektörler ile temsil edilmektedir. Hangi biplotun

seçileceğine veri kümesinde yer alan değişken türlerine (nicel, nitel, sıralı vb.) bakılarak karar verilir (Gower vd. 2011).

İki değişkenli nokta saçılım grafiği genel literatürdeki ifadesiyle x yatay eksen ve y dikey eksen olmak üzere iki eksene sahiptir. Biplotlar ise, değişken sayısı kadar eksene sahiptir. Bir nokta saçılım grafiğinde gözlemler noktalar ile gösterilir ve iki değişken üzerinde gözlemlerin dik izdüşümleri alındığında, gözlemlerin ilgili değişkenler üzerindeki değerleri elde edilir. Benzer olarak bir biplot'da, tüm değişkenler üzerinde herhangi bir gözlemin dik izdüşümü, ilgili gözlemin tüm değişkenler üzerinde aldığı değerleri verir. Nokta saçılım grafiklerinde ise gözlemlerin değişkenler üzerinde aldığı değerlere kesin ulaşılırken biplot'da bu mümkün değildir. Biplotlar indirgenmiş boyutlu bir uzayda (genellikle 2 veya 3 boyutlu) gösterilir. Biplot tekniği, indirgenmiş boyutlu bir uzayda değişkenler arasındaki korelasyonlardan yararlanır.



Şekil 2.2. Biplot Grafiği

Biplot grafiğini STATA'da oluşturabilmek için STATA>Statistics>Multivariate Analysis>Biplot adımları takip edilerek açılan pencereye değişkenler ''Variables'' kutusuna atanır. Aynı pencere üzerinde ''negate the data to the axis'' seçeneği seçilerek veriler x eksenine göre negatifleştirilip grafik ters hale dönüştürülebilir. Ayrıca bunlar için STATA komut satırına,

''biplot X1 X2 X3 X4'' yazılarak da elde edilebilir. Tersi işlem için,

''biplot X1 X2 X3 X4'' xnegate komutu ile elde edilebilir.

2.9.1. Gabriel'in Biplot Yaklaşımı

Sıra sayısı k olan herhangi bir n× p boyutlu $\mathbf{X}^3$ veri matrisinin, her bir gözlem için bir nokta(satır) ve her bir değişken için bir vektör(sütun) olmak üzere minimum bilgi kaybına sahip düşük boyutlu grafiksel bir yaklaşım ile gösterilir. Gabriel'in biplot yaklaşımı tekil değer ayrıştırmasına dayanmaktadır. **X:** n× p matrisinin tekil değer ayrıştırması aşağıdaki gibidir:

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times k}\ \mathbf{\Gamma}_{k\times k}\left(\mathbf{V}_{p\times k}\right)^{T} \qquad (2.2)$$

Burada, k, **X** matrisinin sıra sayısını göstermektedir. **Γ** , köşegenleri **X'**in sıfır olmayan $0 < \gamma_1 \leq \gamma_2 \leq \ldots \leq \gamma_k$ tekil değerlerinden oluşan k×k tipindeki köşegen bir matristir ve **Γ** = diag$\{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ şeklinde gösterilmektedir. **Γ** , $\mathbf{X}^T\mathbf{X}$ veya $\mathbf{X}\mathbf{X}^T$ matrislerinin öz değerlerinin karakökleri alınarak oluşturulan köşegen matristir. **X** matrisinin karesel ve simetrik bir matris olmaması durumunda özdeğer ayrıştırması yerine tekil değer ayrıştırması tercih edilir. Özdeğerler matrisi ise köşegenleri azalan sırada özdeğerlerden oluşan, **Λ** = diag$\{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ köşegen matristir. $\mathbf{X}\mathbf{X}^T$ simetrik matrisinin birimleştirilmiş öz vektörlerine **U** tekil vektörler matrisi karşılık gelirken, $\mathbf{X}^T\mathbf{X}$ simetrik matrisinin birimleştirilmiş özvektörlerine **V** tekil vektörler matrisi karşılık gelir. Eşitlik 2.2'de **U** ve **V** tekil vektörler matrislerinin ilk k sütunu alınmıştır. **X** matrisine k boyutta yaklaşım aşağıdaki şekilde olacaktır:

$$\tilde{\mathbf{X}}^{(r)}_{n\times p} = \mathbf{U}_{n\times r}\ \mathbf{\Gamma}_{r\times r}\left(\mathbf{V}_{p\times r}\right)^{T} \qquad (2.3)$$

X matrisinin bu faktörleştirilmesi, değişkenler arasındaki korelasyon, değişkenlerin varyansları ve gözlemler arasındaki farklılıkların bir sunumunu gösterir. Eğer **X**'e r sıra sayısı ile yaklaşım yeterli oluyorsa, çok değişkenli verinin yorumlanması için kullanışlı bir grafiksel araç elde edilmiş olur.

Eckart-Young teoremi, r ranklı $\widetilde{\mathbf{X}}^{(r)}$ matrisi ile k ranklı **X** matrisine (r≤k) optimal yaklaşımı matrislerin en küçük kareler yaklaşım yöntemi kullanarak bulur. Yani, burada amaç hata kareler toplamını minimize etmektir (Eckart ve Young 1936).

Buna göre;

$$\min_{\text{rank}(B_{n\times p})=r} \|\mathbf{X}-\mathbf{B}\| = \left\|\mathbf{X}-\widetilde{\mathbf{X}}^{(r)}\right\| = \text{tr}\left\{\left(\mathbf{X}-\widetilde{\mathbf{X}}^{(r)}\right)\left(\mathbf{X}-\widetilde{\mathbf{X}}^{(r)}\right)^{T}\right\}$$

$$= \sqrt{\sum_{j=r+1}^{k}\gamma_j^2} = \sqrt{\sum_{j=r+1}^{k}\lambda_j} \qquad (2.4)$$

eşitliği sağlanır.

Sıra sayısı r (r≤k) olan daha düşük ranklı, $\widetilde{\mathbf{X}}^{(r)}$ matrisi ile **X** matrisine yaklaşacağımızı farzedelim. Bunu yapmak için ilk olarak yaklaşım hatası veya uyum iyiliği ölçüsü kavramları atanmalıdır. Yaklaşım hatasının ölçüsü normalde, $\mathbf{E} = \mathbf{X} - \widetilde{\mathbf{X}}^{(r)}$ hata matrisinin Öklid normu olarak verilir. Bir matrisinin Öklid normunun, matris iç çarpımının izi olarak yazılabileceğini bilerek,

$$\|\mathbf{E}\| = \left\|\mathbf{X}-\widetilde{\mathbf{X}}^{(r)}\right\| = [\text{iz}(\mathbf{E}^{T}\mathbf{E})]^{\frac{1}{2}} = \left(\sum_{i=1}^{n}\sum_{j=1}^{p} e_{ij}^{2}\right)^{\frac{1}{2}} \qquad (2.5)$$

eşitliği yazılabilir (Bartkowiak ve Szustalewicz 1995).

Asıl sıkıntı, Öklid normu kullanıldığında minimum hata ile daha düşük ranklı matrislerle **X** matrisine nasıl yaklaşılacağıdır. Bu sorun ilk olarak Householder ve Young (1938) tarafından ele alınmıştır.

Öyleyse, sıra sayısı r (r≤k) olan bir $\widetilde{\mathbf{X}}^{(r)}$ matrisi ile k ranklı bir **X** matrisine en iyi yaklaşım, $\mathbf{E} = \mathbf{X} - \widetilde{\mathbf{X}}^{(r)}$ hata matrisinin Öklid normunun minimizasyonu ile **X** matrisinin

tekil değer ayrıştırmasının ilk r bileşenin kullanılmasıyla sağlanabilir. Yaklaşım hatası, **X** ve $\widetilde{\mathbf{X}}^{(r)}$ matrislerinin tekil değer ayrıştırması ve Eckart-Young teoremi kullanılarak aşağıdaki şekilde yazılabilir.

$$\|E\| = \min_{\text{rank}(B)=r} \|\mathbf{X}-\mathbf{B}\| = \left\|\mathbf{X}-\widetilde{\mathbf{X}}^{(r)}\right\|$$

$$= \left\| \mathbf{U}_{n\times k}\ \mathbf{\Gamma}_{k\times k}\left(\mathbf{V}_{p\times k}\right)^{T} - \mathbf{U}_{n\times r}\ \mathbf{\Gamma}_{r\times r}\left(\mathbf{V}_{p\times r}\right)^{T} \right\|$$

$$= \left\|\gamma_1 u_1 v_1^T + \gamma_2 u_2 v_2^T + \cdots + \gamma_r u_r v_r^T + \gamma_{r+1} u_{r+1} v_{r+1}^T + \cdots + \gamma_k u_k v_k^T - \gamma_1 u_1 v_1^T - \gamma_2 u_2 v_2^T - \cdots - \gamma_r u_r v_r^T\right\|$$

$$= \left\|\gamma_{r+1} u_{r+1} v_{r+1}^T + \cdots + \gamma_k u_k v_k^T\right\| , \left(\|\mathbf{X}\| = \sqrt{\text{iz}(\mathbf{X}^T\mathbf{X})}\right)$$

$$= \sqrt{\sum_{j=r+1}^{k} \gamma_j^2} = \sqrt{\sum_{j=r+1}^{k} \lambda_j}$$

Eşitlik 2.3, p×p tipindeki bir **J** matrisiyle oluşturularak:

$$\mathbf{J}=\begin{bmatrix} \mathbf{I}_r & 0 : r\times(p-r) \\ 0 : (p-r)\times r & 0 : (p-r)\times(p-r) \end{bmatrix} \qquad (2.6)$$

$$\widetilde{\mathbf{X}}^{(r)} = \mathbf{U\Gamma JV}^T = \mathbf{U\Gamma V}^T = \mathbf{UJ\Gamma JV}^T \qquad (2.7)$$

şeklinde yazılır. **J** için $\mathbf{J}^2 = \mathbf{J}$ ve $(\mathbf{I}-\mathbf{J})^2 = \mathbf{I}-\mathbf{J}$ sağlanır.

UJ ve **VJ** matrisleri, sırasıyla **U** ve **V** matrislerinin ilk r sütunlarının alındığını ifade eder. Bazı durumlarda, **U** ve **V** matrislerinin ilk r sütunlarının gösterimi için $\mathbf{U}_r$ ve $\mathbf{V}_r$ gösteriminin kullanılması daha uygundur.

2.9.2. Veri Matrisinin Ayrıştırılması

X veri matrisinin elemanları, satır ve sütunlara karşılık gelen vektörlerin iç çarpımına eşittir. **U** , **Γ** ve **V** matrisleri, Eşitlik 2.3'te verilen tekil değer ayrıştırması sonucunda elde edilmektedir. **A**: n×k matrisinin satırları ve **B**: k×p

matrisinin sütunları sırasıyla gözlemler ve değişkenler için koordinatları sağlar. Buna göre:

$$\mathbf{X}_{n\times p} = \mathbf{A}_{n\times k}\mathbf{B}_{k\times p} = (\mathbf{U}_{n\times k}\,\mathbf{\Gamma}_{k\times k})\left(\mathbf{V}_{p\times k}\right)^{T} \qquad (2.8)$$

eşitliği yazılabilir. Burada, $\mathbf{A}_{n\times k} = (\mathbf{U}_{n\times k}\ \mathbf{\Gamma}_{k\times k})$, $\mathbf{B}_{k\times p} = \left(\mathbf{V}_{p\times k}\right)^{T}$ olduğu açıkça görülmektedir.

Eşitlik 2.7'de r boyutta bir yaklaşımda,

$$\widetilde{\mathbf{X}}^{(r)} = (\mathbf{UJ\Gamma})(\mathbf{VJ})^{\intercal} = (\mathbf{UJ\Gamma Q})\,)(\mathbf{VJQ})^{\intercal} = \widetilde{\mathbf{A}}^{(r)}\widetilde{\mathbf{B}}^{(r)} \qquad (2.9)$$

şeklinde ifade edilir. Eşitlik 2.9, herhangi bir r×r tipinde ortogonal **Q** matrisi için geçerlidir. (Gower vd. 2011).

2.9.3. Grafiksel Gösterimin Özellikleri

Biplot grafiksel yaklaşımı aşağıda verilen özellikleri sağlamalıdır:

1. Grafiksel yaklaşımda herhangi iki gözlem çifti arasındaki uzaklıklar Öklid uzaklığı: $\|\mathbf{a}_i - \mathbf{a}_{i'}\|^2 = (\mathbf{x}_i - \mathbf{x}_{i'})^T(\mathbf{x}_i - \mathbf{x}_{i'})$, burada $\mathbf{a}_i$, **A**'nın i. satırı ve $\mathbf{x}_i$, **X**'in i. satırıdır.

2. Orijinden i. gözleme öklid uzaklığı, $\|\mathbf{a}_i\|^2 = (\mathbf{x}_i^T\mathbf{x}_i)$, i. satırın toplam varyansa katkısını verir.

3. Grafiksel yaklaşımda herhangi iki değişken vektörü arasındaki uzaklıklar Öklid uzaklığı:

$\left\|\mathbf{b}_j - \mathbf{b}_{j'}\right\|^2 = \left(\mathbf{x}_j - \mathbf{x}_{j'}\right)^T\left(\mathbf{x}_j - \mathbf{x}_{j'}\right)$, burada, $\mathbf{b}_j$, **B**'nın j. sütunu ve $\mathbf{x}_j$, **X**'in j. sütunudur.

4. Orijinden j. değişken vektörüne Öklid uzaklığı, j. Değişkenin standart sapması ile doğru orantılıdır. Yani, $\left\|\mathbf{b}_j\right\|^2 = (\mathbf{x}_j^T\mathbf{x}_j) = (n-1)s_j^2$ eşitliği sağlanır.

5. Grafiksel yaklaşımda, $\mathbf{b}_j$ ve $\mathbf{b}_{j'}$ vektörleri arasındaki açının kosinüsü yaklaşık olarak j ve j′ değişkenleri arasındaki korelasyonu verir.

$$r_{jj'} = \frac{\mathbf{b}_j^T\mathbf{b}_{j'}}{\left\|\mathbf{b}_j\right\|\left\|\mathbf{b}_{j'}\right\|} = \frac{s_{jj'}}{\sqrt{s_{jj}\,s_{j'}s_{j'}}} \approx \cos\left(\theta_{jj'}\right) \qquad (2.10)$$

Eşitlik 2.10'a göre, yüksek pozitif korelasyonlu değişken vektörleri arasındaki açı küçüktür. Eğer herhangi iki değişken vektörü arasındaki açı çok küçükse, bu iki değişken arasındaki korelasyonunda yüksek olması beklenir. Vektörler arasındaki açı

90˚ ise bu iki vektör birbirine diktir ve ilgili değişkenler arasında herhangi bir ilişki yoktur. Eğer vektörler arasında açı geniş açı ise, negatif eğilimli ilişkiden söz edilebilir.

