T. C.

ORDU ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

KAYIP VERİ İLE BAŞ ETME YÖNTEMLERİ

FATİH KALE

YÜKSEK LİSANS TEZİ

ZOOTEKNİ ANABİLİM DALI

ORDU 2020

TEZ BİLDİRİMİ

Tez yazım kurallarına uygun olarak hazırlanan ve kullanılan intihal tespit programının sonuçlarına göre; bu tezin yazılmasında bilimsel ahlak kurallarına uyulduğunu, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, tezin içerdiği yenilik ve sonuçların başka bir yerden alınmadığını, kullanılan verilerde herhangi bir tahrifat yapılmadığını, tezin herhangi bir kısmının bu üniversite veya başka bir üniversitedeki başka bir tez çalışması olarak sunulmadığını beyan ederim.

FATİH KALE

Not: Bu tezde kullanılan özgün ve başka kaynaktan yapılan bildirişlerin, çizelge, şekil ve fotoğrafların kaynak gösterilmeden kullanımı, 5846 sayılı Fikir ve Sanat Eserleri Kanunundaki hükümlere tabidir.

ÖZET

KAYIP VERİ İLE BAŞ ETME YÖNTEMLERİ

FATİH KALE

ORDU ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

ZOOTEKNİ ANABİLİM DALI

YÜKSEK LİSANS, 63 SAYFA

(TEZ DANIŞMANI: DR. ÖĞR. ÜYESİ YELİZ KAŞKO ARICI)

Araştırmalarda üzerinde çalışılan değişken(ler) bakımından örneklemi oluşturan en az bir deney ünitesinden planlanan veri elde edilememiş ise "kayıp veri" sorunu ortaya çıkmaktadır. Veriyi elde etme yöntemi, zaman yetersizliği, çalışılan deney ünitelerinde meydana gelen kayıplar gibi çok çeşitli sebepler ile kayıp veri sorunu meydana gelebilmektedir. Sonuçlarını etkilemesi sebebiyle kayıp verilerin varlığı istatistik analiz içeren araştırmalarda önemli bir husustur. Özellikle boylamsal verilerin elde edildiği çalışmalarda bu sorun ile başedilmesi daha büyük önem taşımaktadır.

Araştırmalarda karşılaşılan kayıp veri sorunu ile başetmek amacıyla günümüze kadar farklı yakşalımlar benimsenmiş ve çeşitli yöntemler geliştirilmiştir. Bu tez çalışmasında kayıp veri sorunu, kayıp veri mekanizmaları ve kayıp veri ile başetme yöntemlerinin detaylı olarak anlatılması amaçlanmıştır. Farklı istatistik programlarda uygulama örnekleri hazırlananmış olup bu tez çalışmasının kayıp veri sorunu yaşayan araştırıcılara ışık tutması beklenmektedir.

Anahtar Kelimeler: Kayıp veri, Yaklaşık değer, Veri atama yöntemleri, Rassal Olarak Kayıp

ABSTRACT

MISSING DATA SOLUTION METHODS

FATİH KALE

ORDU UNIVERSITY INSTITUTE OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF ANIMAL SCIENCE

MASTER THESIS, 63 PAGES

(SUPERVISOR: ASSIST. PROF. DR. YELİZ KAŞKO ARICI)

If the planned data could not be obtained from at least one experimental unit that constitutes the sample in terms of the variable(s) studied in the studies, the problem of "missing data" arises. This problem can occur due to various reasons such as method of obtaining data, lack of time, losses in the experiment units studied. The presence of missing data is an important consideration in research involving statistical analysis because it affects the results. Especially in studies where longitudinal data are obtained, it is more important to deal with this problem.

Different approaches have been adopted and various methods have been developed to deal with the missing data problem encountered in the researches. In this thesis, it is aimed to describe the missing data problem, the missing data mechanisms and the missing value solution methods in detail. Application examples have been prepared in different statistical programs and it is expected to shed light on researchers who have missing data problems.

Keywords: Missing value, Approximate value, Data assignment methods, Randomly Lost

TEŞEKKÜR

Tez konumun belirlenmesi, çalışmanın yürütülmesi ve yazımı esnasında başta danışman hocam Sayın Dr. Öğr. Üyesi Yeliz KAŞKO ARICI 'ya, Zootekni bölüm başkanım Sayın Prof. Dr. Sezai ALKAN'a, tezimin değerlendirmesi aşamasında önemli katkılar sağlayan jüri üyesi Sayın Dr. Öğr. Üyesi Yasemin GEDİK'e ve tez yazım aşamasında manevi desteklerini esirgemeyen İlçe Müdürüm Sayın Erkan AYAZ'a teşekkür ederim.

Aynı zamanda, manevi desteklerini her an üzerimde hissettiğim babam İbrahim KALE, annem Hamide KALE, kardeşim Hakan KALE, oğullarım Ahmet Akif KALE ve Yusuf Eren KALE'ye tüm kalbimle teşekkürü bir borç bilirim.

İÇİNDEKİLER

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

SİMGELER ve KISALTMALAR LİSTESİ

Euclid(d)	:	Her Eksik Veri İçeren Satır İçin Öklid Uzaklığı
M	:	Ataması Yapılmış ve Analiz Edilmiş Küme Sayısı
v	:	Nokta Tahmini İçin Varyans Tahmini
$\mathbf{V_i}$	:	Analiz Edilmiş i. Kümeden Varyans Tahmini
$\overline{\boldsymbol{Q}}$	:	Çoklu Atamada Monte Carlo Tekniği
$\widehat{\boldsymbol{Q}}_\mathbf{i}$	:	Analiz Edilmiş i. Kümeden Tahmin
TROK	:	Tamamen Rassal Olan Kayıp
ROK	:	Rassal Olan Kayıp
MAR	:	Missing at Random
MNAR	:	Missing Not at Random
$\mathbf{X_{ij}}$	:	i. Durumun j. Değişkeni
$\mathbf{X_i}$	:	Tamamlanmış Veri Kümesi Matrisi
$\mathbf{Y_{ij}}$	:	i. Durumun j. Değişkeni
$\mathbf{Y_i}$	:	Tamamlanmamış Veri Kümesi Matrisi
NA	:	Not Available (Kayıp Verinin R Programında Gösterimi)
NI	:	İhmal Edilemez Kayıp (Non-Ignorable)

1. GİRİŞ

Araştırmalarda, üzerinde çalışılan değişkenler bakımından örneklemi oluşturan tüm deney ünitelerinden yapılan veri toplama işlemi tamamlanabilmiş ise elde edilen veri seti kayıpsız yada tam (tamamlanmış) veri seti olarak adlandırılır. Ancak en az bir deney ünitesinden veri elde edilememiş ise o veri setine kayıp veri içeren veya tamamlanamamış veri seti denilmektedir. Veri setlerinde bulunan eksik gözlem, eksik veri, kayıp gözlem, tamamlanamamış gözlem vb. olarak da adlandırılabilen "kayıp veri" istatistik analiz içeren araştırmalarda önemli bir husustur. Verinin elde edilme yöntemi, araştırma süresinin yetersizliği, çalışılan deney ünitelerinde meydana gelen kayıplar vb. sebepler ile kayıp veri sorunu meydana gelebilmektedir. Çalışmayı yürüten araştırmacının veri derleme hataları ve çalışma yürütülen grubun kendinden veya çevresel etkilerden kaynaklı olarak da kayıp gözlem değerleri oluşabilmektedir. Sonuçlarını etkileyecek düzeyde kayıp veri varlığı istatistiksel analizlerin istenmeyen faktörlerinden biri olarak kabul edilmektedir.

Kayıp veri sorunuyla karşılaşıldığında ilk adım kayıp veri mekanizmasının belirlenmesi, ikinci adım ise kayıp veri miktarının belirlenmesidir. Kayıp veri oluşum mekanizmaları literatürde 3 ana sınıf altında toplanmıştır. Bunlar; Rassal Kayıp (Missing at Random-MAR), Tamamen Rassal Kayıp (Missing at completely at random–MCAR) ve Rastgele Olmayan Kayıptır (Missing not at random-MNAR) (Rubin, 1976; Little ve Rubin, 1987). Büyük örneklemlerde tamamen rastlantısal kayıp veri oranı ≤%5 ise ciddi sorunlar ortaya çıkmamakta ve kayıp verilerin çözümünde kullanılan yöntemler kazıp veriler göz ardı edildiğinde elde edilen ile benzer sonuçlar verebilmektedir. Fakat küçük ve orta büyüklükteki örneklemlerde kayıp veri oranı arttıkça ciddi sorunlar ortaya çıkabilmektedir. Rastlantısal olmayan kayıp veriler sonuçların genellenebilirliğini etkilediğinden kayıp veri oranları az olsa bile tamamen rastlantısal kayıp verilere oranla daha önemli sorunlara sebep olabilmektedir (Tabachnick ve Fidell, 2001; Arıkan ve Soysal, 2018).