2.9.4. Veri Matrisinin Varyans Ayrıştırması

X veri matrisinin varyans ayrıştırmasını göstermek için ilk olarak bu veri matrisinin normu kavramı üzerine yoğunlaşmak gerekir. Buna göre norm yapısı kullanılarak,

$$\|\mathbf{X}\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2 = iz(\mathbf{X}^T\mathbf{X}) = \sum_{k=1}^{p} \gamma_k^2 = (n-1)\sum_{j=1}^{p} s_j^2 \qquad (2.11)$$

eşitliği yazılabilir. Burada γ_k, k. Tekil değerdir.

$$\sum_{j=1}^{p} s_j^2 \; ise \text{ X veri kümesinin p değişkenine ilişkin toplam varyansı}$$

2.10. Diğer Çok Değişkenli Grafikler

2.10.1. Ağaç Diyagramları

Değişkenlerin kümelenmesi sürecinde çok sık kullanılan grafiklerden biri ağaç diyagramlarıdır. Dikey çizgiler bir araya gelmiş kümeleri temsil eder. Bir Ağaç grafiği, hiyerarşik bir ağaçtaki veri noktalarını birbirine bağlayan birçok U şekilli çizgiden oluşur. Her U'nun yüksekliği, bağlanan iki veri noktası arasındaki mesafeyi temsil eder. Üzerindeki çizginin pozisyonu kümeler bir araya geldiğinde uzaklıkları gösterir. Ağaç diyagramı, nesneler arasındaki hiyerarşik ilişkiyi gösteren bir şemadır. Genellikle hiyerarşik kümelemeden bir çıktı olarak yaratılır. Bir ağaç diyagramının temel kullanımı, nesneleri kümelere ayırmanın en iyi yolunu bulmaktır

Ağaç diyagramlarında, değişkenlerin hangi aşamalarda ve hangi uzaklık (ya da benzerlik) düzeyinde bir araya geldiklerini görmek mümkündür. Ağaç diyagramlar dikey çizilebileceği gibi yatay da çizilebilir. Bu diyagramların bir başka özelliği ise, aynı yere bağlanan herhangi iki gözlemin yer değiştirebilmesidir. Ağaç grafiği soldan sağa okunur. Ağaç grafiklerinde birbirine yakın olan veriler analizin başında birleşirler. Düşey formdaki bir diyagram baş aşağı bir ağaç şeklindedir. En altta yapraklar olarak veri seti elemanları bulunur. Ağaçta yukarıya doğru çıkıldıkça birbirine benzer olan gözlemler dallar halinde birleşmeye başlarlar. Bunlar da daha üst düzeylerde birleşirler.

Düşey eksende gözlenen birleşim yüksekliği elemanların benzemezliğinin göstergesidir. Daha uzun bir dal daha az benzeyen elemana karşı gelir. Birleşmiş iki elemanın ne kadar benzer veya benzemez olduğu sadece dal yüksekliğinden anlaşılır. Yatay eksendeki pozisyonlar benzerliği yansıtmazlar. Alt grupları belirlemek için grafik belirli bir yükseklikten kesilebilir.

Hiyerarşik kümeleme yöntemindeki kümeleme süreci, aşamalı bir yapıda olup bir alt aşamadaki küme alt grupları bir sonraki aşamadaki kümeleri oluşturmak için bir araya getirir. Hiyerarşik kümeleme yöntemleri gözlemleri kümelemek amacıyla uygun uzaklık veya benzerlik ölçülerini dikkate alırken, aşamaların ve kümelenmelerin kolay anlaşılması için ağaç diyagramlarından veya buz saçağı grafiklerinden yararlanılır (Alpar, 2011, s. 314).

Ağaç Diyagramı size kaç tane kümenizin olması gerektiğini söyleyemez. İnsanların bu grafikleri okurken yaptığı yaygın bir hata, grafiğin şeklinin kaç tane küme olduğuna dair bir ipucu verdiğini varsaymaktır. Genel olarak, verilerdeki küme sayısını belirlemek için ağaç diyagramlarını bir araç olarak kullanmak bir hatadır. Açıkça "doğru" kümelerin olduğu yerlerde bu genellikle bir ağaç diyagramında belirginleşir. Bununla birlikte, ağaç grafikleri, sonucu destekleyen gerçek bir kanıt bulunmadığında, genellikle doğru sayıda küme önermektedir.

2.10.2. Buz Saçağı Grafikleri

Şekli buz saçağına benzemesi nedeniyle buz saçağı grafiği olarak adlandırılan grafikte, küme sayısını belirlemek oldukça kolaydır. Buz saçağı grafikleri, yatay ya da dikey olarak çizilebilir. Bu grafik yardımıyla, her aşamadaki küme sayısını ve kümeleri görmek mümkündür. Dikey Buz Saçağı Grafiği'nde kümeler arasındaki En Küçük Öklid Mesafesi baz alınarak aşağıdan yukarı doğru okunur. Diğer taraftan yatay saçağı grafiği, bir ağaç diyagramına benzer. Burada, grafik sağdan sola okunur ve vakaların seçimi Öklid mesafesine göre yapılır. Sonraki adımlarda, en yakın değişkenler (veya kümeler) birleştirilir. Buz saçağı grafikleri en benzerlerinin birinci aşamada, ikinci aşamada en benzerlerinin bir araya getirildiği, tüm birimler büyük bir küme oluşturmak için çizilene kadar hiyerarşik değerlerin kümelenme işlemini göstermektedir.

Özellikle kümeleme analizi sonuçlarında kullanılan bu yöntemde sütunlar kümelenen objelere, satırlar kümelerin sayılarına denk gelir. Buz saçağı diyagramı

aşağıdan yukarı doğru okunur. Kaç sınıf olması gerekirse ona göre sınıflandırma yapılır.

Burada yapılacaklar;

- Önce değişkenlerin tanımlanarak kümeleme problemleri formülüze edilir.
- Uygun mesafe ölçüsü seçilir.
- Bu mesafe ölçüsü kümelenen objelerin ne kadar benzer ya da benzer olmadıklarına karar vermede kullanılır.
- Birçok kümeleme yöntemi geliştirilmiştir. Uygun olanı seçilir.
- Kümelerin sayısına araştırmacı karar verir.
- Oluşturulmuş kümeler değişkenlerin kümelenmesi bakımından yorumlanmalı ve güvenli değişkenler bakımından profile edilmelidir.

2.10.3. Path Diyagramı

Path diyagramı açıklanan(bağımsız) ve açıklayıcı(bağımlı) değişkenler arasındaki ilişkileri bir şema şeklinde ifade etmeye yarayan görsel bir araçtır. Path diyagramı şekil itibariyle akış şemasına benzemektedir. Aralarındaki farka bakacak olursak akış şemaları sebepsel bir akışı görselleştiren bir doğruyla değişkenleri birbirine bağlarken, path diyagramı ise, değişkenlerin diğer değişkenler üzerinde sebepsel değişimini gösteren bir şema olarak düşünülebilir. Birbirleriyle sebep-sonuç ilişkisi içinde olduğu görülen değişkenler arasındaki bağlantı, path diyagramları ile gösterilebilir. Path diyagramlarında tek yönlü oklar kullanılır ve bu oklar her bağımsız değişkenden kendisine bağımlı olan değişkene doğru çizilir. Path katsayılarının sembolik (path girdi diyagramı) veya sayısal değerleri (path çıktı diyagramı) çizilen oklar üzerine yazılır. Girdi path diyagramı, olası nedensellik ilişkilerini gösteren henüz tahmin edilmemiş temsili parametreleri ifade ederken, çıktı path diyagramı istatistiki analizin sonuçlarını ifade eder ve tahmin edilen parametre değerlerini (katsayıları) gösterir. Sistem içerisinde diğerlerine bağımlı olmayan değişkenler arasındaki korelâsyonlar ise, iki yönlü oklar tarafından gösterilir ve birleştirici eğri biçiminde çizilir. İki yönlü eğri biçimindeki ok durumunda ise basit korelasyon katsayılarının sembolik veya sayısal değerleri yazılır.

A değişkeni B değişkenini etkiliyor ancak B değişkeni A değişkenini etkilemiyorsa A→B şeklinde gösterilir. Değişkenler birbirini etkiliyor ise ve her iki

yönlü ok A↔B şeklinde gösterilir. Bu ilişki eğrisel olarak çizilen iki başlı ok şeklinde de gösterilebilir. Eğrisel olarak çizilen iki başlı ok ise değişkenler arasında ilişki olduğunu ancak neden-sonuç ilişkisi olmadığını gösterir. Diğer taraftan regresyon modellerindeki hata terimi de path diyagramlarında yer alır. Hata terimi de bağımlı değişkene doğru tek yönlü ok ile çizilir.

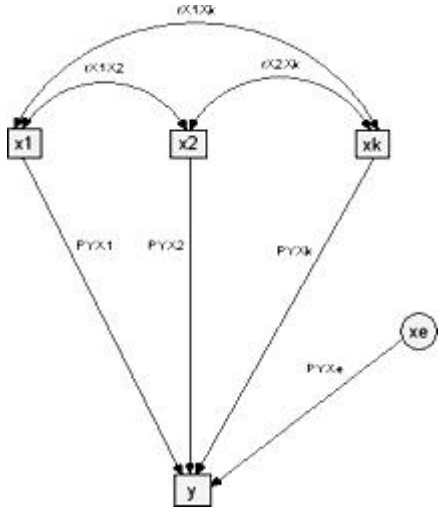
Path diyagramının çizilmeden önce araştırmacının konuyla ilgili muhakkak ön bilgiye sahip olması gerekir. Bu doğrultuda diyagramda bulunan değişkenler arasında sebep-sonuç ilişkisinin belirlenmesi ve bu ilişkilerin doğrudan mı yoksa dolaylı olarak mı oluştuğunun ortaya konulması gerekir. Bundan dolayı araştırmacı konuyu çok iyi bilmeli, değişkenler arasındaki ilişki konusunda ayrıntılı bilgiye sahip olmalıdır. Bu bilgiler eşliğinde path diyagram oluşturulduktan sonra path katsayıları hesaplanır ve yorumlanır.

Örneğin X_1, X_2,.., X_k bağımsız değişkenleri ile X_e hata değişkeni ve bunların oluşturduğu Y bağımlı değişkeninin meydana getirdiği sebep-sonuç ilişkisi:

$$Y=b_0+b_1X_1+b_2X_2+...+b_kX_k+X_e \qquad (2.12)$$

doğrusal regresyon modeli ile gösterilebilir. Bu modelde b_0 sabit terim (regresyon sabiti), b_i katsayıları ise kısmi regresyon katsayılarını temsil etmektedir. X_e, Y bağımlı değişkenine ait hata değişkenidir. X_e'nin sıfır ortalamalı [$E(X_e)=0$] ve $\sigma^2_{X_e}$ varyanslı normal bir dağılım gösterdiği ve diğer X_i bağımsız değişkenlerinden bağımsız olduğu varsayılır. X_i bağımsız değişkenlerinin ise hatasız ölçüldüğü kabul edilir. Yani X_i değişkenlerine ait hatalar dikkate alınmaz.

Eşitlik 2.12'de doğrusal model ile belirlenen neden-sonuç ilişki sisteminde X_i, (i=1,2,...,k) bağımsız değişkenlerine sebep, Y bağımlı değişkenine ise sonuç denir. X_e ise sistemde görülmeyen diğer etki faktörlerinin tümünü içerir. Eşitlik 2.12'deki doğrusal modelin belirlediği sebep-sonuç ilişkileri Şekil 2.3'deki gibi gösterilebilir ve bu diyagramda görüldüğü gibi, sebep değişkenlerinin sonuca olan etkileri oklarla gösterilmiştir. Okun yönü, etkinin yönünü gösterir. Bağımsız değişkenler arasındaki ilişki ise iki tarafta ok başlıklı eğriler ile ifade edilmiştir.



Şekil 2.3. X_1, X_2,..,X_k sebep değişkenleri ile Y sonuç değişkeni arasındaki ilişki

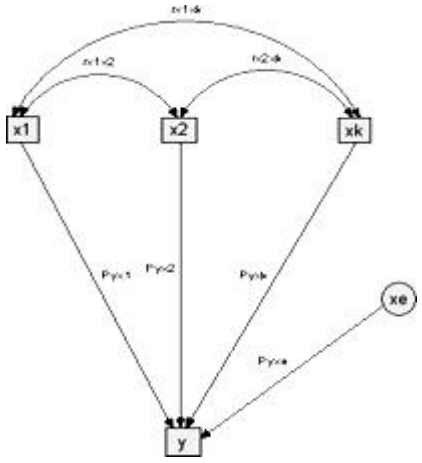
Path Analizi Tekniği, standartlaştırılmış değişkenler arasındaki ilişkileri incelediğinden Eşitlik 2.12'deki doğrusal modeldeki değişkenler önce standartlaştırılır, daha sonra standartlaştırılan değişkenler için doğrusal regresyon modeli baştan yazılarak aşağıdaki eşitlik elde edilir:

$$y=P_{yx_1}x_1 + P_{yx_{12}}x_2 + \cdots + P_{yx_k}x_k + P_{yx_e}x_e \qquad (2.13)$$

Burada, y ve x_i değişkenleri standartlaştırılmış değişkenleri, x_e ise hata terimini temsil etmektedir. Kısmi regresyon katsayıları, tanım gereği path katsayılarıdır. Path katsayıları P_{yx_i} ile ifade edilip, x_i sebep değişkeninden y sonuç değişkenine giden etki miktarını gösterir. Bu durumda, sebep-sonuç sisteminde standartlaştırılmış değişkenler arası ilişkileri inceleyen analize "Regresyon Analizi" yerine "Path Analizi" adı verilir. x_i sebep değişkeni ile y sonuç değişkeni arasındaki path katsayısı da:

$$P_{yx_i} = \frac{\sum x_k y_{ki}/n}{\sum x_k^2/n} = E(x_k y) = Cov(x_k y) = r(x_k, y) \ = b_k \frac{\sigma_{x_k}}{\sigma_y} \qquad (2.14)$$

Şekil 2.3'de verilen sebep-sonuç diyagramı buraya kadar açıklanan kurallar yardımıyla Şekil 2.4'deki path diyagramına dönüştürülür. Buradan da anlaşılacağı gibi path diyagramı ile sebep-sonuç diyagramı arasında şekilsel bir farklılık yoktur. Ancak path diyagramında, değişkenler standartlaştırılmış ve pathları gösteren oklar üzerine de path katsayıları yazılmıştır.

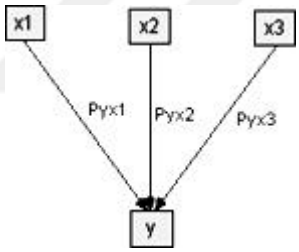


Şekil 2.4. Sebepler arasında korelasyon olduğu durumda $\mathbf{x_i}$ sebep değişkenleri, $\mathbf{x_e}$ hata değişkeni ve y sonuç değişkeni arasındaki ilişkiyi gösteren path diyagramı

Değişkenler arasındaki ilişkiye göre belli başlı path diyagramları aşağıda gibidir:

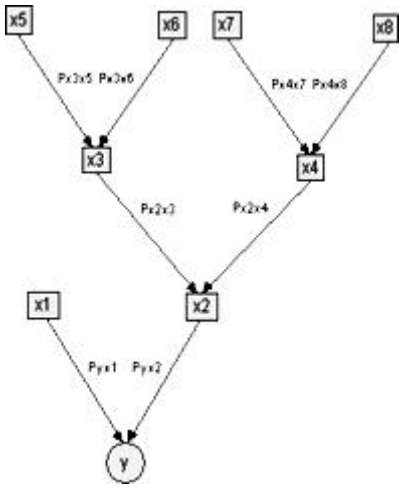
a) Sebep Değişkenleri Arasında Korelasyonun Olmadığı Sistemler

Burada, P_{ij}, x sebep değişkeni ile y sonuç değişkeni arasındaki path katsayısını göstermektedir.



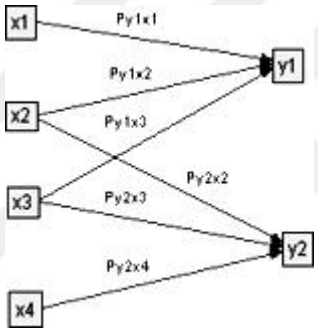
Şekil 2.5. $\mathbf{x_1}$, $\mathbf{x_2}$ ve $\mathbf{x_3}$ gibi birbirinden bağımsız sebeplerin y sonucuna etkilerini gösteren path diyagramı

b) Korelasyonsuz (Bağımsız) Sebepler Zinciri



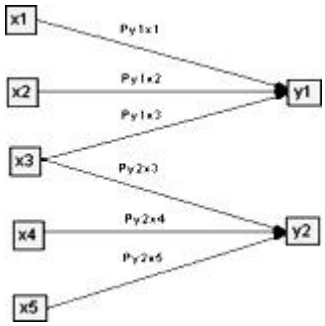
Şekil 2.6. Korelasyonsuz sebepler zincirini gösteren path diyagramı

c) Sebepler Arasında Korelasyonun Olmadığı Ortak Sonuçlar Sistemi



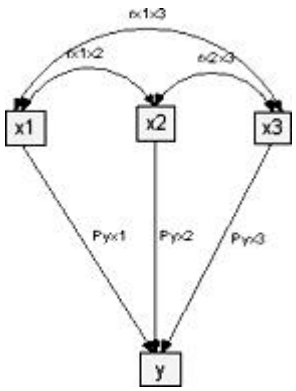
Şekil 2.7. İki ayrı sonucun birbirinden bağımsız ortak sebepler tarafından etkilenmesini gösteren path diyagramı

d) Korelasyonlu Ortak Sebep İçeren Sonuçlar Sistemi



Şekil 2.8. Sebepler arasında korelasyonun olduğu sonuçlara ait path diyagramı

e) Birbirine Bağımlı (Aralarında Korelasyon Bulunan) Sebepler Sistemi



Şekil 2.9. Birbirine bağımlı değişkenlerin y sonucuna etkilerini gösteren path diyagramı

Not: $r_{xx_{ij}} \neq 0$ => değişkenler birbiriyle yer değiştirebilir. Korelasyon vardır.

$r_{xx_{ij}} = 0$ => değişkenler birbiriyle yer değiştiremez. Korelasyon yoktur.

BÖLÜM 3

UYGULAMALAR

Çalışmanın bu aşamasında toplanmış olan veriler grafiksel olarak gösterilecektir. Grafiksel gösterimler için yararlanılan programlar: SPSS, Statistica, STATA, R Studio.

Uygulama 3.1. Sağlıklı ve uzun bir ömür herkesin en büyük isteğidir. Günlük hayatta yapılabilecek küçük değişikliklerle hastalıklardan korunup sağlıklı, mutlu ve uzun bir hayat yaşamak mümkün olabilir. Bunun için çaba göstermek gerekir.

Yapılan sağlık testinde seçilen 50 kişinin ankete verdiği cevaplara ilişkin aldıkları puanlar ve sağlıklı yaşam için yapılması gerekenlerle ilgili 5 değişken rassal olarak seçilmiştir. Bu değişkenler ve açıklamaları aşağıdaki gibidir:

X1: Dengeli beslenmeye ilişkin puan;

X2: Düzenli uyumaya ilişkin puan;

X3: Spor yapmaya ilişkin puan;

X4: Temiz kalmaya ilişkin puan;

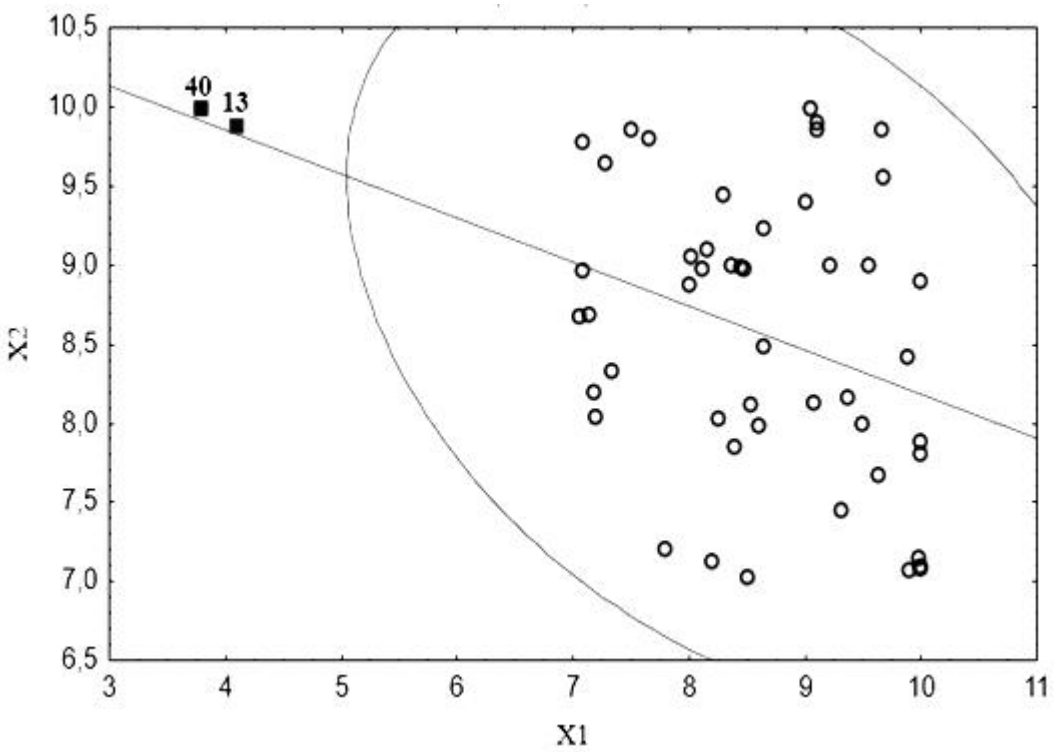
X5: Zamanında yapılan aşılara ilişkin puan;

Bundan sonraki süreçte grafiksel gösterimler, bu veriler ışığında çizilecektir.

Tablo 3.1. Sağlık Testi Yapılan 50 Kişinin 5 Değişkene İlişkin Puanları

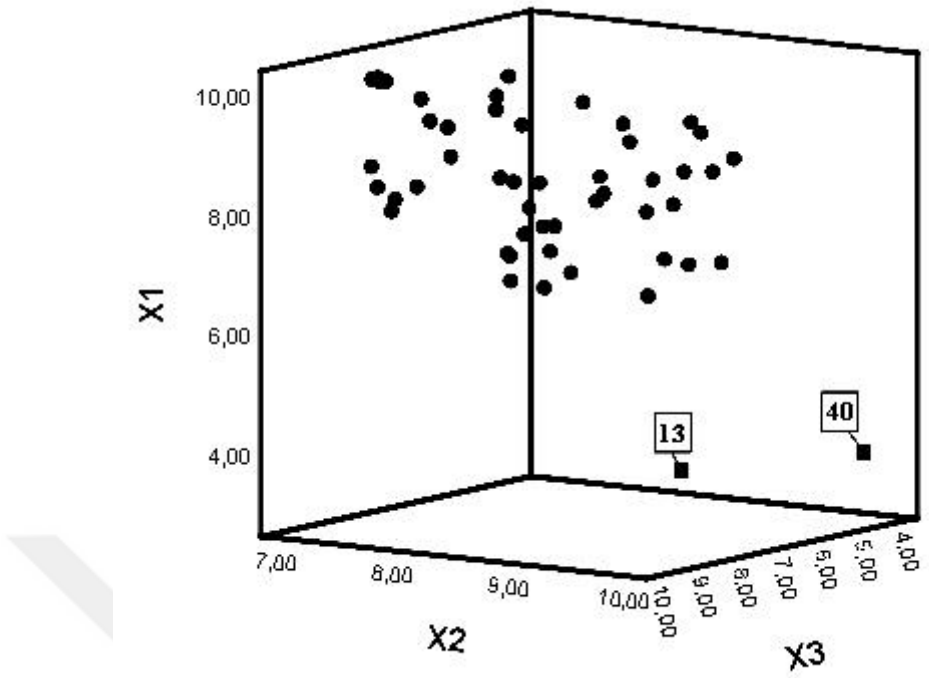
	X1	X2	X3	X4	X5
1	7,2	8,04	7,7	9	10
2	8,5	7,02	8,14	8	9
3	10	8,9	8,3	7	9
4	9,1	9,85	7,28	8	8
5	6,32	6,99	8,16	9	7
6	8,12	8,98	9,5	8	8
7	7,5	9,85	8,4	7	9
8	9,05	9,99	8,2	9	10
9	8,4	7,85	9,93	9	8,5
10	8,25	8,03	6,88	8	8
11	6,36	9,44	5,33	8	9
12	7,14	5,98	6,32	8	7,5
13	4,1	9,88	8,66	8	7
14	9,55	9	6,84	9	9
15	9,98	7,15	8,32	9	8
16	8,54	8,12	8,1	10	9
17	7,08	8,96	5,87	10	10
18	8,61	5,45	9,76	9	7,5
19	9,22	9	7,42	8	10
20	6,44	9,64	6,39	8	8
21	8,2	7,12	8,28	7,5	7
22	5,24	8,67	7,94	6	4
23	9,32	7,45	7,92	9	9
24	9,63	7,67	6,91	8,5	7
25	10	7,88	9,37	10	8
26	8,45	8,99	8,04	7	9
27	10	7,08	8,16	8	10
28	8	8,88	9,68	9	8,5
29	8,48	8,97	9,82	5	9
30	9,9	7,07	7,92	8	8
31	10	7,8	6,99	8	7
32	7,65	9,8	8,85	7	6
33	6,18	8,2	7,1	9	5
34	9,1	9,9	8,66	9,5	7
35	9,88	8,42	9,07	8	8
36	8,65	9,23	7,51	8	8
37	8,36	9	5,9	10	7
38	9,68	9,55	7,49	9	9
39	7,33	8,33	8,46	7	10
40	3,8	9,99	4,48	4	4,5
41	8,16	9,1	9,56	9	7
42	8,65	5,48	8,8	8	9
43	9,37	8,16	7,68	9	8
44	7,4	9,4	9,3	9	10
45	9,66	9,85	8,1	10	10
46	7,08	9,78	9,2	6,5	8
47	8,02	9,05	7,15	9	9
48	9,08	8,13	9,35	10	8
49	10	7,09	8,32	9	9
50	9,49	7,99	9,03	9	10

Grafik 3.1.a.'da X1 ile X2 arasındaki saçılım dağılımına göre, merkeze göre gözlemlerin % 95'ini kapsayan sınırların kontur grafiği çizilmiştir. 13. ve 40. gözlemler bu kontorun kapsama alanı dışında kalmıştır.

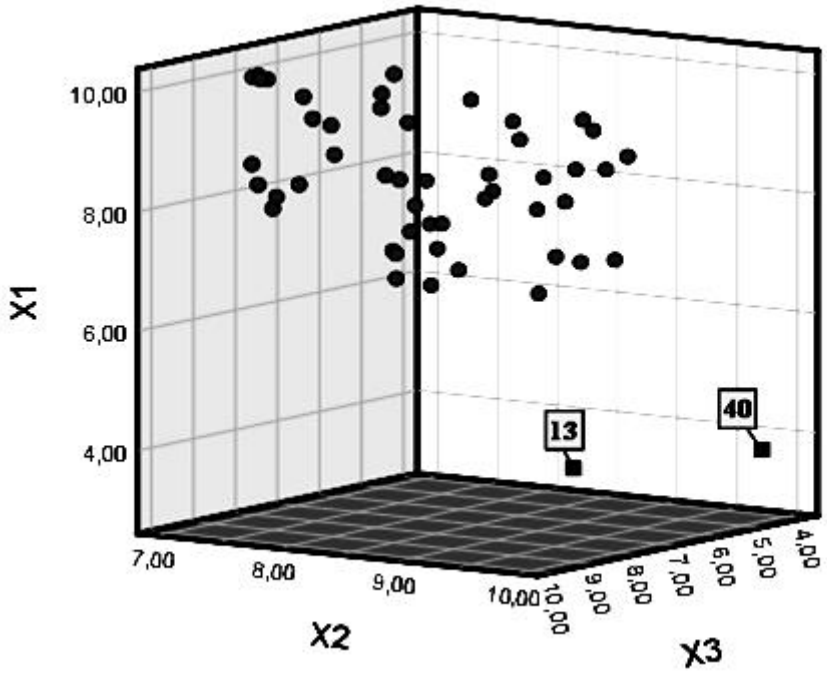


Grafik 3.1.a. X1 ve X2'ye Ait Saçılım Grafiği ve % 95 Konturu (STATISTICA)

Grafik 3.1.b ve aynı örneğe ilişkin Grafik 3.1.c'de üç boyutlu saçılım grafikerinden yararlanılmış ve kişilere ait X1, X2 ve X3 baz alınarak; anket sonucu kişilerin sağlıkla ilgili verdiği spor yapmaya ilişkin puanlar, düzenli uyumaya ilişkin puanlar ile dengeli beslenmeye ilişkin puanlar ilişkilendirilmiştir. Ayrıca kişilerin verdiği X1 değişkenine ait cevaplar yani aldığı puanlar, X2 ve X3 değişkenlerinin sonuçlarına gore doğru orantılı olarak azalıp artmaktadır. Bu üç değişkenli veride 13 ve 40 nolu gözlemler aşırı gözlem olarak karşımıza çıkmaktadır. Aşırı gözlemlere ilişkin ayrıntılı bilgiye Bölüm 1'de yer verilmiştir.