Kayıp veriyi ifade eden alanda yeni bir gözlem değeri oluşturulmamış, ikame bir değer atanmamış veya veri silme yöntemlerinden herhangi biri uygulanmamışsa, verilerin analizinde doğru sonuçlardan uzaklaşılabilme ihtimali bulunmaktadır. Kayıp veriler, analizlerde kullanılacak olan istatistiksel yöntemlerin hemen hemen hepsi için

önemli bir sorun oluşturur, çünkü tüm yöntemler veri setinin eksiksiz olduğu varsayımı altında gelişmiştir (Pigott, 2001; Allison, 2003; Osborne, 2013;). Kayıp veri varlığı ile tüm bilim dalarında karşılaşılabilmektedir. Örneğin sağlık alanında yapılan çalışmalarda örneklemi oluşturan grupta bazı hastaların takipden çıkması sorunu ile sıkça karşılaşılmaktadır. Çok sayıda denek üzerinde çalışma yapabilme imkanına sahip sosyal bilimlerde de katılımcıların her soruya cevap vermek istememesi kayıp veri sorununu önemli bir sorun haline getirmektedir. Özellikle büyük örneklemlerle çalışılan araştırmalarda eksiksiz veri setlerini oluşturmak mümkün olmamaktadır (Cool, 2000).

Kayıp veri miktarı, kalan verilerin analizlerdeki yeterliliği ve değişkenlerin dağılım şekli dikkate alındığında birden fazla kayıp veri değerlendirme yöntemine ihtiyaç duyulmaktadır. Kayıp veriyi tespit etmek ve kayıp gözlem değeri içeren veri setlerini analiz etmek için kullanılan yöntemler yıllar içerisinde değişim göstermekte olup geliştirilen yeni yöntemler ile kayıp veri sorununa yeni yaklaşımlar sunulmaktadır. Kayıp veri ile başetmek amacıyla çok sayıda yöntem geliştirilmiş ve farklı yaklaşımlar önerilmiştir. Veri analizlerinin yorumlanabilir düzeyde sonuçlar verebilmesi için kayıp veri değerlendirme yöntemleri seçilirken doğru yaklaşımların ortaya konulması gerekmektedir. Literatürde önerilen kayıp veri ile başetme yolları dört ana başlık altında incelenebilir. Bunlar; veri silme yöntemleri, tüm gözlem değerlerinin kullanılması yöntemi, gözlem değerlerinde yer değiştirme yöntemleri ve veri atama atama yöntemleridir.

Kayıp veri sorunu ile başetmede doğru yaklaşımı belirlemek çoğu zaman kolay olmamaktadır. Kayıp verinin yoğunluğunun az olduğu ve tamamen rassal olarak dağılım gösterdiği durumlarda basit veri atamaya dayalı yöntemler kullanılabilir. Kayıp olan veri grupları, analize dahil edilen diğer değişkenlere bağlı olduğu durumlarda yapılan veri silme işlemi sonuçlarda büyük yanlılığa neden olabilir (Tabachnick ve Fidell, 1996; Schafer, 1999; Osborne, 2013). Fakat diğer durumlarda çoklu atamayla bağlantılı yöntemlerin daha sağlıklı sonuçlar üreteceği bildirilmektedir (Schafer, 1999; Osborne, 2013). Boylamsal verilerin elde edildiği uzun yıllara dayalı yada farklı periyodlara sahip çalışmalarda çeşitli nedenlerden dolayı kayba uğramış verilerin yerine veri tahsisi genellikle mevcut veriler üzerinden yeni bir modelleme

oluşturularak, yani veri tahmini veya veri silme yöntemlerinden herhangi biri ile analizler gerçekleştirilebilmektedir.

Kayıp verilere sahip deney ünitelerinin istatistik analizlerde analiz dışı bırakılması da sıklıkla kullanılan bir yöntemdir ve istatistik programlar bunu otomatik olarak yapmaktadır. Düşük oranlarda kayıp veri içeren veri setlerinde (yaklaşık olarak ≤%5) kayıp verilerin analiz dışı bırakılmasının sonuçların güvenilirliğini etkileyemeceği ancak kayıp veri miktarının fazla olması durumunda kayıp verileri silmenin önemli bir bilgi kaybı oluşturabileceği bildirilmektedir. Bunun sebebi veri silme yaklaşımı ile örneklem genişliğinin küçülmesi ve araştırma sonuçlarının genellenebilirliğinin azalmasıdır. Bu durumda hipotez kontrollerinde II. Tip hatanın artması ve hedef testin gücü değerinin azalması söz konusu olacaktır (Schafer, 1997; Cool, 2000; Mertler ve Vannatta, 2005; Groves, 2006; Garson, 2015; Carpita ve Manisera, 2011; Soysal ve Akın Arıkan, 2017).

Bu tez çalışmasında kayıp veri sorunu, kayıp veri mekanizmaları ve kayıp veri ile başetme yöntemleri incelenmiştir. Yöntemlerin birbirinden farklı yönleri, avantajları, dezavantajları ve karmaşık yada basit veri gruplarında nasıl sonuçlar vereceği, hangi veri grubu için hangi yöntemin uygun olduğu detaylıca açıklanmıştır. Böylece araştırıcıların kayıp veri sorunu ile karşılaştıklarında başvurabilecekleri bir kaynağın hazırlanması amaçlanmaktadır. SPSS v26 ve R v3.6.2 istatistik programları ile uygulama örnekleri hazırlananmış olup bu tez çalışmasının kayıp veri sorunu yaşayan araştırıcılara ışık tutması beklenmektedir.

2. GENEL BİLGİLER

Araştırmalarda toplanması planlanan veriler, toplanabilen veri arasındaki farka kayıp veri adı verilmektedir (Longford, 2005). Literatürde kayıp verilerin oluşumu ve nedenlerine ait birçok tanımlama yapılmıştır. Ortaya çıkma şekline göre kayıp veriler madde yanıtlamama (item nonresponse), birim yanıtlamama (unit nonresponse), dalgalı kayıp (wave nonresponse) şeklinde adlandırılmaktadır (Akbaş ve Koğar, 2020). Madde yanıtlamama şeklindeki kayıplar araştırmaya dahil edilen bir bireyin en az bir verisinin elde edilemediği durumları ifade ederken, birim yanıtlamama şeklindeki kayıplar bir deney ünütesine ait ölçümlerin hiçbirine ulaşılamadığı durumları ifade etmektedir. Dalgalı ve dönüşsüz kayıplarla ise aynı deneklerden tekrarlanan ölçümlerinin alındığı boylamsal araştırmalarda karşılaşılmaktadır. Dalgalı kayıpta, katılımcılara bazı ölçüm zamanlarında ulaşılabilmekte bazılarında ise ulaşılamamaktadır. Verinin bir noktadan sonraki ölçüm noktaları için elde edilememesi ise dönüşsüz kayıp olarak adlandırılmaktadır (Graham, 2012).

Bir plan çerçevesinde ilerlerken araştırıcının müdahale edemediği birçok faktör sebebiyle kayıp veriler oluşabilmektedir. Ölçüm yöntemleri ve hataları, verinin bilgisayar ortamına aktarılmasındaki yazım hataları, hayvan deneylerindeki kayıplar, zirai çalışmalarda iklim ve çevresel faktörler, araştırma süresindeki kısıtlılıklar ve özellikle anket çalışmalarında soruların yanıtsız bırakılması, kurumsal alanda hizmet veren kurum kuruluşların belli yıllara ait tutulmayan kayıtları, doğru tutulmayan veya tahrip olan kayıtlar gibi sebepler kayıp verinin oluşum sebepleri arasında sıralanabilir.

Bazı istatistik analiz yöntemlerin yapılabilmesi verinin eksiksiz olması ile mümkün olabilmektedir. Özellikle tekrarlanan ölçümlerin analizinde kullanılan yöntemler ve çok değişkenli analiz yöntemleri için veri setinin tam olması önemli bir husustur. Ancak uzun yıllara dayalı büyük veri gruplarında yapılan çalışmalarda kayıp veri içermeyen örnek grupları oluşturmak neredeyse imkansız olmaktadır (Cool, 2000). Gözlem değerlerinden bir veya bir kaçının kaybolması, kayıp verili matrislerin oluşmasına neden olmaktadır. Bu da veri setlerinin satır ve sütun sayısının değişmesine ve istatistik analizlerde sorunlarla karşılaşılmasına neden olmaktadır (Little ve Rubin, 1987).