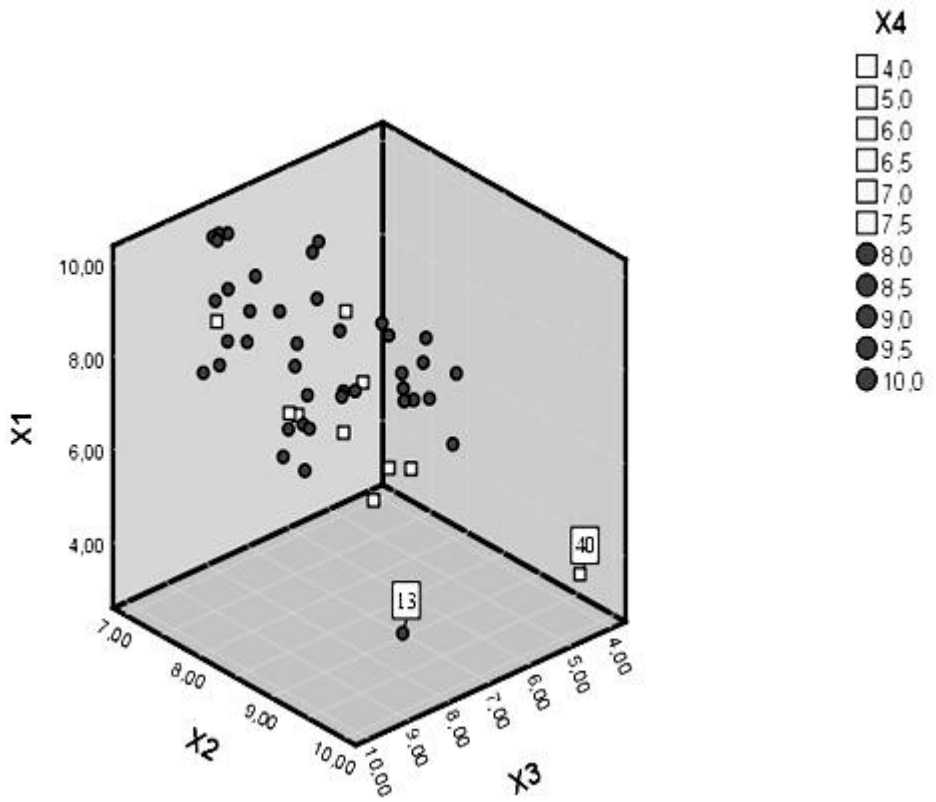
Grafik 3.1.b. Tablo 3.1'deki X1, X2 ve X3 Değişkenlerine İlişkin 3 Boyutlu Saçılım Grafiği(SPSS)



Grafik 3.1.c. Tablo 3.1'deki X1, X2 e X3 Değişkenlerine İlişkin 3 Boyutlu Saçılım Grafiği (Çizgilerle Desteklenmiş)

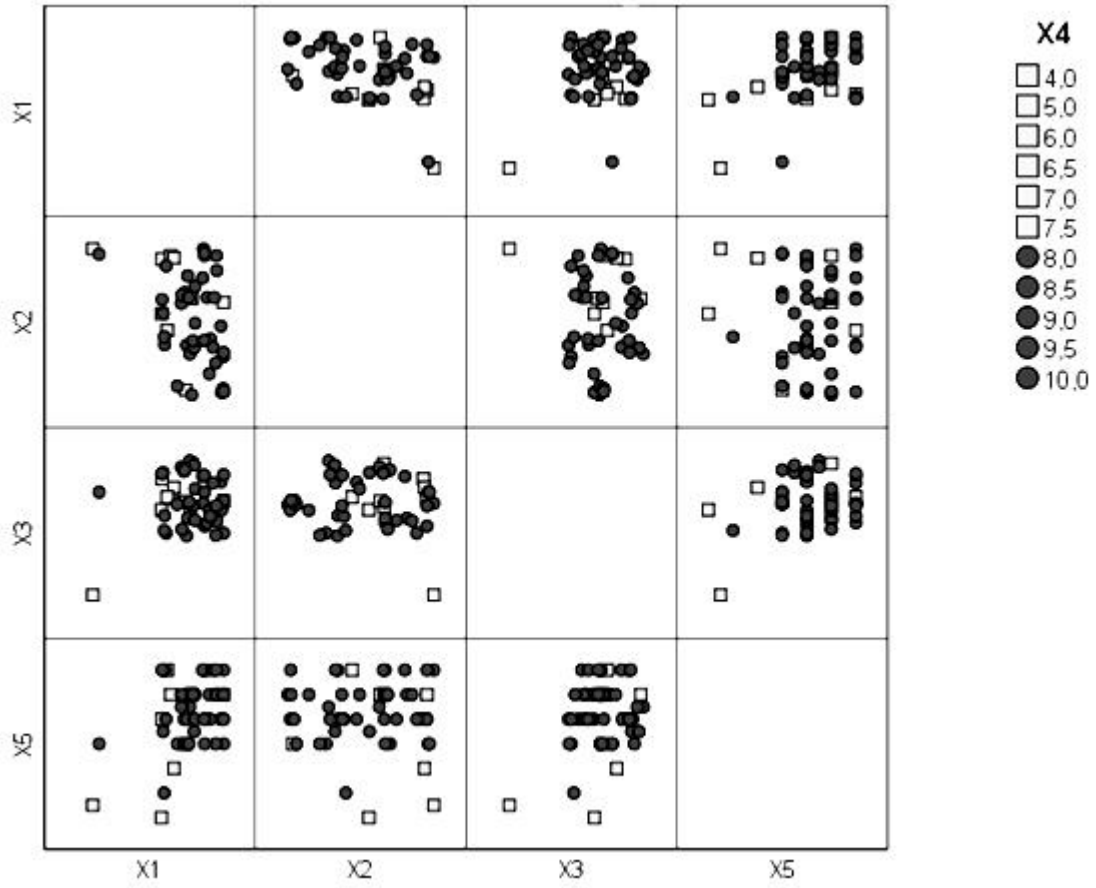
Grafik 3.1.b ve Grafik 3.1.c'ye bir başka değişkeni ekleyebiliriz. Bu ekleme söz konusu olduğunda, eklenen değişkenin sınıflandırılması için saçılım noktalarının işaretlerinin ayrı gösterilmesi gerekecektir. Bu yüzden, eklenecek olan X4 değişkeni

puanları: <8 olanlar için 1, >=8 olanlar için de 2 olarak kodlama yapıldıktan sonra aşağıdaki Grafik 3.1.d oluşmuştur.



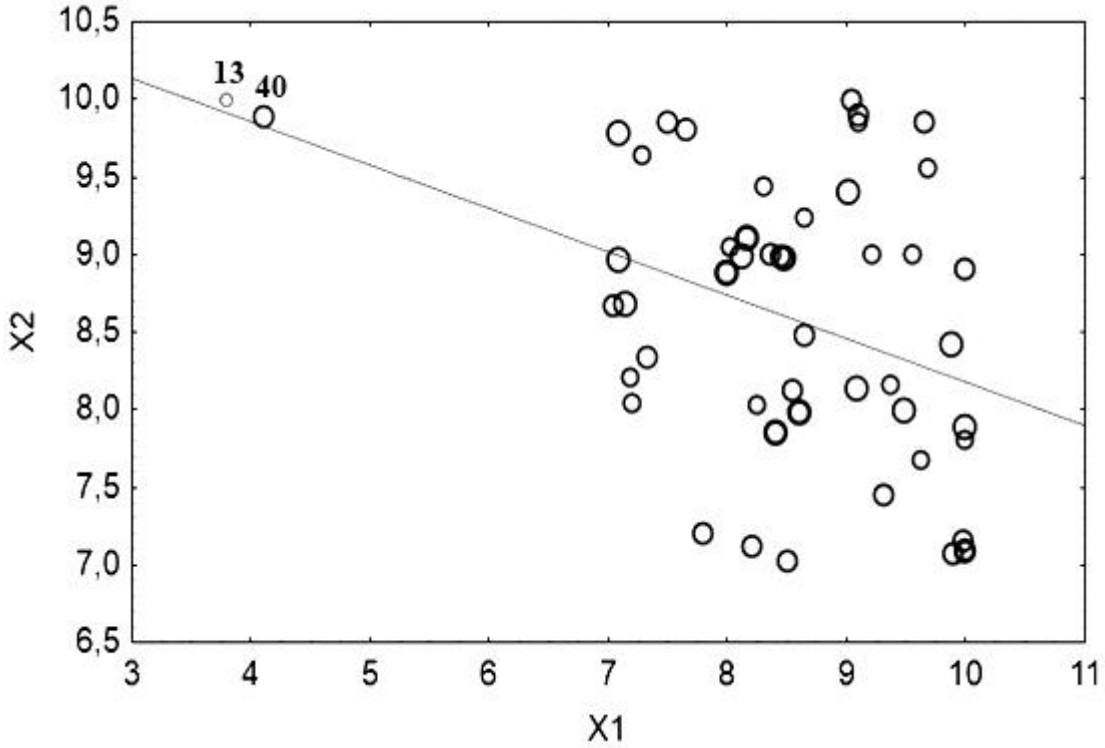
Grafik 3.1.d. Tablo 3.1'deki X1, X2 ve X3 Değişkenlerine X4'ün Eklenmesi İle 3 Boyutlu Saçılım Grafiği (SPSS)

Saçılım grafikleri ile ilgili son örnek aşağıdaki gibidir. Bu grafik, matris şeklinde olup saçılım grafiklerinin farklı bir uygulamasıdır. Bu ikişerli grafiklerde X4 değişkenine ilişin puanlar Grafik 3.1.d'deki gibi sınıflandırılmıştır. Bu grafikte her değişkenin değerleri de eksenlerde gösterilmiştir.



Grafik 3.1.e. Tablo 3.1'deki X1, X2, X3 ve X5 Değişkenlerinin X4'e Göre Matrissel Saçılım Grafiği (SPSS)

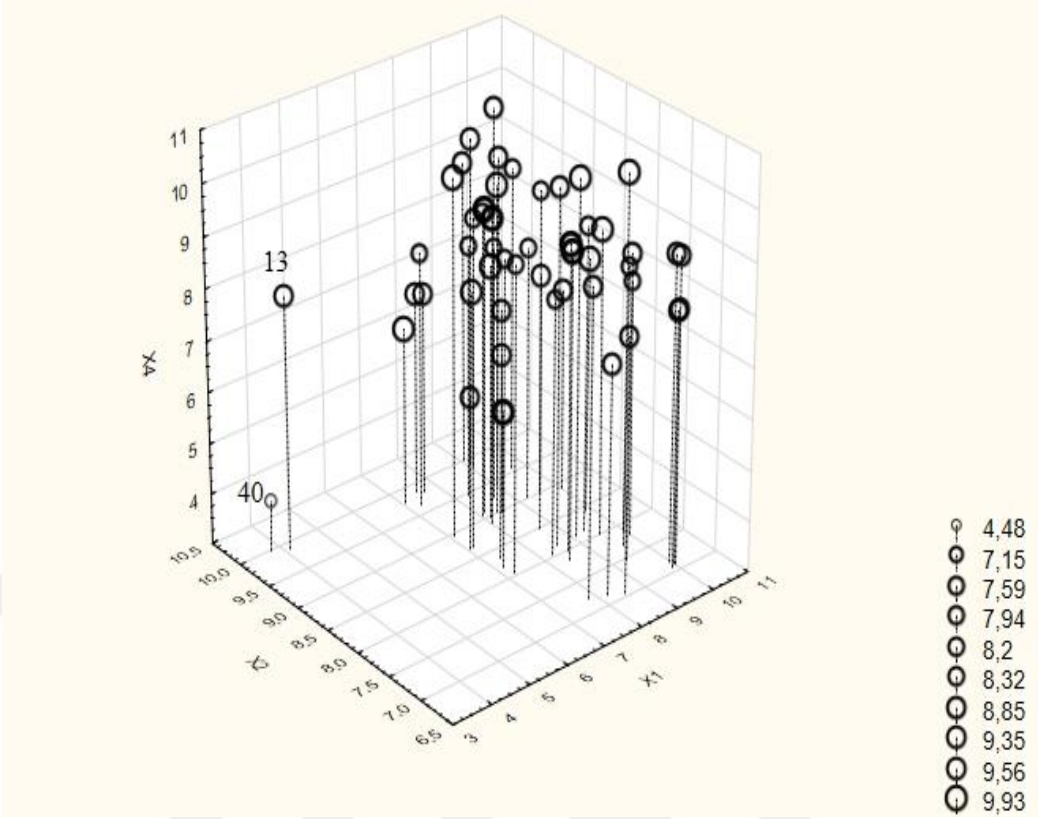
Tablo 3.1'deki veriler için ilk üç değişkene ilişkin çizilen kabarcık grafiği, Grafik 3.2'deki gibidir. Bu grafikte ilk aşamada, X1 ve X2 değişkeni arasında negatif korelasyon olduğu görülmektedir. Ayrıca X1 ve X2 değerleri artarken X3 değerleri de aynı yönde artış göstermektedir. Grafiğe bakacak olursak, 13. ve 40. gözlemlerin çapları diğerlerine göre çok küçüktür. Bu durum eklenen 3. değişkenin yani, X3'ün de aşırı gözlem olmaya devam etmesine sebep olmuştur.



Grafik 3.2. Tablo 3.1'deki X1, X2'ye göre X3 Değişkeninin Kabarcık Grafiği (STATISTICA)

Grafik 3.2'yi çizebilmek için STATISTICA programından Graphs>2D Graph>Scatterplot… aşamaları takip edilerek açılacak olan pencerede Advanced komutunu tıkladıktan sonra Graph Type'dan Bubble seçeneği seçilir. Bunlar yapıldıktan sonra Variables komutuyla değişkenler aktarılır. Kabarcık değişkeni, Weight değişkeni olarak atanacaktır.

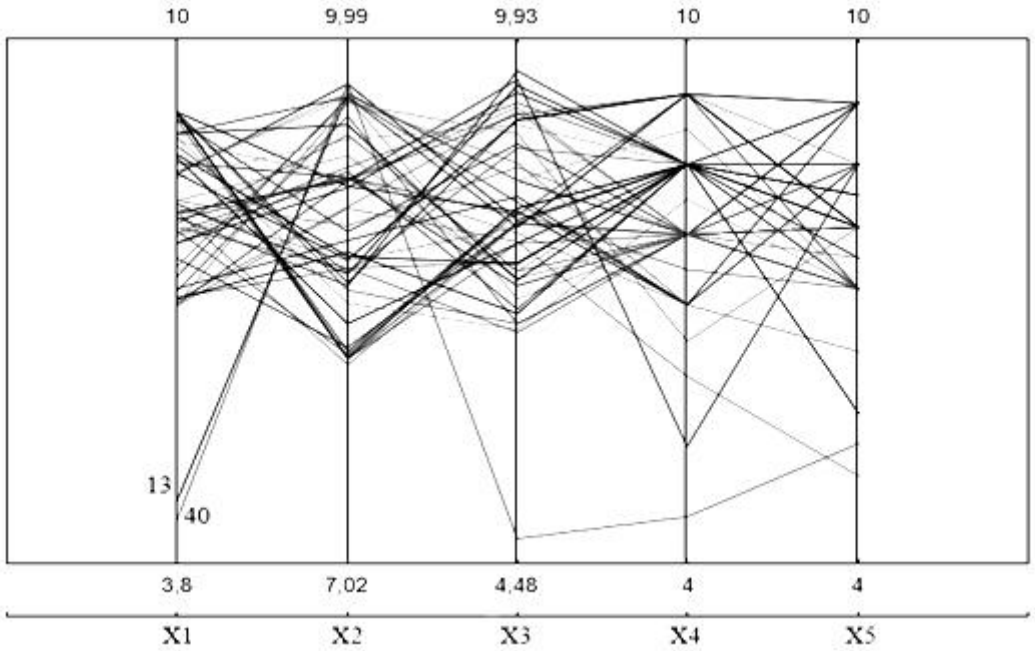
Kabarcık grafikleri veri sayısı üçten fazla olduğu durumlarda kullanılan en kolay yöntemlerden biridir. Böyle bir veri kümesi örneğinde, 3 değişkene ek olarak kabarcık büyüklüğünün de kullanıldığı 3 boyutlu grafiklerden yararlanılmaktadır. Buna ilişkin örnek Grafik 3.3'te verilmiştir.



Grafik 3.3. Tablo 3.1'deki X1, X2, X4'e göre X3'ün Kabarcık Grafiği

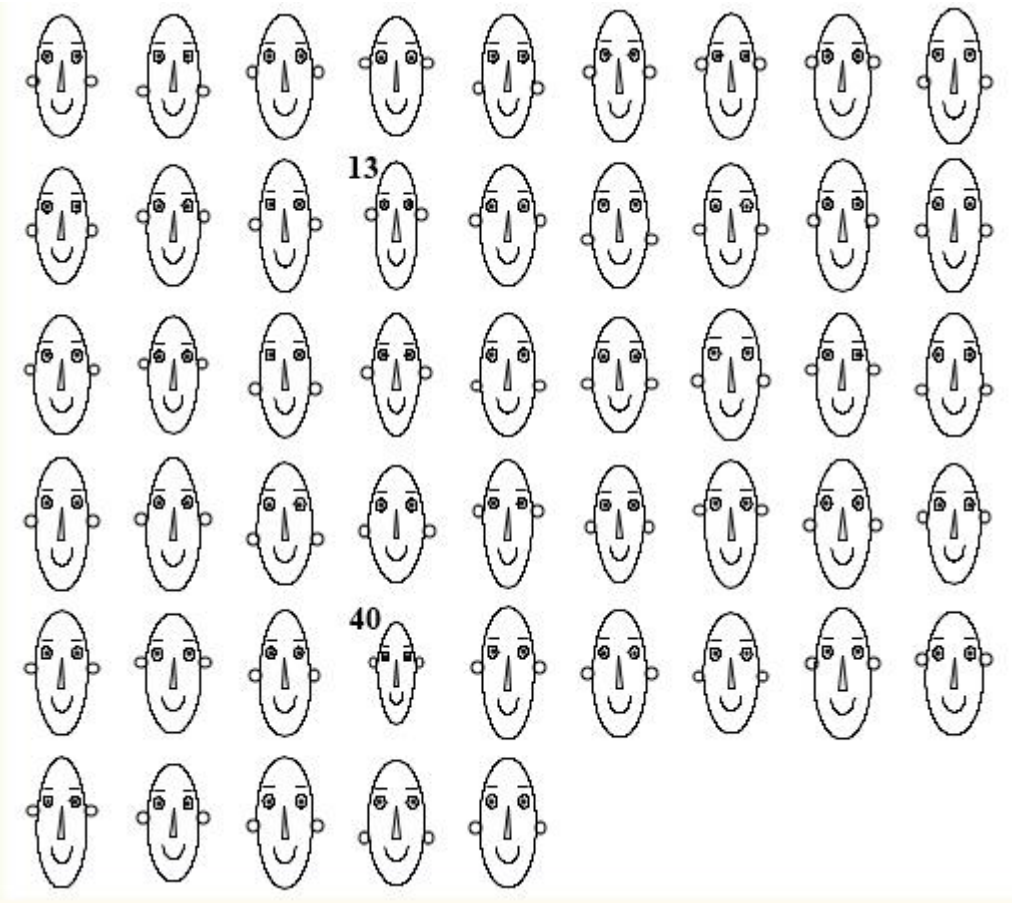
Grafik 3.3'te görüldüğü gibi X1, X2 ve X4'ün değerleri arttıkça X3 değişkenine ilişkin değerler de doğru orantılı olarak artmaktadır. Bu grafiği çizebilmek için, STATISTICA programından Graph>3D XYZ Graphs>Scatter Image Plots… adımları izlenerek açılacak olan pencerede Variables komutu tıklanır. Ardından açılan pencerede X, Y ve Z eksenlerine değişkenler atanır. Weight kısmına da Bubble'ı çizilecek olan değişken atandıktan sonra OK tıklanır. Elde edilen grafik noktaları tıklandıktan sonra açılan pencerede All Options komutu tıklanır ve bu pencerede Scatter Plot seçilir. Aynı zamanda, açılan penceredeki Type seçeneklerinden Regular yerine Bubble seçilmelidir.

Grafik 3.4'te Tablo 3.1'deki veriler kullanılarak oluşturulan paralel koordinatlar aşağıdaki gibidir:



Grafik 3.4. Tablo 3.1'deki X1, X2, X3, X4 ve X5'e İlişkin Paralel Koordinatlar (STATISTICA)

Tablo 3.1' deki gözlemlerin, beş değikene ilişkin Chernoff Yüzleri'ne Grafik 3.5'te yer verilmiştir. Grafikte görüldüğü gibi, 13. ve 40. gözlemlere ilişkin yüzler diğerlerine göre daha küçüktür. Diğer özelliklerinin yanında, kulak yerleri de birbirine benzerdir. Eğer bir kümeleme analizi yapılacak olursa bu iki gözlemin ayrı bir küme oluşturması büyük bir olasılıktır.

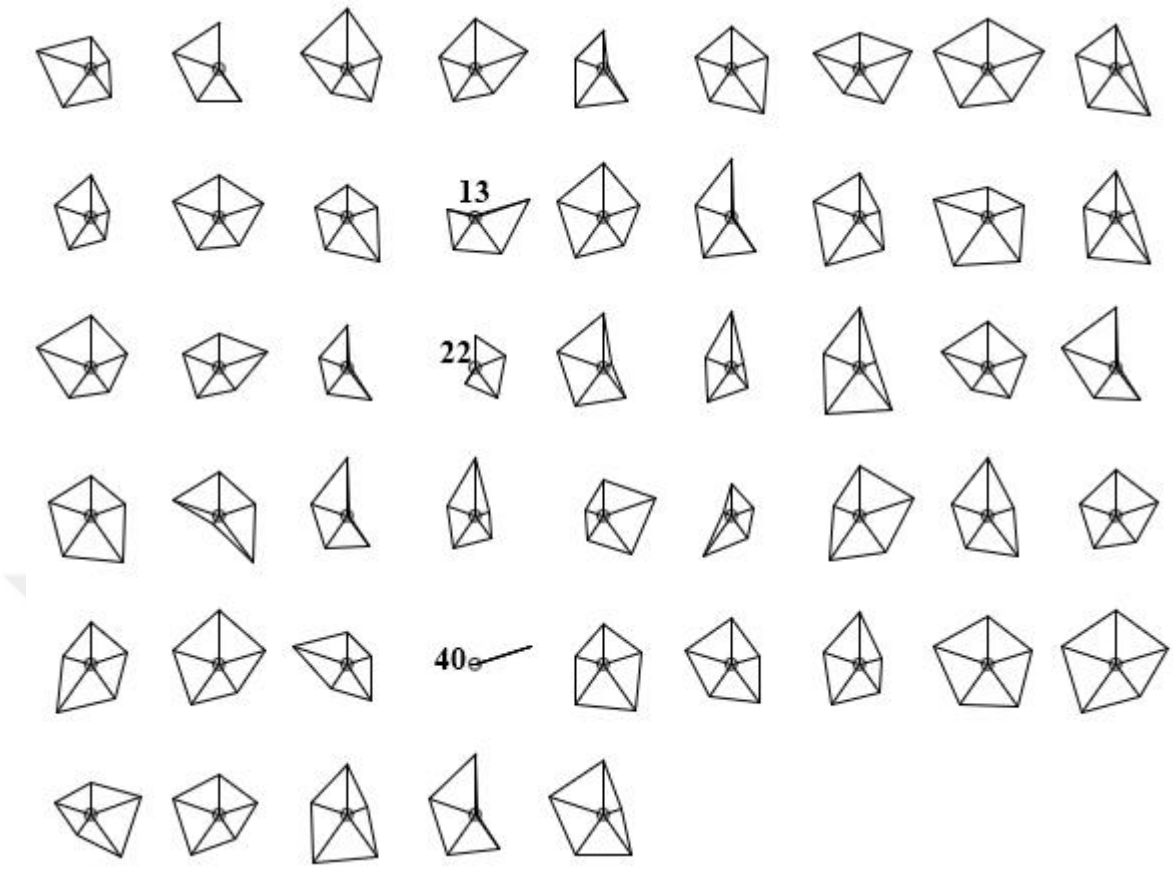


Grafik 3.5. Tablo 3.1'deki 5 Değişkene İlişkin Chernoff Yüzleri Grafiği

(X1: Yüz genişliği, X2: Kulak seviyesi, X3: Yarı yüz yüksekliği, X4: Üst yüz yayvanlığı, X5: Alt yüz yayvanlığı)

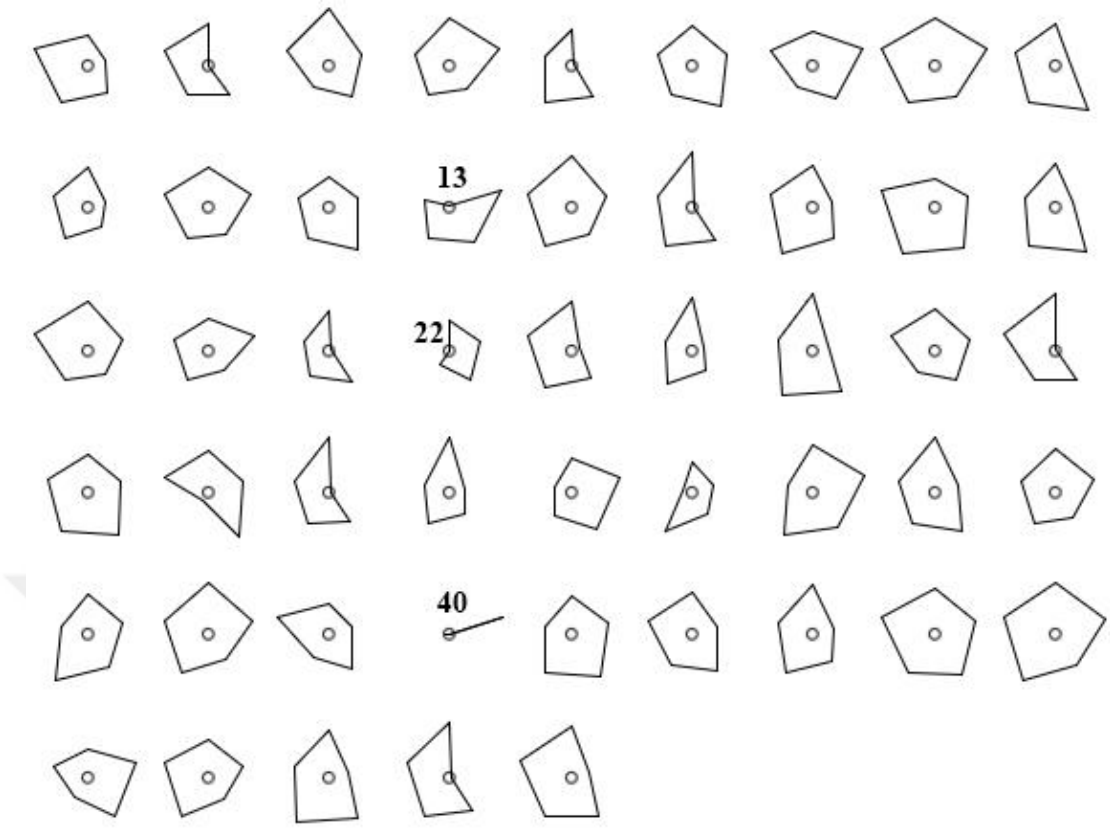
Burada, 13. ve 40. gözlemlere ilişkin değerlerde ciddi bir farklılık söz konusudur. Bu iki değişken ikonik olarak da diğerlerinden ayrılmaktadır.