Peng ve ark. (2006), kayıp verilere bağlı olarak ortaya çıkabilecek problemleri dört ana madde altında özetlemektedir:

- İstatistiksel analizlerde yapılan kestirimlerde kayıp veriler sebebiyle yanlılık oluşabilir. Örneğin bir anket çalışmasında cevaplamayı reddedenlerin profilinin cevaplamayı Kabul edenlerin profilinden farklı olması örneklemin, randomization özelliğini kaybetmesine sebep olabilir.

- Kayıp veriler istatistiksel analizin gücünün azalmasına sebep olabilirler. Örneğin, örneklem genişliğinin azalması ile serbestlik derecesinin azalması dolayısıyla standart hata değerinin artması söz konusu olur.

- Kayıp veriler içeren veri setleri ile bazı istatistik testlerin yapılması mümkün değildir. Özellikle matrisler aracılığı ile yürütülen istatistiksel testler kayıp veriler ile analizin yapılmasını mümkün kılmaz. Örneğin, çok değişkenli analiz yöntemlerinde ve tekrarlanan ölçümlerin analizinde kullanılan testlerde kayıp verileri olan denekler analizlere dahil edilemezler.

Kayıp verinin oranı arttıkça kayıp veri sorunu tüm istatistiksel değerlendirmeler için önem kazanmaktadır. Bayhan (2018) R programında yazılan kodlarla rassal olarak kayıp veri mekanizmasına uygun olarak %5, %10 ve %20 oranında değer silinerek Cronbach α güvenirlik değerini hesaplandığı çalışmasında, Cronbach α değerindeki değişimin dağılım şekline bağlı olmaksızın örneklem büyüklüğü, testin uzunluğu ve dağılım biçimi farklı olsa dahi kayıp veri oranı arttıkça Cronbach α değerindeki değişim de arttığını bildirmiştir.

Araştırmalarda karşılaşılan kayıp veri sorunu ile başetmek amacıyla günümüze kadar farklı yaklaşımlar benimsenmiş ve çeşitli yöntemler geliştirilmiştir. Kayıp veri ile analize devam etmek için kayıp verileri analiz dışı bırakma veya çeşitli istatistiksel yöntemlerle kayıp verileri tamamlama gibi metodlar kullanılmaktadır (Kürşad ve Nartgün, 2016). Kayıp veriler ihmal edilerek yapılan analizlerde hatalı sonuçlar ortaya çıkabileceği ve bu sonuçların araştırılan populasyon için genelleştirildiğinde hata miktarının artış göstereceği düşünülmektedir (Byrne, 2000).

Kayıp veri analizlerinde; kayıp veri grubunu analiz dışı bırakma veya silme gibi yöntemlerin yanında, belirli istatistiksel yaklaşımlarla veri tahmin etme yöntemleriyle de çözümler üretilebilmektedir. İstatistiksel programlamalarda yer

bulan ve sıklıkla kullanılan yöntem kayıp verilerin analiz dışı bırakılmasıdır (Schafer, 1997; Garson, 2015). Bu yöntemin yoğun olarak kullanılması veri setinde %5 ve altındaki kayıp oranı dahilinde gerçekleştirilebilmektedir (Mertler ve Vannatta, 2005; Groves, 2006; Garson, 2015). Şayet kayıp veri oranı fazla ise veri silmenin bilgi kaybına neden olabileceği (Carpita ve Manisera, 2011), örneklem genişliğini önemli düzeyde azaltacağı, anket çalışması ise sorulara cevap veren bireyler arası sistematik olarak bir yanlılığa sebep olacağı (Schafer, 1997; Carpita ve Manisera, 2011), istatistiksel analizlerde testin gücünü azaltacağı ve dolayısıyla hesaplanacak p-değerinin olması gerekenden farklı çıkmasına sebep olacağı (Cool, 2000; Garson, 2015) bildirilmektedir.

Veri atama yöntemlerinin araştırmalarda tam verili matrislerin kullanılmasına izin vermesi ve çalışılan veri grubunun büyüklüğünü koruması nedeniyle diğer yöntemlere göre araştırmacılar tarafından kayıp veri analiz yöntemleri arasında daha çok tercih edilmektedir (Huisman, 2000; Fox-Waslylyshyn ve El-Masri, 2005). Eksik veri içermeyen tam veri setleri ile çalışmak istatistik analizlerin güvenilirliğini olumlu yönde etkileyen unsurların başında gelmektedir.

Kayıp veri mekanizmaları literatürde 3 ana sınıf altında toplanmıştır; Bunlar Rassal Kayıp (Missing At Random-MAR), Tamamen Rassal Kayıp (Missing At Completely At Random–MCAR) ve Rastgele Olmayan Kayıptır (Missing Not At Random-MNAR) (Rubin, 1976; Little ve Rubin, 1987). Kayıp veri mekanizmasının belirlenmesi, kayıp veri sorununu giderecek doğru analizin seçilmesinde, o verinin değerlendirilmesinde ve doğru sonucun elde edilmesinde etkin rol oynar (Baygül, 2007).

2.1 Kayıp Veri Çalışmalarının Tarihçesi

Kayıp veriler özellikle uzun yıllara dayalı boylamsal (longitudinal) çalışmalarda her zaman bir sorun olmuştur. Kayıp veriler ile çalışmalar Afifi ve Elashof (1966)'ın "Çok Değişkenli İstatistiklerde Eksik Gözlemler" adlı çalışması ile başlamıştır. Kayıp veri sorununun çözümü için temeller 1970'li yıllarda atılmıştır. Kayıp veri mekanizmaları ile ilgili çalışmayı ilk olarak Rubin (1976) yapmış ve kayıp veri mekanizmalarına yönelik yaptığı sınıflamadan sonra Dempster ve ark. (1977)'ın önerdiği beklenti maksimizasyonu ve Rubin'in (1987) önerdiği çoklu atama yöntemleri

kayıp veri ile başetme yollarının temellerini oluşturmuştur. Tamamen Rassal Kayıp, İhmal Edilemez Kayıp ve Rassal Kayıp olarak bu mekanizmaları sınıflandırma Little ve Rubin (1987) tarafından gerçekleştirilmiştir. Bu sınıflandırmaya göre hangi yöntemin hangi mekanizmaya uygun olduğu belirlenmiştir. Little ve Rubin (1987) tarafından yayınlanan "Kayıp Veriler ile İstatistiksel Analiz" adlı kitap ile kayıp veriler istatistikte bir araştırma alanı olmuştur.

Klinik çalışmalarda; Son Gözlem Değerini İleri Taşıma, Tamamlanmış Olgular Analizi gibi atama ve çoklu atama yöntemleri ile ilgili çalışmalar Heyting, Tolboom ve Essers (1992) tarafından yapılmıştır. Kayda değer bir düşüşle birlikte sürekli boylamsal veriler için bir model öneren Diggle ve Kenward (1994) ve Molenberg ve Verbeke (2005) Rastgele Olmayan Kayıp Veri mekanizması (MNAR) üzerine çalışmalarda bulunmuştur.

Kayıp veriler için yöntemler geliştirildikçe istatistik yazılımlarda yer bulmaya başlamış ve böylece kullanımları da yaygınlaşmıştır. Eberle ve Toutenburg (1999), kayıp veri analiz yöntemleri içeren Minitab, LogXact, SAS, SPSS, S-PLUS STATISTICA, StatXact, Stata, SYSTAT ve JMP istatistik programlarını birçok değerlerndirme ölçütü kullanarak incelenmiştir. İnceleme sonucunda hiçbir istatistik programın kayıp veri yöntemlerinin tamamını içermediği ve genellikle sınırlı sayıda yöntemin yer aldığı belirlenmiştir.

Kayıp veriler üzerine yöntemler geliştirildikçe istatistiksel açıdan yöntemlerin avantaj ve dezavantajları ortaya çıkmıştır. Cool (2000) veri silme yöntemlerinin örneklem genişliğini etkilemesi sebebiyle istatistiksel hata oranını arttıracağını, veri atama yöntemlerinin ise testlerin gücünü arttırdığını gösteren çalışmalar yapmıştır. Enders ve Bandalos (2001), liste bazında silme ve çift yönlü silme yöntemleri ile bazı atama yöntemlerini incelemişlerdir. Kayıp veri ile ilgili istatistiksel değerlendirmelerin yanlı ve tarafsız sonuçlara hangi yöntemlerle ulaşıldığı ile ilgili çalışmalar artmış olup, Allison (2001) geleneksel yöntemlere göre maksimum benzerlik ve çoklu veri atama yöntemlerinin daha başarılı olduğunu ortaya koymuştur.