Grafik 3.6'da Tablo 3.1'deki verilerden yararlanarak Yıldız İkon grafiği çizilmiştir. STATISTICA programı kullanılarak çizilen grafikte Chernoff Grafiği'nde olduğu gibi 40. gözlemin diğer gözlemlerin aldığı değerlerden farklı olduğu görülmektedir. Diğer kişilere göre oldukça düşük puan aldığı da anlaşılmaktadır.



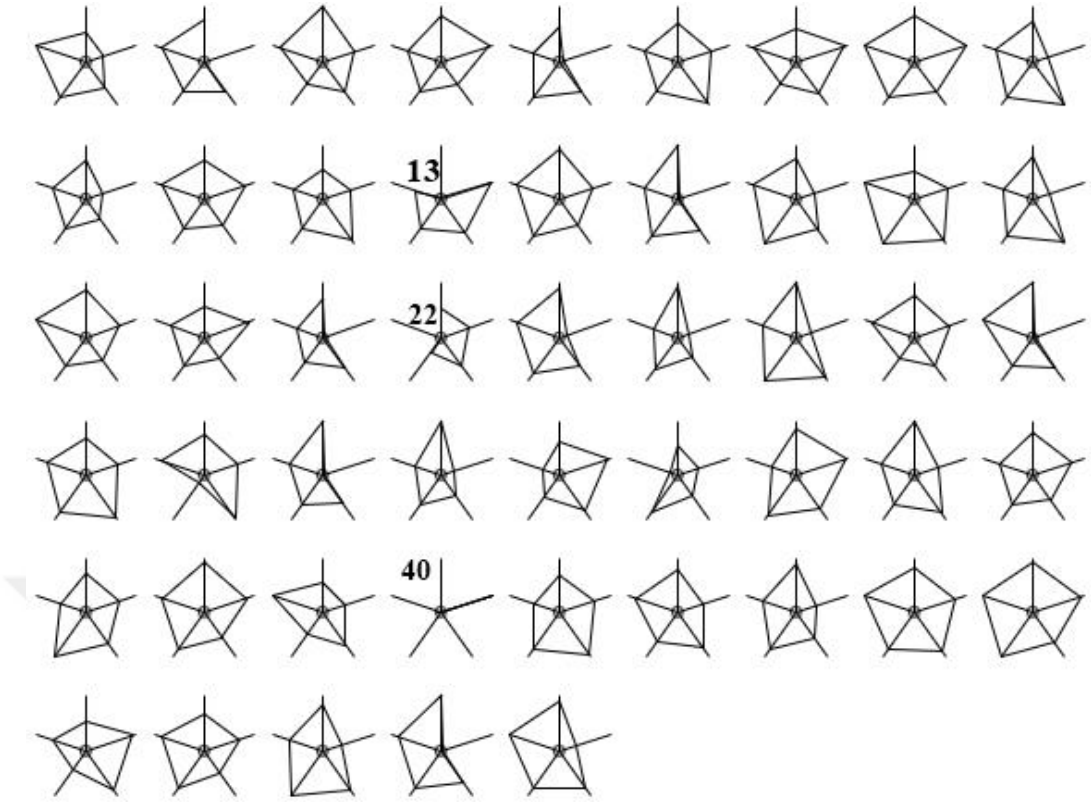
Grafik 3.6. Tablo 3.1'deki 5 Değişkene İlişkin Yıldız İkon Grafiği (STATISTICA)

Tablo 3.1'deki verileri poligon ikon grafiğinde de kullanacağız. Yıldız İkon grafiğinde olduğu gibi burada da diğer gözlemlerden farklı olarak göze çarpan 40. gözlem değerini görmekteyiz.



Grafik 3.7. Tablo 3.1'deki 5 Değişkene İlişkin Poligon İkon Grafiği (STATISTICA)

Tablo 3.1 verisindeki beş değişkene ilişkin gözlemler, güneş ışığı grafikleri için de kullanılacaktır.

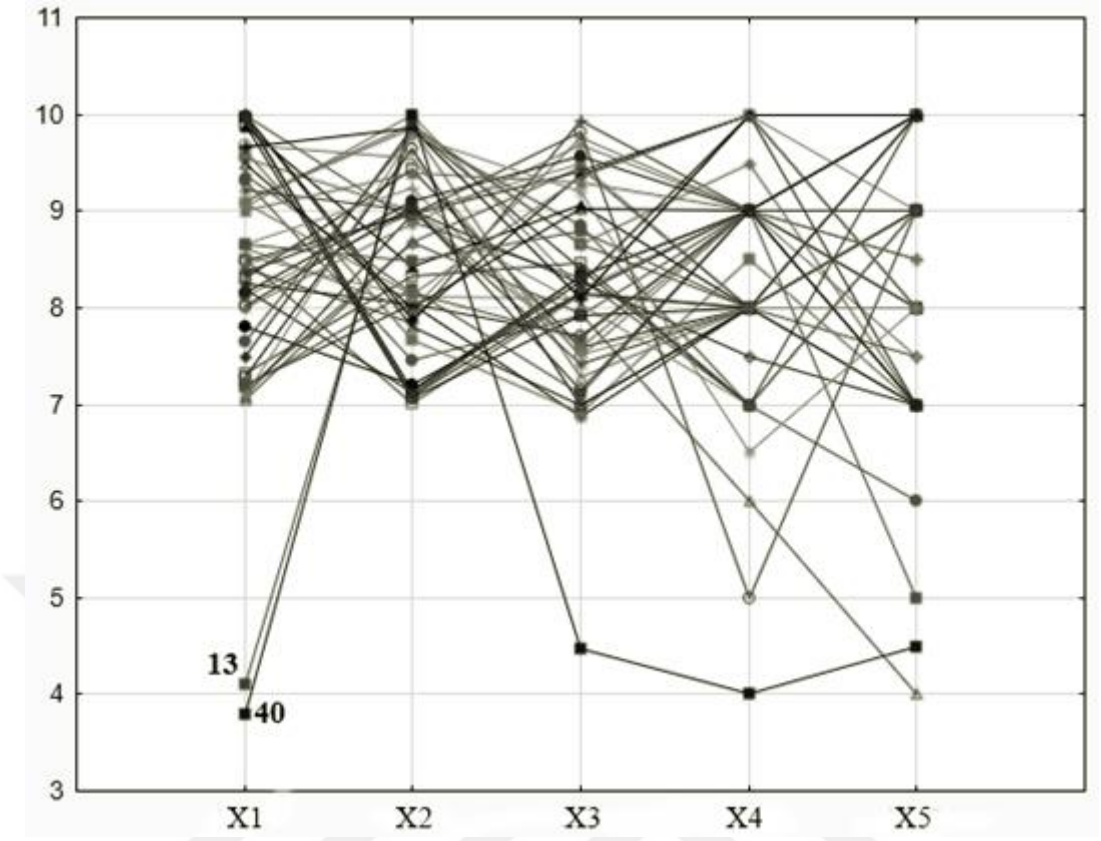


Grafik 3.8. Tablo 3.1'deki 5 Değişkene İlişkin Güneş Işığı Grafiği (STATISTICA)

Tablo 3.1'deki verilere ilişkin tüm gözlem değerlerine ait Profil Grafiği, Grafik 3.9'da verilmiştir. 13. ve 40. gözlemlerin diğer gözlem değerlerinden farklı bir yapıda olduğu görülmektedir. Aynı verilere ilişkin diğer bir sunum ise, doğrular yerine çubukların kullanılması ile oluşturulan çubuk-profil grafikleridir. Bu grafik de Grafik 3.10'dur.

STATISTICA programı ile çizilen bu grafiklerde, Grafik 3.9'u oluşturabilmek için, Graphs>2D Graphs>Line Plots (Case Profiles) adımları izlenerek değişkenler seçilir. Graph Type'dan Multiple seçilir. Grafik 3.10'u oluşturabilmek için de Graphs>Icon Plots adımları izlenilir. Daha sonra Variables komutuyla değişkenler atanır ve en son Graph Type tıklandıktan sonra Columns seçeneği tıklanır.

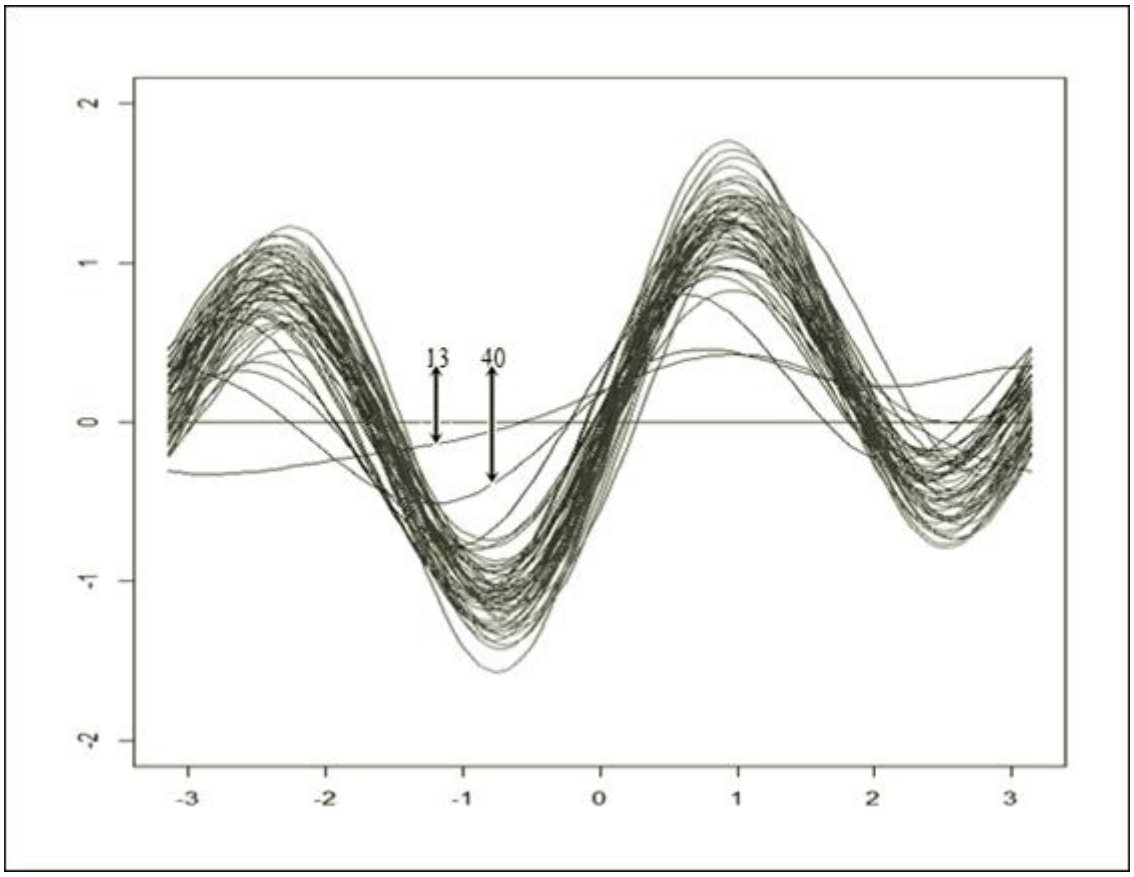
Grafik 3.9 ve grafik 3.10'da görüldüğü gibi gözlemlerden 13. ve 40. gözlem değerleri benzer yapıda iken, bu gözlemler diğerlerinden farklılık göstermektedir. Bu iki grafik birbirinin benzeri sonuçlar vermektedir. Bu nedenle, ikisinden birini seçmemiz bizi aynı sonuca götürecektir.



Grafik 3.9. Tablo 3.1'deki 5 Değişkene İlişkin Profil Grafiği

Grafik 3.10. Tablo 3.1'deki 5 Değişkene İlişkin Çubuk-Profil Grafiği

Tablo 3.1'deki 50 gözleme ilişkin 5 değişken değerlerini Andrews eğrileri çizerek oluşturduğumuzda grafiğimiz aşağıdaki gibi şekillenecektir:



Grafik 3.11. Tablo 3.1'deki Verilere İlişkin Andrews Grafiği (R Studio)

Daha önceki grafik çizimlerinde de kullandığımız Tablo 3.1 değerlerinde karşımıza çıkan 13. ve 40. gözlemlerdeki farklılık yani aykırı değerlik bu grafikte de belirgin olarak karşımıza çıkmaktadır. Bu grafik türünde eğriler oluşturulurken dikkat edilmesi gereken husus değişkenlerin benzer birimlerde olmasıdır. Değişkenlerin farklı birimlerde olması durumunda verilerin standartlaştırılması yoluna gidilecektir.

Grafiği oluşturmak için R Studio programına veriler matris şeklinde girilir ve daha sonra excel olarak atılan dosya ismiyle komut yazılarak datalar çağrılır. Örneğin; veri ismi x ise data(x) yazılır ve enter tıklanarak veriler çağrılır. Daha sonra andrews(x,clr=5,ymax=3) şeklinde komut yazılarak elde etmek istediğimiz grafiğe erişmiş oluruz.

Uygulama 3.2. Dünyadaki en iyi turizm şehirlerini belirlemek için yapılan bir çalışmada, rasgele Van şehri belirlenmiştir. Bu şehir için de rasgele 5 tane değişken seçilmiştir. Bu değişkenlerin standartlaştırılmış değerleri de aşağıdaki gibidir:

X1: Ülkelere gitmek için vize başvurusunda bulunup vizesi kabul edilenlerin genel içindeki payı,

X2: Ülkeleri ziyaret edenlerin yıllık yaptıkları harcamaların ortalaması,

X3: Gidenlerin aylık gelirleri,

X4: Seyahat edenlerin yaş ortalaması,

X5: Deniz turizminin yapılan tüm turizm içindeki yüzdesi,

Tablo 3.2. Van Şehri İçin Seçilen Kişilere İlişkin Değişken Değerleri

	X1	**X2**	**X3**	**X4**	**X5**
1	0,8	1,66	3,5	44,8	0,15
2	0,65	1,82	3,65	38,7	0,26
3	0,72	2,22	4,2	36,5	0,18
4	0,87	2,16	4	29,4	0,6
5	0,56	2,44	4,8	56,2	0,52
6	0,63	2,88	4,2	48,2	0,4
7	0,9	1,55	5,2	38,8	0,38
8	0,84	1,78	4	42	0,8
9	0,66	2,98	4,4	56	0,47
10	0,7	2,72	2,8	42,6	0,56
11	0,75	3,5	4,3	21,2	0,68
12	0,91	3,42	4,6	26,5	0,92
13	0,69	2,56	3,9	41	0,88
14	0,77	2,54	3,4	40,3	0,65
15	0,84	3,12	3,2	65,4	0,54
16	0,75	3,56	3,5	26,6	0,45
17	0,82	4,08	3,85	25,66	0,6
18	0,55	0,96	2,95	48,12	0,5
19	0,33	0,88	2,75	47,36	0,58
20	0,45	1,1	3	28,95	0,8
21	0,54	1,23	3,1	43,33	0,75
22	0,65	1,95	3,5	40,5	0,76
23	0,5	2,3	3,45	27,6	0,25
24	0,46	1,87	3,86	29,1	0,48
25	0,88	3,21	4,25	33,33	0,62
26	0,96	3,48	4,1	30,2	0,54
27	0,6	1,73	4	36	0,36
28	0,52	1,05	3,6	41	0,74
29	0,86	3,15	3,95	25,1	0,86
30	0,85	3,88	5,1	28,4	0,89

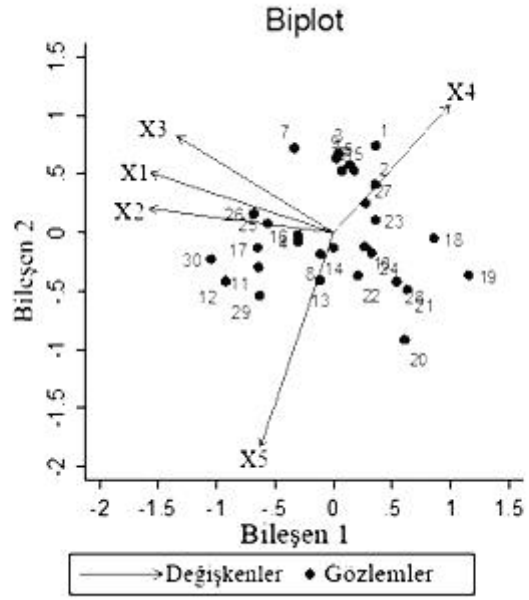
Tablo 3.3. Van Şehri İçin Seçilen Kişilere İlişkin Standartlaştırılmış Değişken Değerleri

	X1	X2	X3	X4	X5
1	0,61823	-0,79687	-0,53376	0,64797	-1,9879
2	-0,31222	-0,62285	-0,29618	0,06996	-1,47014
3	0,12199	-0,1878	0,57494	-0,1385	-1,84669
4	1,05244	-0,25306	0,25817	-0,81126	0,13023
5	-0,87049	0,05148	1,52527	1,72818	-0,24633
6	-0,43628	0,53004	0,57494	0,97013	-0,81116
7	1,23853	-0,91651	2,15881	0,07944	-0,9053
8	0,86635	-0,66635	0,25817	0,38265	1,07162
9	-0,25019	0,6388	0,89172	1,70922	-0,48168
10	-0,00207	0,35602	-1,64247	0,43951	-0,05805
11	0,30808	1,20437	0,73333	-1,58825	0,50678
12	1,30056	1,11736	1,20849	-1,08605	1,63645
13	-0,0641	0,182	0,09978	0,2879	1,44817
14	0,43214	0,16024	-0,69215	0,22157	0,36557
15	0,86635	0,79107	-1,00893	2,59992	-0,15219
16	0,30808	1,26963	-0,53376	-1,07658	-0,57582
17	0,74229	1,8352	0,02059	-1,16565	0,13023
18	-0,93252	-1,55821	-1,40489	0,96255	-0,34047
19	-2,29717	-1,64522	-1,72167	0,89054	0,03609
20	-1,55281	-1,40594	-1,3257	-0,8539	1,07162
21	-0,99455	-1,26455	-1,16731	0,50868	0,83627
22	-0,31222	-0,48146	-0,53376	0,24052	0,88334
23	-1,24266	-0,10079	-0,61296	-0,98182	-1,51721
24	-1,49078	-0,56847	0,03643	-0,83969	-0,43461
25	1,11447	0,88896	0,65414	-0,43887	0,22436
26	1,61071	1,18262	0,41656	-0,73546	-0,15219
27	-0,62237	-0,72074	0,25817	-0,18588	-0,99944
28	-1,1186	-1,46033	-0,37538	0,2879	0,7892
29	0,99041	0,8237	0,17898	-1,21871	1,35403
30	0,92838	1,61767	2,00043	-0,90602	1,49524

Tablo 3.3 verisine ait korelasyon matrisi Tablo 3.4'teki gibidir. Buna ilişkin Biplot grafiği ise Grafik 3.12'deki gibidir. Bu grafikler STATA programı ile çizilmiştir.

Tablo 3.4. Tablo 3.3. Verisinin Korelasyon Matrisi

Değişkenler	X1	X2	X3	X4	X5
X1	1	0,644	0,508	-0,236	0,146
X2		1	0,443	-0,319	0,183
X3			1	-0,198	0,053
X4				1	-0,216
X5					1



Grafik 3.12. Tablo 3.3'teki Verilere İlişkin Biplot Grafiği

Oluşturulan Biplot grafiği hakkında yapacağımız yorumlar:

1. Grafiğe bakıldığı zaman X3 ile X4 arasındaki açı 90 derecenin üzerindedir. Yani $r_{X2,X3}<0$'dır. Bunu, yukarıda elde etmiş olduğumuz korelasyon matrisinde de görebiliriz.

2. Hiçbir vektörün ucuna yakın gözlem değeri yoktur.

3. 9, 3, 15, 5, 6 ve 1 nolu gözlemler X4 (Seyahat Edenlerin Yaş Ortalaması) ile benzerlikleri olduğu için bir arada verilmiştir.

4. X1 ile X3 arasındaki korelasyon X5 ile X3 arasındaki korelasyondan daha düşüktür.

5. Grafikte birbirine yakın olan gözlem değerleri benzerlik itibariyle yanyana görüntülenmektedir.

6. Gözlemler merekeze yaklaştıkça özellikle 16 ya da 4 değerine bakacak olursak gözlemler X2'ye de X5'e de yakındır. Bu yüzden, tam anlamıyla X2 ile ilgilidir söyleyemeyiz.

7. 26 nolu gözlem 90 dereceden küçük olan iki eğri arasında kaldığından negatif korelasyona sahiptir. Tersi bir durum için de 23 nolu değişken 90 derecenin üzerindeki iki açı arasında kaldığından yüksek korelasyona sahiptir.

Uygulama 3.3. Türkiye'deki 81 il arasından alınan 30 il için, 5 değişkene ilişkin değerler aşağıdaki gibidir. Veriler, Türkiye İstatistik Kurumu veri bankasından alınmıştır.

X1: İstihdam oranı (%)

X2: İşsizlik oranı (%)

X3: Ortalama günlük kazanç (TL)

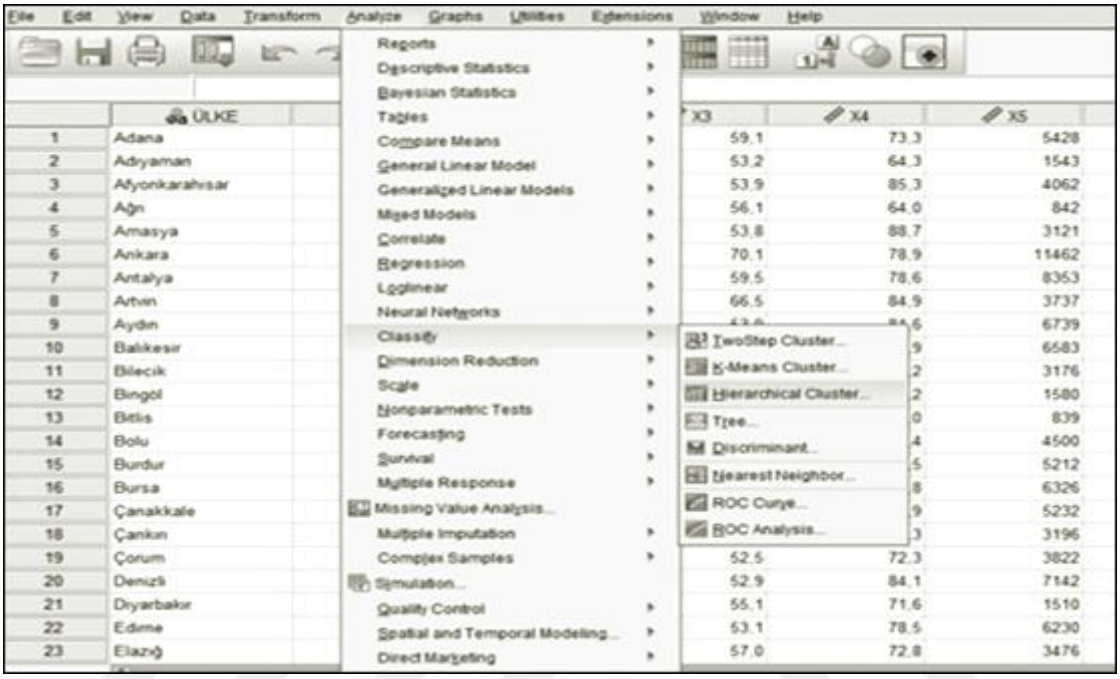
X4: İşinden memnuniyet oranı (%)

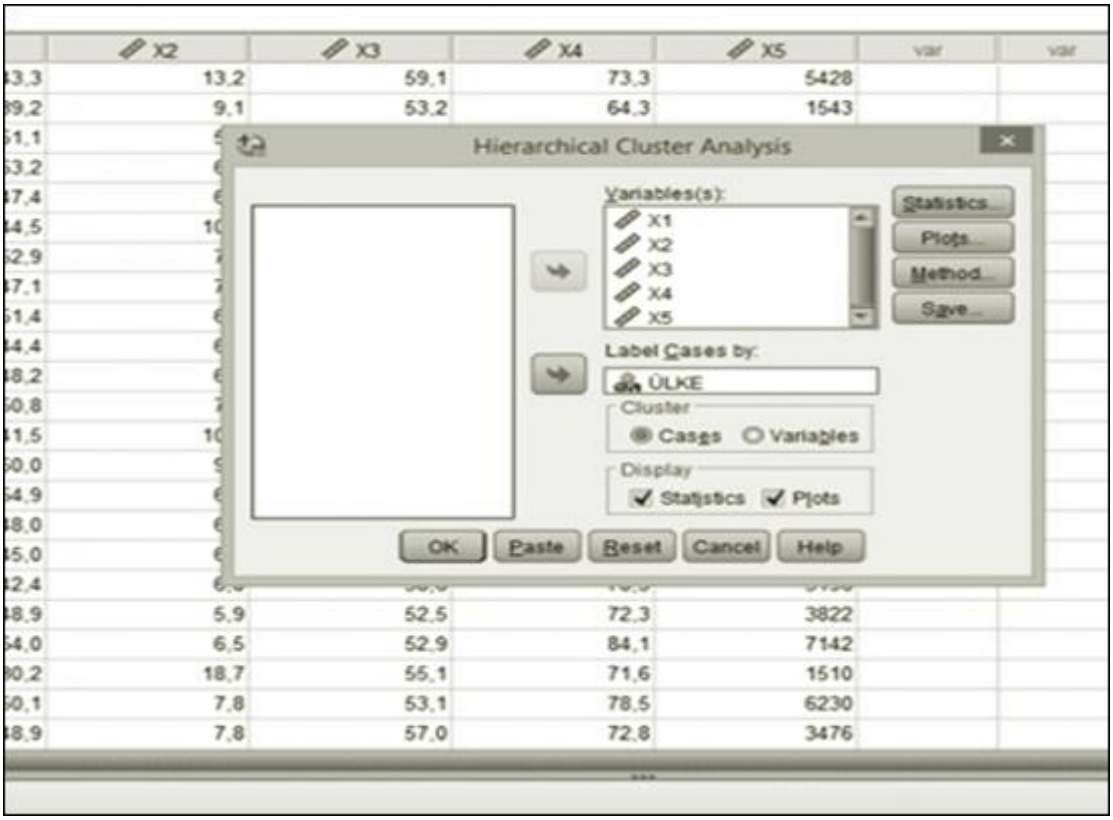
X5: Kişi başına düşen tasarruf mevduatı (TL)