Örneklemde bulunan katılımcıların ölçek maddelerinden bazılarına cevap vermemesi sonucunda ortaya çıkan kayıp veriler ölçek kullanılan araştırmalarda sık karşılaşılan bir sorundur. Peng ve ark. (2006) eğitim araştırmalarında kayıp veri

sorununu incelemek amacıyla 1998-2002 yılları arasında 1087 çalışmayı incelemişlerdir. Sonuç olarak, çalışmaların %54'sinde kayıp veri sorununun çözümüne yönelik bazı yöntemlerin kullanıldığı, %28'inde kayıp veri sorunundan bahsedilmediği ve %18'inde ise kayıp verilerden bahsedildiği ancak başetmek amacıyla herhangi bir yaklaşım sergilenmediği bildirilmiştir.

Horton ve Kleinman (2007) kayıp veri analiz yöntemlerinde istatistiksel yazılımların kullanılması gerektiğini vurgulamıştır. R paket programı ile ilgili çalışmaları Missztal (2012) ele almış olup olup, SAS ve STATA gibi istatistik programlarının kullanılabildiğini Soley Bori (2013) ortaya koymuştur. Bu istatistik programları genellikle çoklu atama yönteminin etkisini belirlemek için kullanılmıştır.

Ülkemizde Kayıp veriler ile ilgili çalışmalar 2000'li yıllarda başlamıştır. Oğuzlar (2001) kayıp veriler ile ilgili sorunları ve çözüm önerilerini ortaya koymuştur. Kayıp veri mekanizmalarının avantaj ve dezavantajları Bal (2003) tarafından değerlendirilmiş olup basit istatiksel değerlendirmelerin veriler içindeki değişimlerini incelemiştir. Alkan (2012) tam veri setleri ile yapılan analizler ile kayıp olan grubun analizlerini karşılaştırarak aralarındaki anlamlı farkı göstermiştir. Sezgin ve Çelik (2013) ise veri atama yöntemleri değerlendirilirken, mevcut yöntemlerin birbirine üstünlüğünün olmadığı, verilerin yapısal özelliklerine göre uygun yöntemin seçilmesi gerektiğini bildirmiştir.

Sıddıkoğlu (2019) gerçek veri parametreleri üzerinden benzetim çalışması kullanılarak gerçekleştirdiği çalışmasında kayıp veri sorununu verinin eksiksiz olduğu durumun standart ölçüt olarak kullanıldığı karşılaştırmalarla incelenmiştir. Çalışmada yanıt fonksiyonu ile kayıp veri yerine değer atama yöntemi ile parametrelere ilişkin güvenirlik ölçütleri tam veriden bulunanlara çok benzer bulunmuştur. Tahminlerinin güvenirliği açısından, yanıt fonksiyonu ile atama yapılan veri setleri kayıp veri setlerinden daha yüksek tahmin güvenirliğine sahip olduğunu bildirmiştir.

2.2 Kayıp Verinin Oluşumu

Little ve Rubin (1987) kayıp verinin oluşumunu üç kategoriye ayırmıştır. Bunlar;

- Kayıplık, gözlem değerinden bağımsız olarak gelişim gösteriyorsa, bu tamamen rastgele kayıptır.

- Bir bağımsız değişken diğer bağımsız değişken üzerinde bir baskı kurup etkiliyorsa mevcut değişkene bir etkisi bulunmuyorsa bu rassal kayıptır.

- Değişkenlerin dağılımda faktör olarak karşılıklı bir etkileşim söz konusu olduğunda ise bu ihmal edilemez kayıplık olarak kayıp veri mekanizması oluşum ve değerlendirme kriterleri arasında yer almaktadır.

Kayıp verilerin oluşum mekanizmalarının analizler de 3 önemli etkisi vardır. Bunlar;

- Analiz edilecek veri kümesindeki bireylerin aynı sayıda ölçümlere sahip olmaması veri setinde dengesizliklere yol açabilmektedir.

- Kayıp veri kümesi için eksik bilgi alma gibi olumsuzlukları oluşturacağından, kayıp veri ile gözlemlenen ölçümler arası hassasiyet artabilmektedir.

- Kayıp veriler ile ölçüm yapılan bireyler kümesindeki analizler yanıltıcı çıkarımlara neden olabilmektedir.

Bu sonuçlar dikkate alındığında kayıp veriler mevcut verilerden faydalanmayı asgari düzeye indirebilmektedir.

2.3 Kayıp Veri Mekanizmaları

Kayıp veri grubu için oluşturulan çözüm yöntemleri güvenilir sonuçlara ulaşmada önem arz etmektedir. Kayıp verinin teşkil ettiği veri gruplarındaki problemlerin çözümü için kayıp verinin hangi mekanizmaya dahil olduğunu belirlemek gerekmektedir. Little ve Rubin (2002) kayıp veri mekanizmalarını oluşum süreci ve oluşum şekline göre üç sınıfa ayırmıştır. Bunlar;

1. Tamamıyla Rassal Olarak Kayıp (TROK; Missing Completely at Random, MCAR)

2. Rassal Olarak Kayıp (ROK; Missing at Random, MAR)

3. İhmal Edilemez Kayıp (İEK; Noignorable, NI) olacak şekilde üç kategoriye ayırmıştır.

2.3.1 Tamamen Rassal Kayıp (MCAR)

Eğer kayıp veri mekanizmasında kayıp veri X ve Y gibi iki gözlem değeri gibi değişkenlerden oluştuğu varsayılırsa ve bu iki değişken birbirinden bağımsız olarak

meydana gelmiş ise bu Tamamen Rassal Kayıp olarak adlandırılmaktadır. Tamamen Rassal Kayıp ihtimale dayalı bilinmeyen parametrelere göre oluşur. Yani, şans faktörünün bu mekanizma da etkili bir rolü vardır (Sinharay ve ark., 2001). Kayıp veri mevcut veriler ile diğer kayıp verilerden tamamen bağımsız olarak oluşum göstermiş ise Tamamen Rastgele kayıplık özelliği gösteren bir yapıya sahip olabilmektedir.

TROK mekanizmasına dahil olmayan veriler yanlı ve taraflı sonuçlar üretmeye daha yakın olacağından, kayıp verilerin çözümünde daha güçlü yöntemlerin kullanılmasını zorunlu kılmıştır (Little ve Rubin, 2002; Alpar, 2003; Yazıcı, 2005). TROK yapısına dahil olan kayıp veri, kayıp veriye diğer değişkenlere ait ölçeklerden bağımsız olarak gelişen ve istek dışı ortaya çıkan bir gözlem değerindeki eksilme durumu olarak gerçekleşebilmektedir (Enders, 2011).

2.3.2 Rassal Kayıp (MAR)

X ve Y değişken gözlem değerleri arasında; X'in Y değerine bağlı olarak bir değişimi söz konusu olduğu halde Y değerinin X değerini etkilememesi, yani bağlı olmaması durumudur. Örneğin, yapılan bir anket çalışmasıda cinsiyetlere bağlı olarak sorulara cevap vermeme eğilimi rassal olarak kayıp mekanizmasına dahil bir kayıplık durumunu göstermektedir.

Kayıp veriler için Rastgele Kayıp veri mekanizması oluştuğunda kayıp veriye bakmak yerine gözlenen veri üzerinden bir değerlendirme yapmak gerekebilmektedir. Çünkü kayıplık, kayıp olan veri ile ilgili olmayıp gözlenen veriler üzerinden bir bağlantı oluşturmaktadır (Little ve Rubin, 1987).

2.3.3 İhmal Edilemez Kayıp (NI)

Kayıp veri mekanizmasında X gözlem değerinin Y'ye Y değerinin de X gözlem değerine bağlı olmasından kaynaklanan ve verilerdeki kaybın ROK veya TROK mekanizmasına dahil olmama durumudur. Veri kaybı rassal olmayıp veri grubundaki diğer değişkenler üzerinden bir kayıp veri tahmini gerçekleştirilememektedir (Allison, 2001). Örneğin, yapılan bir anket çalışması için hazırlanan soruların hedeflenen kitleye sorulamaması veya hedef kitle için doğru soruların belirlenmemiş olması sonucunda cevapsız kalan sorular, ihmal edilemez kayıp veri mekanizmasına dahil bir kayıplık olarak değerlendirilebilmektedir.

2.4 Kayıp Veri Sürecinde Rastgeleliğin Sorgulanması

Herhangi bir veri grubundan kayıp veriler için uygun analiz yöntemleri belirlenirken rastgele dağılıp dağılmadığı hakkında bilgi sahibi olmak gerekmektedir. Kayıp verinin rastgele dağılım gösterip göstermediğini sorgulamak için 3 ayrı yöntem geliştirilmiştir. Bu yöntemlerden ilk olarak, mevcut değerlendirmeye tabi tutulmuş veriler kayıp veriye ait değer içerip içermeyen olmak üzere iki ana grup altında toplanmıştır. Daha sonra kayıp veri setlerindeki ilgilenilen değer grupları dikkate alınarak bu iki grup ortalaması hakkında anlamlı bir fark olup olmadığı ile ilgili sonuçlara ulaşmak için t testi uygulanmaktadır. Elde edilen sonuçlar dikkate alındığında bu iki grup arasında elde edilen anlamlı fark bize tamamen rastgele olmayan kayıp mekanizmasının bir göstergesi ve analizlerde yol gösterici bir yöntem olarak ön plana çıkmaktadır (Alpar, 2003; Baygül, 2007).

İkinci değerlendirme yönteminde, ilk değerlendirme yöntemine benzer olarak tekrar bir gruplandırma yapılarak tam veriler 1, kayıp olan veriler 0 olarak numaralandırıldıktan sonra değişkenler arasındaki Pearson korelasyon katsayısı hesaplanır. Hesaplama sonucu oluşan pearson katsayısı her bir değişken çiftinin ilişki derecesini bize gösterir. Azalan korelasyon katsayısı bize rastgeleliği göstermektedir (Alpar, 2003; Baygül, 2007).

Üçüncü olarak TROK mekanizmasında rastgeleliğin araştırılmasında çok sık başvurulan bir yöntem olan ki-kare testidir. Test sonucunda H_0 hipotezi ret edildiğinde ($p<0,05$) mevcut verilerin yapısının TROK mekanizmasına dahil olmadığı sonucuna ulaşılır (Little, 1998).

3. MATERYAL VE YÖNTEM

3.1 Kayıp Veri İle Başetme Yöntemleri

Kayıp veri oluştuktan sonra sorun çözücü birden fazla yaklaşım bulunmaktadır. İstatistik analizlerde kayıp verileri dikkate almamak ve veri silme yöntemlerini kullanmak veya değer atama yöntemlerinden biri ile kayıp veri setlerine uygun şekilde kayıp veriyi ikame etmek yaklaşımları benimsenebilir.

Kayıp veri yerine atama yapılırken bu değer ortalama değer olabilir veya değişkenin daha önce tanımladığı veri setine yakın veya bazı tahminlere dayalı olarakta veri oluşturulabilir. Değer koyma metodunda ise eksik olan değerler için ikame veri oluşturularak veya regresyon modeli kullanılarak kayıp veri analizi gerçekleştirilebilir. Hot Deck yerine koyma ve Cold Deck yerine koyma yöntemlerinde, kayıp olan veriye en yakın bir tahmini değer oluşturularak bir modelleme yapılmıştır. Amaçlanan; veriler eksiksizmiş gibi hareket edilerek kayıp veri değerlendirme yöntemlerinden en uygun olanı kullanabilmektir. Bu doğru yöntemi tespit etmek için izlenecek metodu belirleme;

1. Mevcut veri setine kayıp olan değer için yeni bir gözlem değeri ilave etme,

2. Veri setinden veri veya veri gruplarının silinerek eksiksiz örneklem oluşturulması,

3. Kayıp veriye yakın veya yaklaşık bir değerin, tahmin yoluyla kayıp veri olarak belirlenmesi,

4. Mevcut olan gözlem değerlerinin kullanılması,

5. Gözlem değerlerinin yer değiştirmesi,

gibi kriterler kullanılarak gerçekleştirilebilmektedir.

Veri setlerine yeni gözlem değerlerinin ilavesi her zaman mümkün olmamakla birlikte hem iş gücü hem de zaman kaybına neden olabilmektedir. Kayıp verilere ek olarak mevcut örneklemden gözlem değerlerinin çıkarılması ise büyük hem very hem de kaybına neden olacağından uygulanacak istatistiksel yöntemlerin gücünü azaltacaktır (Roth, 1994; Alpar, 2011).

Tahmini değer atama yönteminde, veri setlerine tahmini olarak değerler atanırken dikkatli davranılmazsa çözümden uzaklaşılarak yeni problemlerin ortaya çıkmasına sebebiyet verilebilir (Little, 1987; Rubin, 1987).

Veri grupları değerlendirilirken doğru yöntem ile çalışmak veriler için anlamlı sonuçlar elde etmede önemlidir. Bu sebeple doğru analiz için karar verilirken yöntemlerin yapısı, yoğun olarak kullanıldığı veri grupları, analizlerde hangi yöntemlerin daha çok tercih edildiği ve yöntemlerin maaliyeti gibi soruların cevaplanarak uygun olan yöntemin belirlenmesi kayıp veri ile başetmede öncelikli kriterler arasında yer almaktadır.

3.1.1 Veri Silme Yöntemleri

Veri Silme Yöntemleri çiftler düzeyinde veri silme ve liste durum düzeyinde veri silme yöntemleri olarak iki ana başlık altında toplanmaktadır. Bu iki yöntemde veri kaybını en aza indirecek şekilde bir veri silme işlemi gerçekleştirilebilmelidir. Amaçlanan yeterli sayıda verinin istatistiksel analizler gerçekleştirilmeden önce veri seti veya setlerinden uzaklaştırılmasıdır. Fazla sayıda veri örneklemden uzaklaştırılırsa analizler için olumsuz sonuçlar ortaya çıkabilmektedir. Silme yöntemi kullanıldığında yeterli ve uygun sayıda verinin analiz için örneklemde kalması önem arzetmektedir (Baygül, 2007).

3.1.1.1 Liste Durum Düzeyinde Veri Silme

Liste durum düzeyinde veri silme yönteminde kayıp veri gözardı edilerek veriler analiz edilir. Bu yöntemde ikame veri tahsis edilmemektedir. Mevcut istatistiksel analizler haricinde ilave yeni analizlere ihtiyaç duymadan veri setlerindeki örneklemlerde parametre tahmini gerçekleştirilebilmektedir. Kayıp veri ile ilgili değer, mevcut veri grubu içinden çıkarılır. Veriler Tamamen Rastgele Kayıp özelliği gösterdiğinde bu yöntem yansız ve tarafsız sonuçlar üreten ve araştırmalarda çok sıklıkla başvurulan bir yöntemdir (Allison, 2009). Liste durum düzeyinde veri silme yönteminde, kullanılan veri sayısı azalacağından standart hatanın büyük çıkma olasılığıda yüksek olacaktır. Küçük olan örnek gruplarında parametre tahminlerinde yanlılığa yol açabilmektedir (Demir, 2013).

Liste durum düzeyinde veri silme yöntemi; istatistiksel yöntemlerle değerlendirilirken Tamamen Rastgele Kayıp veri mekanizmasında sağlıklı sonuçlar

verir ama Rastgele Kayıp mekanizmasında gerçek değerlendirmelerden uzak, taraflı sonuçlar verme ihtimalini yükseltebilmektedir. Amaçlanan, veri seti için tamamlanmış bir örneklem oluşturularak analiz yapabilmektir.

3.1.1.2 Çiftler Düzeyinde Veri Silme

Bu yöntemde kayba uğramamış tüm gözlem değerlerinden yola çıkılarak bazı istatistiksel parametreler önce hesaplanır. İki değişkenli fonksiyonlarda X ve Y gibi iki örneklemde korelasyon, ortalama, gibi parametrik tüm değerler analiz edilir (Allison, 2002). Mevcut yöntem rastgele kayıp mekanizmasında anlamlı ve daha olumlu sonuçlar verebilmektedir. Diğer yöntemlere göre daha çok veri ile çalışıldığı için sonuçların gözlenebilir şekilde doğruluğu ayırt edilebilmektedir.

Liste durum düzeyinde veri silme yönteminden farklı olarak mevcut veri setindeki gözlem değerinden kayba uğramayan bölümü çift değere sahipse bu yöntem kullanılmaktadır. Birden fazla hesaplamaya gerek duyulan bir yöntem olarak kayıp veri analizlerinde kullanılıp, ortalama, standart sapma, korelasyon matrisi gibi tespit edilebilen tüm veriler hesaplanabilmektedir. Diğer bir fark ise kayıp veriye dahil olan birey veya gözlemin analizden uzaklaştırılması yerine kayıp verinin dahil olduğu durumun analizden çıkarılabilmesidir (Howell, 2007).

Liste durum düzeyinde veri silme de olduğu gibi bu yöntemde de Tamamen Rassal Kayıp mekanizmasında tarafsız sonuçlar elde edilebilmektedir (Allison, 2009). Gözlem değerinde daha az değişkenlik olduğundan yapılan analizler sonucunda daha düşük değerde standart hataya sahip olunabilmektedir. Korelasyon yüksek ise liste durum düzeyinde veri silme, korelasyon düşük ise çiftler düzeyinde veri silme yöntemi ile istatistiksel olarak anlamlı sonuçlar oluşturulabilmektedir (Baygül, 2007; Demir, 2013; Öztemur, 2014).