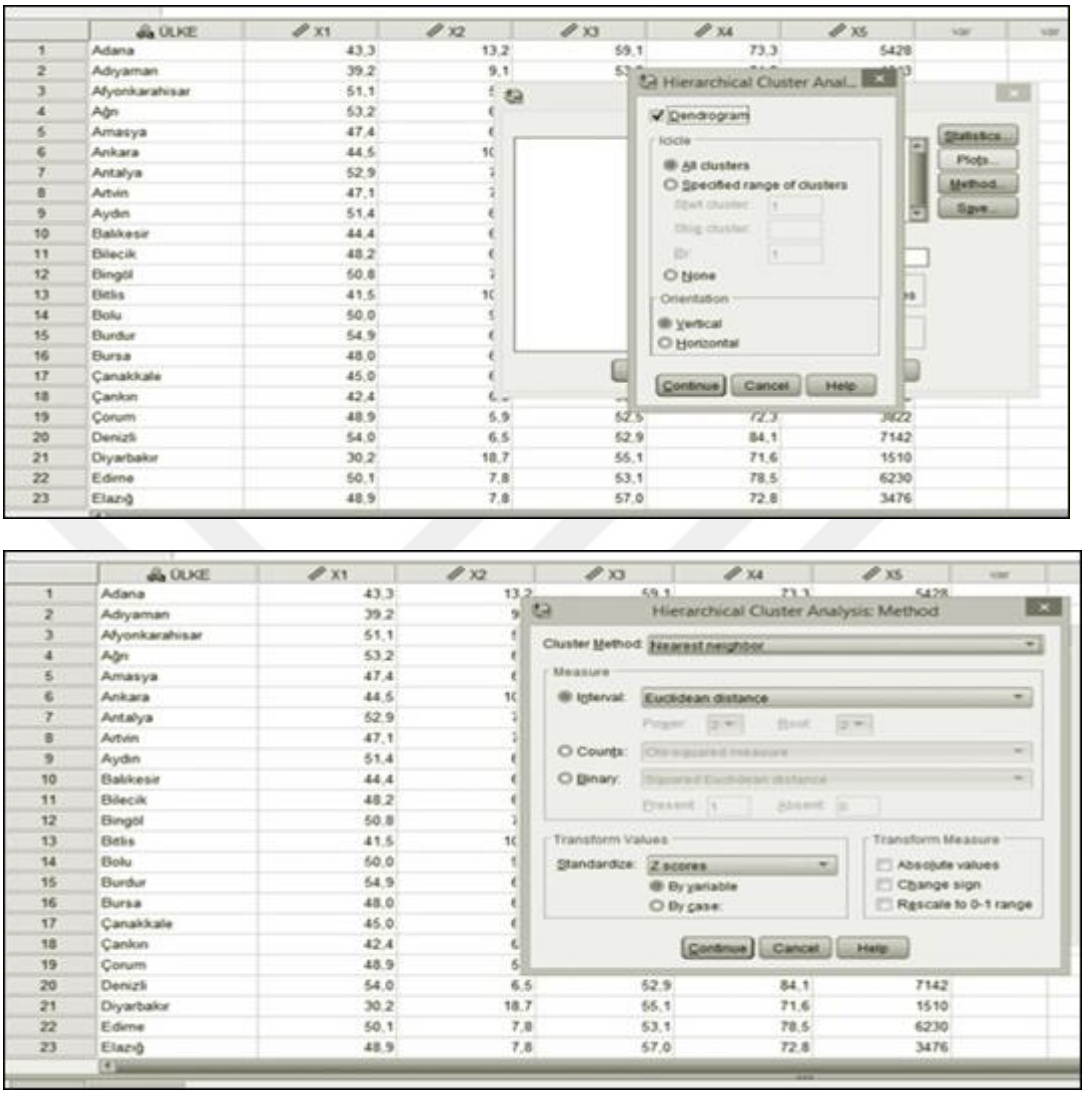
Tablo 3.5. 30 İle Ait 5 Değişken Değerlerine İlişkin İş-Gelir Verileri

İller	X1	X2	X3	X4	X5
Adana	43,3	13,2	59,1	73,3	5 428
Adıyaman	39,2	9,1	53,2	64,3	1 543
Afyon	51,1	5,6	53,9	85,3	4 062
Ağrı	53,2	6,8	56,1	64,0	842
Amasya	47,4	6,6	53,8	88,7	3 121
Ankara	44,5	10,2	70,1	78,9	11 462
Antalya	52,9	7,9	59,5	78,6	8 353
Artvin	47,1	7,1	66,5	84,9	3 737
Aydın	51,4	6,9	53,9	81,6	6 739
Balıkesir	44,4	6,0	56,0	86,9	6 583
Bilecik	48,2	6,5	67,1	85,2	3 176
Bingöl	50,8	7,0	54,2	71,2	1 580
Bitlis	41,5	10,6	54,9	71,0	839
Bolu	50,0	9,6	58,2	85,4	4 500
Burdur	54,9	6,9	55,5	82,5	5 212
Bursa	48,0	6,6	64,1	82,8	6 326
Çanakkale	45,0	6,1	58,0	81,9	5 232
Çankırı	42,4	6,8	58,6	78,3	3 196
Çorum	48,9	5,9	52,5	72,3	3 822
Denizli	54,0	6,5	52,9	84,1	7 142
Diyarbakır	30,2	18,7	55,1	71,6	1 510
Edirne	50,1	7,8	53,1	78,5	6 230
Elazığ	48,9	7,8	57,0	72,8	3 476
Erzincan	47,7	6,7	63,3	79,8	4 627
Erzurum	46,7	6,6	59,9	79,2	1 801
Eskişehir	42,9	8,5	66,8	80,0	6 570
Gaziantep	43,6	6,9	52,7	73,6	2 468
Giresun	47,9	6,5	49,1	82,3	4 701
Gümüşhane	46,3	7,2	59,1	82,7	2 668
Hakkari	39,9	11,7	58,3	72,1	689

Verilerin ölçüm birimleri farklı olduğundan z standartlaştırılması uygulanır. Örnekte veri sayısı az olduğundan aşamalı kümeleme yöntemleri kullanılır. Öklit uzaklık ölçüsü ve SLINK (en yakın komşuluk) yöntemlerine göre kümeleme analizi aşamaları;







İlk olarak bağımsız değişkenler Variable kutusuna atanır, ardından string değişken Label Case alanına aktarılır. Metod butonuna basarak yöntemler belirlenir. Bu aşamada en yakın komşuluk, 'Nearest neighboor' ve uzaklık yöntemi olarak 'Euclidean Distance' seçilir. Analiz verilerinin ölçüm biçimleri farklı olduğundan analiz öncesi z standardizasyonu yapılması istenir.

Kümeleme analizi sonuçlarını görüntülemek için plot menüsünden açılan pencerede 'dendogram' kutucuğu işaretlenir, grafiğin yatay veya dikey olması tercihe bağlıdır.

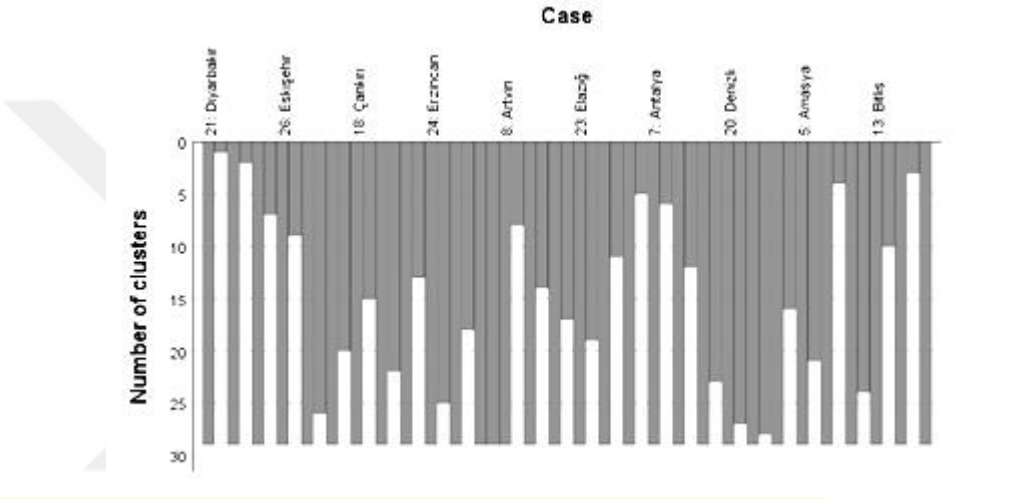
Seçilen algoritmalar sonucu elde edilen dendogram grafiği ile,

Tablo 3.6. Öklid Uzaklığı Değeri

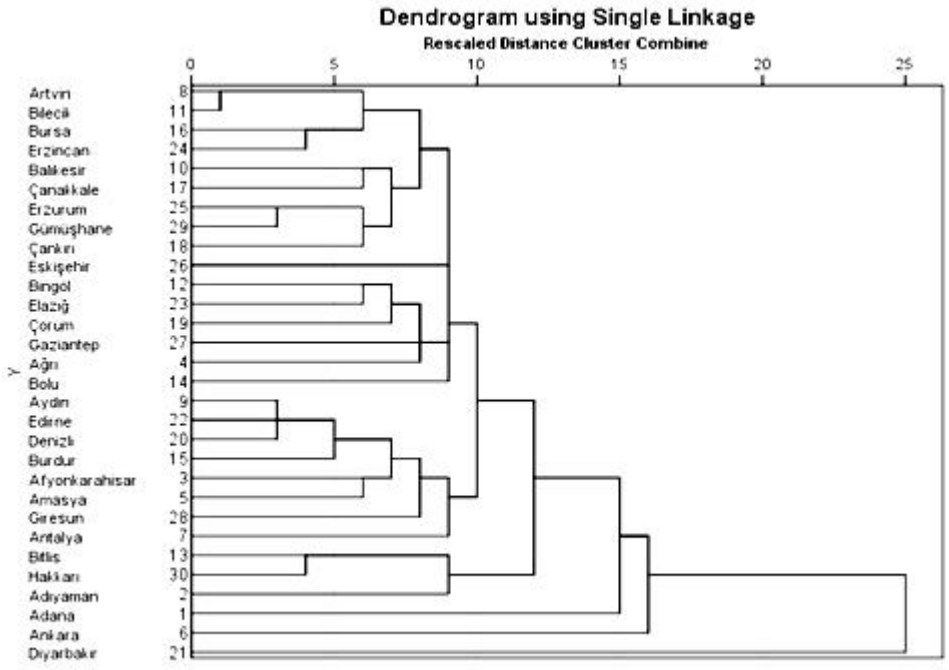
Case Processing Summary[a]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
30	81,1%	7	18,9%	37	100,0%

a. Euclidean Distance used



Grafik 3.13. 30 Şehre Ait İş-Gelir Değişkenlerine İlişkin Buz Saçağı Grafiği



Grafik 3.14. 30 Şehre Ait İş-Gelir Değişkenlerine İlişkin Ağaç Diyagramı Grafiği

30 şehrin genel olarak 5 kümede toplandıkları görülmektedir. Diyarbakır'ın tek bir küme, Ankara ve Adana'nın ikinci grubu, Antalya, Giresun, Amasya, Afyonkarahisar, Burdur, Denizli, Edirne ve Aydın'ın üçüncü grubu, Adıyaman, Hakkâri ve Bitlis'ın dördüncü grubu, Bolu ve yukarıya doğru olan illerin ise, beşinci grubu oluşturdukları görülmektedir.

Bir dendrogramı yorumlamanın anahtarı, iki nesnenin bir araya geldiği yüksekliğe odaklanmaktır. Gözlemlere ilişkin bir örnek verecek olursak, Bitlis ve Adıyaman'ın en çok benzer olduğunu görebiliriz, çünkü onları birleştiren bağlantının yüksekliği aynıdır. Yukarıdaki dendrogramda, dendrogramın yüksekliği kümelerin birleştirildiği sırayı gösterir. Yüksekliklerin kümeler arasındaki mesafeyi yansıttığı, daha bilgilendirici bir dendrogram oluşturulabilir.

KAYNAKÇA

Alkan, Bilal Barış, (2011) ‘‘*Çok Değişkenli İstatistiksel Yöntemlerde Biplot Tekniği’’*, Doktora Tezi, Ankara Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Alpar, Reha, (2017), *‘‘Uygulamalı Çok Değişkenli İstatistiksel Yöntemler’’*, (Beşinci Baskı), Detay Yayıncılık, Ankara.

Bilgin, T. Tugay, Çamurcu, A. Yılmaz, (2008), *‘‘Çok Boyutlu Veri Görselleştirme Teknikleri’’*, Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, 30 Ocak - 01 Şubat 2008 Maltepe Üniversitesi, Bilgisayar Mühendisliği Bölümü, İstanbul - Marmara Üniversitesi, Bilgisayar ve Kontrol Eğitimi Bölümü, İstanbul.

Üçkardeş, F., Şahinler, S., Ercan, E., *‘‘Aykırı Gözlemlerin Belirlenmesinde Kullanılan Bazı İstatistikler’’*, KSÜ Doğa Bilimleri Dergisi, 2010/13(1), 42-45.

Çok Değişkenli Veri Analizine Giriş, <https://math.illinoisstate.edu/day/courses/old/312/notes/twovar/twovar01.html> (11.03.2019)

Chernoff Faces, <http://de.wikipedia.org/w/index.php?title=Datei:Chernoff_faces_construction.svg> (11.01.2017)

Chernoff, H., *“The Use of Faces to Represent Points in K-Dimensional Space Graphically”*, Journal of the American Statistical Association, Vol. 68, No. 342, p. 361-368, June, 1973

Gharibnezhad, F., Mujica, L., Rodellar, J., (2011), *‘‘Damage detection using Andrew plots’’*, Universitat Politecnica de Catalunya, Spain.

Group 3 Hedwig Höller, Matthias Eichhaber, Thomas Nuschy, and Christof Steinkellner, (2013) *‘‘Chernoff Faces’’*, Graz University of Technology, Austria.

Hardle, W., Simar, L., (2003), *''Applied Multivariate Statistical Analysis''*, (22th. Edition), Berlin.

K Boyutlu Uzayda Noktaları Temsil Eden Yüzlerin Grafiksel Olarak Kullanılması, <http://dx.doi.org/10.2307/2284077 > (12.04.2017)

Khattree, R., Naik, N., *''Andrews plots for multivariate data: some new suggestions and applications, Journal of Statistical Planning and Inference 100''* , (2002), ss. 411–425.

Koğar, Hakan, (2010), *''Farklı Örneklem Büyüklüklerinde Uç Değerlerle Baş Etme Yöntemlerinin Puanlarının Geçerlik ve Güvenirlik Kanıtları Üzerindeki Etkisi''*, Yüksek Lisans Tezi, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Michael Friendly, (2006), *''A brief history of data visualization''*, Psychology Department and Statistical Consulting Service York University, Canada.

Path Analizi ve Kümeleme Analizi İncelemesi, <http://bssupgrade.oceaninfo.ru/library/files/39503.pdf > (17.05.2017)

Reyes Núñez, José Jesús, (2009), *''Ideas For The Use Of Chernoff Faces In School Cartography''*, Eötvös Loránd University Department of Cartography and Geoinformatics, Budapest, Hungary.

Rıchard A. Johnson, Dean W. Wıchern, (2007), *''Applied Multivariate Statistical Analysis''*, (Sixth. Edition), Pearson Prentice Hall Pearson Education Inc., U.S.A.

Sarı, İsmet Kürşat, (2012), *''Karma Ayrıştırma Analizinde Kayıp Gözlem Tahmin Yöntemlerinin Değerlendirilmesi''*, Yüksek Lisans Tezi, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Konya.

Vatansever, Metin, (2008), *''Görsel Veri Madenciliği Tekniklerinin Kümeleme Analizinde Kullanımı ve Uygulanması''*, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Winnie Wing-Yi Chan, (2006), *''A Survey on Multivariate Data Visualization''*, Department of Computer Science and Engineering Hong Kong University of Science and Technology Clear Water Bay, Kowloon, Hong Kong.