Veri silmede, silinen veri grubunun küçük olması yöntemin başarısını, sonucun doğruluğunu ve hata payını olumlu yönde etkileyebilmektedir. Testin gücünün artması için silinen veri grubunun az olması elde edilen sonuçları anlamlı hale getirebilmektedir.

3.1.2 Eldeki Değerlerinin Kullanılması Yöntemleri

Bu yöntemler sırasıyla eldeki tüm gözlem değerlerinin kullanılması yöntemi ve tam gözlem değelerinin kullanılması yöntemi olup istatistik paket programlarına uyumlu olarak sonuçlar elde edilebilmektedir.

3.1.2.1 Eldeki Tüm Gözlem Değerlerinin Kullanılması Yöntemleri

Uzun zamana yayılmış veri setlerindeki kayıp veriye ait değer mevcut tüm gözlem değeri olarak değerlendirilir ve analiz edilir. Yaygın olarak kullanılan bir yöntemdir. Mevcut analize uygun ölçütlerden kovaryans, korelasyon, and standart sapma ve ortalama gibi değişkenlerin sonucuna ulaşılır. İstatistiksel yazılımlardan SPSS paket programı kullanılarak veriler değerlendirilebilmektedir. SPSS paket programında ‘‘pairwise’’ seçeneğiyle veriler analiz edilerek analiz sonuç menüsünde tüm veriler için elde edilmiş değerler belirtilmektedir (Çiğdem, 2011).

Bu nedenle, kayıp veri içeren araştırmalarda eldeki tüm bilginin kullanılması yöntemi çok sık kullanılmaktadır. Bu atama yöntemi, gerçekte kayıp verilerin yerine konulması ya da kayıp verilerin atanmış değerlerle doldurulması değil, eldeki tüm veriler dağılımı yardımıyla tanıtıcı istatistiklerin (ortalama, standart sapma vb.) ve bağlantılı ölçülerin (korelasyon, kovaryans) elde edilmesidir. İstatistik programların büyük bir bölümünde bu yönteme ilişkin menüler olduğu için araştırmacılar tarafından sıklıkla kullanılmış ve bu yaklaşımla elde edilen korelasyon ya da kovaryans matrisleri (faktör analizi) girdi matrisi olarak kullanılmıştır (Çiğdem, 2011).

Gözlem değerindeki farklı sayıda bütün verilerin bir arada değerlendirildiği bu yöntemde kayıp veriler TROK mekanizmasına dahil değilse analiz süreci sonucunda elde edilen verilerde yanlı ve örneklemi temsil düzeyi yeterli olmayan sonuçlar elde edilebilmektedir. Bu da analizin başarısını olumsuz yönde etkileyebilmektedir. Bir diğer dezavantajı ise elde edilen sonuçlara ait korelasyonların belirlenen sınırların dışında gerçekleşmesi ve korelasyon matrisleri arasında tutarsızlık oluşturabilmesidir. Kayıp verinin az olduğu veri gruplarında bu yöntemin kullanılması elde edilecek sonuçlar için önem arzetmektedir (Çiğdem, 2011).

Bu yöntemde mevcut bilgiler kullanılarak tahmini kayıp veri analizi ve kovaryans çözümlemesi yapılabilmektedir. Tekrarlanabilen ölçümlerin analizlerinde ve eşit olmayan uzunluktaki örnek dizilerinde kolayca kullanılabilen bir yöntemdir.

Bu analiz yönteminde en çok başvurulan yöntem çift taraflı silme metodudur. Analizde kayıpsız veri seti oluşturularak tam vaka analizine göre korelasyon matrisi hesaplanır. Örneklem büyüklüğü maksimize edilir.

Kayıp olan veri setlerinin bilgilerini analizlerde sunmasından dolayı; etkili bir yöntem olarak ön plana çıkmaktadır. Bu yöntemin dezavantajı, kayıp veri modeline uygun oluşan değişkenlere göre farklı sonuçlar ortaya çıkarabilmesidir. Diğer dezavantajı ise, örnek boyutunun ve bununla bağlantılı olarak analizlerde ortaya çıkan serbestlik derecesinin değişkenlik gösterebilmesidir. Kayıp veriler Tamamen Raslantısal Kayıp özelliği göstermediği sürece sonuçlar, önyargılı ve yanlı olacaktır.

3.1.2.2 Eldeki Tam Gözlem Değerlerinin Kullanılması Yöntemi

TROK mekanizmasına dahil olan veriler için kullanılan bu yöntemde kayıp veri içermeyen değerler analiz edilir. Kayıp veri içeren gözlemler değerlendirilmeye alınmadan veri gruplarından uzaklaştırılır. Bu yöntemde analizler sonucunda tarafsız ve anlamlı sonuçlar elde etmek için rassal yapıda olmayan yani TROK yapısına dahil olmayan veriler değerlendirilmeye alınmamalıdır.

İstatistik paket programları üzerinden sıklıkla kullanılan bu yöntemde ortalama vektörü, kovaryans matrisi ve korelasyon matrisi gibi bağlı istatistiksel değerler elde edilebilmektedir. Tam gözlem değerlerinin kullanılması yönteminin kullanımındaki dezavantajı, kayıp veriye sahip gözlem değerlerinin, veri gruplarından uzaklaştırılması sonucu analiz edilecek değerlerin azalarak eksik bilgiler üzerinden değerlendirme yapılmasıdır. Bu durum yanlı sonuçlar elde edilmesi gibi olumsuz etkilere neden olmaktadır. Bundan dolayı TROK yapısına dahil olan veriler değerlendirilerek elde edilen sonuçlarda güven aralıklarında bir artış sağlanmalıdır (Çiğdem, 2011).

Kayıp verileri analiz etmek için geliştirilen bu yöntemde, kayıp olan veriler atlanarak bir değerlendirme yapılır. Bu yöntemin avantajları sıralandığında; Her türlü istatistiksel değerlendirmeye uygun bir metod olup, ayrıca özel bir hesaplama yöntemi gerektirmeyebilmektedir. Dezavantajları sıralandığında; Tamamen Rastlantısal Kayıp özelliği gösteren verilerde tahmini yanıt eğiliminin tarafsız bir değerlendirmeden uzak olmasıdır. Bu durumda örnekte temsil edilen kısmındaki değerlendirmeler önyargılı olup sonuç yanlı olabilmektedir. Kayıp olan tüm veri setlerinin silinmesinden dolayı

istatistiksel hassasiyet ve testin gücü azalabilmektedir. Kayıp veri sorunu, örneklemin küçük bir kısmını temsil ediyorsa, bu yöntem oldukça iyi sonuçlar verebilmektedir.

3.1.3. Kayıp Veri ile Tam Gözlem Değerinin Yer Değiştirmesi Yöntemi

Veri kaybının yoğun olduğu durumlarda örnekte yer bulamamış benzer özellik gösteren verilerle kayıp verinin bulunduğu alanların yer değiştirmesi yöntemidir (Çiğdem, 2011). Örnek grubunu temsil edecek özellikteki verilerin kayıp veriler yerine dahil edilmesi ile kayıp veriler için en uygun veri seti oluşturulabilmekte ve anlamlı sonuçlar elde edilebilmektedir. Gözlem değerleri arasında yer değişimi söz konusu olduğunda kayıp veri grubu temsil edecek örneklemin içinden veya benzer özellikteki veriler örnek dışarıdan seçilerek bir atama işlemi gerçekleştirilebilmektedir.

3.1.4 Son Gözlem Değerini İleri Taşıma Yöntemi

Bu yöntemde kayıp veri yerine bir önceki değer kullanılarak kayıp veri telafi edilir. Her kayıp veri için son gözlemlenen değer ile bir yer değişimi söz konusudur. Yöntem sanayide, sağlık alanındaki tedavi gruplarında ve grup olarak değerlendirilen yatay parsel denemelerinde kullanılabilmektedir. Son Gözlem Değerini İleri Taşıma yöntemi basit ve sonuç verilerinin analizlerinin güçlü olduğu varsayımına dayanarak gerçekleştirilen bir yer değiştirme metodu olarak ön plana çıkmaktadır.

Son gözlem değerini ileri taşıma yöntemi; uzun süreli çalışmalarda kullanılan bir yöntem olarak ön plana çıkmaktadır. Kayıp olan verinin yerine kayıp olan veriden bir önceki değer atanarak kayıp veri kısmı belirlenmiş olur. Dezavantajı; Oluşturulan veri setleri analiz edilirken zaman ve değerlendirme yöntemleri itibarıyla bu yöntemde hata oranı yüksek çıkabilmektedir. Kayıp veriler ROK ve TROK yapısına dahilse elde edilebilecek sonuçlar yanlı ve taraflı olabilmekte ve sayısal olarak kayıp veri sayısı arttıkça sonuçlarda hata oranı artabilmektedir. Ayrıca son gözlem olarak taşınan veri diğer veriler içerisinde abartılı bir yapıya sahipse örneklemde marjinal veri grupları çoğalabilmeltedir (Şeker ve Eşmekaya, 2017).

3.1.5 Veri Atama Yöntemleri

Birbirinden bağımsız dört farklı değerlendirme yöntemi bulunmakta olup bu yöntemler sırasıyla;

1. Çoklu Veri Atama Yöntemleri

2. Tekli Veri Atama Yöntemleri

- Regresyon Veri Atama Yöntemleri
- Cold Deck Veri Atama Yöntemleri
- Hot Deck Veri Atama Yöntemleri
- Ortalama Atama Yöntemi
- Stokhastik Regresyonla Veri Atama

3. Kayıp Veri İle Diğer Başetme Yöntemleri

- Beklenti Maksimizasyonu Algoritması
- Karar Ağaçları
- Markov Zincirleri Monte Carlo Yöntemi
- En Küçük Kareler Yaklaşımı
- Yapay Sinir Ağları
- Bayesci Veri Atama
- Mahalanobis Uzaklığı Ataması

3.1.5.1 Çoklu Veri Atama Yöntemleri

Bu yöntemde; veri setlerindeki iki veya ikiden fazla kayıp veri için değerlendirme yapılarak ve birden fazla analiz yöntemiyle en doğru atanacak değer elde edilmeye çalışılmaktadır. Eksiksiz veri seti elde etmek; çoklu atama için her atama sonunda oluşturulan kayıpsız veri setleri üzerinden gerçekleştirilebilmektedir (Enders, 2010). Çoklu Veri Atama yönteminde amaç, kayıp olan iki veya daha fazla örneklemin yerine olasılık dağılımına en uygun verilerin seçilip atanmasıdır. Bu yöntemin avantajı birden fazla sonucun kayıp veri için ikame veri oluşturmada kullanılmasıdır. Diğer değer atama ve veri silme yöntemlerine göre veri tahlilini en doğru değere yaklaştıran bir yöntemdir ve basit paket programlarda dahi kayıp veriler için uygun çözümler geliştirilebilir niteliktedir. Elde edilen çıkarımlar sonucu standart hata ve p değerinden anlamlı sonuçlar alınabilmektedir. Buna bağlı olarak veri setlerinden sonuca yönelik saygınlığı yüksek anlamlı ve varyansı büyük olmayan değerler bulunabilmektedir.

Çoklu Veri Atama yönteminin dezavantajlı olduğu durumları da söz konusudur. İlk olarak bazı veriler yerine ikame veri atadığımızda kayıp verilerin değişken olmasına izin verildiğinden verilerdeki bireysel değişimler gözardı edilebilmektedir. İkincisi gözlemlenen değerler ile kayıp veriler arasındaki ayrımın analizlerde yok sayılmasına bağlı olarak tarafsız sonuç oluşturamamasıdır. Sonuç olarak işgücü ve zaman kaybı fazla olabilmektedir.

Birden fazla kayıp veri grubunda, mevcut veri kümeleri dikkate alınarak tahmini değerler elde edilebilmektedir. Ortalama, standart hata, varyans v.b istatistiksel değerlendirmelere göre kayıp veriler için ikame veri tahmini yapılır. Bu atamalar sırasında mevcut veriler ile atanan veriler arasındaki değişimler incelenir. Kayıp veri grubunda analiz sonuçları kabul edilebilir seviyelerde ataması yapılan hangi grubu gösteriyorsa o veri tahmini olarak seçilebilmektedir. Bu yöntemde; Örneklemdeki genel değişimleri atanan verilerle sabit duruma getirmek amaçlanmaktadır.

Çoklu veri atama için m>1 sayıda veriden kayıpsız veri seti elde edilecek şekilde veri ataması gerçekleştirilmesi ve m kadar verinin standart istatistiksel analizlerle elde edilen sonuçların birleştirilerek değerlendirilmesi süreci içerisinde takip edilebilmektedir (Schafer ve Graham, 2002). Analizlerde; Ortalama, standart hata varyans v.b verileri koruyan veri setleri oluşturmak, genel olarak amaçlanan çoklu veri atama değerlendirme kriterleri arasındadır. Çoklu atama yöntemi, atanan değerler grubunu iyi bir şekilde temsil yeteneğine sahip olduğundan, kullanılabilirliği kolay bir yöntem olarak dikkat çekmektedir. Hesaplamalarda diğer yöntemlere göre daha kolay ve anlaşılabilir yöntemlerdir. Araştırmalarda veri kaybı olmadan analiz sonuçlarına gitmesi ve bu sonuçları mevcut verilere en yakın istatistiksel sonuçlarla değerlendirmesi, çoklu atama yöntemini cazip kılmaktadır.

Kayıp veriler tamamlanırken Markov Zinciri Monte Carlo yöntemi kullanılmaktadır. Mevcut tahmin yöntemleri ile elde edilen veri grupları analiz edilerek grup temsiline yatkınlığı en yüksek ve ortalamalara en yakın olan tahmini değerler atanır.

Tek değişkenli bir varsayım üzerinden değerlendirme yapıldığında regresyon modeli oluşturularak kayıp sayısına göre hesaplama yapılır ve atama gerçekleştirilir. Modelleme koşullu dağılıma bağlı olarak gerçekleştirilir.

Çoklu atama yöntemi analizi yapılırken;

1.Çoklu atama için birden fazla veri seti oluşturulur. Bunlar, standart istatiksel yöntemlerle analiz edilerek belirli bir sayıda sonuç elde edilebilmektedir.

2. Bu analiz sonuçları toplanarak mevcut veri grubuna ait çoklu atama yöntemi tahmini yapılır. Daha sonra yeni bir veri grubu hazırlanır ve kayba uğramamış veriler ile kayıp veriler arasındaki ilişkiyi devam ettiren yapıyı korumak için bir modelleme oluşturulur.

Çoklu atama yöntemi ile veri derlemede ve kayıp veri analizinde yeterli şartlar sağlanamazsa iyi bir yöntem olarak ortaya çıkmayabilir. Standart hata $p<0,05$ olarak kabul edilerek eksik veri kümesi %5'den küçük olarak kabul edilirse çoklu atama yöntemi daha az tercih edilen bir yöntem olarak ortaya çıkmaktadır.

Tekli atamada modelleme yapılırken ve analiz aşamasında, mevcut veriler için özel düzeltmeler gerekmektedir. Maksimum olabilirli tahmini değerlendirme yöntemi; Beklenti Maksimizasyonu Ortalaması ve Çoklu Atama Yöntemiyle veri atamada benzer özellikler göstermektedir. Monte Carlo yöntemine göre seçilen verilerde ise mevcut verilerin dağılımını bozmayan bir tahmini veri ortalaması belirlenir. Rastgele kayıplık mekanizmasının kullanıldığı veri gruplarının çoklu atama yöntemlerinde belirleyici bir rol oynadığı Rubin (1976), Little ve Rubin (1987) tarafından ortaya konulmuştur.

Çoklu atamalardan elde edilen nokta tahmini her analizden elde edilenin (4.4.1)

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i \bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i \tag{3.1}$$

Nokta tahmini için varyans tahmini (4.4.2)

$$V = \frac{1}{m}\sum_{i=1}^{m}\hat{V}_i + \frac{m+1}{m}\left[\frac{1}{m-1}\sum_{i=1}^{m}\left(Q_i - \bar{Q}\right)^2\right] \tag{3.2}$$

Şeklinde olup burada,

$\overline{Q}$: Çoklu atamada Monte Carlo Tekniği,

V : Nokta tahmini için varyans tahmini,

m : Ataması yapılmış ve analiz edilmiş küme sayısı,

$\hat{Q}_i$: Analiz edilmiş i. kümeden tahmin,

V_i : Analiz edilmiş i. kümeden varyans tahminidir. (Baygül, 2007)

3.1.5.2 Tekli Veri Atama Yöntemi

Bu yöntem; Kayıp veri temel alınarak gözlemlenmeyen kısmına, gerçek değerin tahmin edilerek kayıp veri olarak atanması usulüne dayanır. Uygulama da yaygın olarak kullanılmasının sebebi hesaplamasının basit olmasıdır. Dezavantajı; analiz için değer oluşturulurken; örneklemde özel bir düzenlemeye gereksinim duyulabilmesidir. Bu düzenleme olmadan tekli atama için oluşturulan model kayıp veriyi içeren örneklemdeki alana cevap vermede zorluk yaşayacaktır. Bu yöntemdeki diğer bir dezavantaj; Tek değer üzerinden tahmini gerçek değerin ele alınmasıdır. Bu metod; Verilerde özel düzenlemeler gerçekleştirilmeden oluşturulan modele cevap vermede yetersiz kalabilmekte, yanlı sonuçlar ve hatalı değerlendirmelere neden olabilmektedir.

3.1.5.2.1 Regresyon Veri Atama Yöntemleri

Regresyon atama yönteminde, mevcut veriler üzerinden regresyon modeli oluşturularak regresyon değeri hesaplanabilmektedir. Bu değer bizim kayıp veri veya veri setlerini tahmin etmede kullanmamız gereken analiz sonuçlarını ortalama değer olarak gösterebilmektedir. Kayıp veriler için regresyon denklemini, birbiri ile ilişkili olan değişkenler ve tüm değişkenler üzerinden kurarak bir veri ataması gerçekleştirilebilmektedir (Oğuzlar, 2001). Bu yöntem kısıtlayıcı analiz sonuçları içermediğinden daha tarafsız sonuçlar oluşturmada önemli bir avantaj sağlayabilmektedir. Verilerde bağımsız değişken bağımlı değişkeni açıklama oranı yüksek olduğu sürece regresyon veri atama yöntemi kullanılabilir bir teknik olarak ön plana çıkabilmektedir.

Regresyon yönteminde kayıp veriler arası ilişki tahmin edilebilir ve istatistiki analizler bu tahmin değerlerine göre gerçekleştirilebilmektedir. Kayıpsız bir veri seti için, tahmin edilen kayıp verilere ait değişkenlere ait denklem oluşturularak gözlenemeyen kısımlara veri ataması gerçekleştirilebilmektedir (Enders, 2010). Bu

yöntemin kullanılmasında en temel ilke, bağımsız değişkenlerin bağımlı değişkenleri açıklama oranının yüksek tutulmasının istenebilmesidir. Örneğin bir regresyon modeli varsayımında kayıp verilere x dersek, bu kayıp x değişkeni tahmininde diğer bağımsız değişkenler kullanılarak bir çözüm üretilebilmektedir. Regresyon analizinde veriler tamamen rastlantısal kayıp mekanizmasına dahil olduğunda, atanan veriler kayıp veriye sahip olmayan diğer bağımsız değişkenlere bağlı olduğunda en küçük kareler yöntemine göre oluşan katsayılar tutarlı olabilmektedir. Yani sonuçlar yansız ve tarafsız özellik gösterebilir. Regresyon analizinde kullanılan en küçük kareler yöntemi, iki fiziki büyüklük arasındaki matematiksel bağlantıyı kurmak için gerçeğe uygun yazılan regresyon modelini verir.

Bu yöntemde amaçlanan, regresyon denklemi kurularak, mevcut veriler ile kayıp olan veri, tahmini olarak hesaplayabilmektir. Kayıp veri içeren değerler yerine örneklemde kayıp veri içermeyen veri setlerinden regresyon modeli geliştirilir ve bir veri grubu elde edilir. Örnek büyüklüğü genişledikçe sonuçlar tarafsıza yakın olacaktır. Kayıp veriler veya verilere ait değişken tahminleri elde edilebilir. Kayıp veriler yerine bu tahmini değerler kullanılır. Böylece tamamlanmış ve kayıp veri içermeyen örnek grupları oluşturulur. Tamamen Rassal Kayıp Mekanizmasına dahil verilerde ise gözlem değeri en küçük kareler yöntemi kullanılabilir. Böylece değerler yanlı sonuç verme durumunda olmayabilirler.

Regresyon atama yönteminin dezavantajları sıralanırsa; Hata teriminin modele dahil edilmediği durumlarda, varyansı küçük gösterebilmektedir (Baygül, 2007). Gözlenen ve gözlenmeyen değerlerin arasındaki korelasyon zayıf ise bu yöntem kullanılabilir bir yöntem olararak değerlendirilmeyebilmektedir.

3.1.5.2.2 Ortalama Atama Yöntemi

Bu yöntemde, kayıp verileri doldurmak için kayıp olmayan verilerin ortalaması alınır. Yöntemi modellemede; Kayıp verileri tahminde mevcut verilerin ortalaması alınarak kayıp verilerin yerine bu değerin kullanılması amaçlanmaktadır. Uygulaması basit olsada, standart sapmanın ihmal edilmesi durumunda, veri setlerindeki değişkenlerin dağılımında sonuçtan çok uzak yanlı yaklaşımlar ortaya çıkabilmektedir. Tamamen Rastlantısal Kayıp veri mekanizmasında bu yöntem, olumlu ve anlamlı sonuçlar verebilmektedir.

Uygulanabilirliği pratik olan bir diğer yöntemde kayıp veri ile ortalamanın yer değiştirmesi yöntemidir. Bu yöntem ile veri setlerinin tam ortalaması ile kayıp veri yer değiştirilerek tam bir veri seti oluşturulması amaçlanmıştır (Çiğdem, 2011). Anlaşılabilir analiz sonuçları elde edebilmek için ortalamanın kayıp verilerin yerine ikame edilebilecek kadar örneklemde yer bulabilmesi önem teşkil etmektedir.

Ortalama ve ortalamaya dayalı veri atama yöntemleri uygulandıkları veri setlerinde merkezlere doğru bir yığılmaya yol açmakta ve bu nedenle varyasın düşüşüne neden olarak sonuçlarda yanlığığa neden olabilmektedir. (Little ve Rubin, 1987) Aritmetik ortalamadan elde edilen değer kayıp veriler için veri atamada genel olarak kullanılmaması gerektiği belirtilmektedir (Enders, 2010).

3.1.5.2.3 Cold-deck Veri Atama Yöntemi

Bu yöntemde kayıp veriye ait değer yerine veri tahminini kolaylaştıracak sağlam ve güvenilir kaynaklardan ortalama veya ortalamaya benzer özellikteki sayısal verilerin yerine konulması amaçlanmaktadır. Bu yöntemdeki veriler geçerliliği yüksek sağlıklı veri kaynaklarından alınmış değer gruplarıdır. Bu yöntem kullanılırken kayıp gruba ataması gerçekleştirilecek analiz değerlerinin veri seti dışından alınmasına ve kayıp veri grubuna bu değerlerin uygunluğunun fazla olmasına dikkat etmek gerekir.

Ortalama atamaya benzer olarak sonuçlarda varyansın düşük olması ve verilerde merkezlerde yığılmalara neden olarak yanlı sonuçlar oluşturabilmesi dezavantajıdır (Alpar, 2003).

3.1.5.2.4 Hot Deck Veri Atama Yöntemi

Cold Deck yöntemine benzer özellikler gösteren Hot Deck atama yönteminde ataması gerçekleştirilecek veri grupları aynı veri grubundan seçilir ve seçilen bu veriler Cold Deck yönteminde olduğu gibi aynı yoğunlukta olmalıdır. Hot Deck veri atama yönteminde kayıp veri ataması gerçekleştirilirken tamamlanmış değerler için satırlar arası uzaklık hesabı olan k-en yakın komşu hesabı kullanılmaktadır. Bu yöntemin uygulanabilmesi için aşağıdaki adımlar gerçekleştirilir:

- Veriler tamamlanmış veri ve kayıp veri kümeleri olmak üzere iki ayrı gruba ayrılır.

- X_i tamamlanmış veri kümesi matrisini, X_{ij} i. durumun j. değişkeni; Y_i tamamlanmamış veri kümesinin matrisi, Yij i. durumun j. değişkenini belirtmektedir.

- Bu iki küme değerlendirilerek, her kayıp veri içeren satır için Öklid (Euclid (d) uzaklığı hesaplanır (Sezgin ve Çelik, 2013).

Bu değerler kayıp veri setleri içinden gözlemlenen veri havuzundan tahmini olarak seçildikten sonra ortalamaya yakın sonuçlar oluşturulabilmelidır. Kayıp veriler için numunedeki benzer cevap birimlerinden faydalanılır. Kayıp veri yerine ikame edilen değer, verilerin dağılımını etkilemez. Hot Deck veri atama yöntemi genellikle ayrıntılı soru dağılımı olan anket çalışmalarında kullanılabilmektedir. Dezavantajı, örneklemdeki kayıp veri birimlerinin zor bulunmasıdır. Korelasyondaki çarpıtmalar ve kovaryans, bu yöntemin ciddi dezaavantajları arasında yer almaktadır.

$$\text{Öklid}(d) = \sqrt{\sum_{j=1}^{n}(X_{ij} - Y_{kj})^2} \qquad (3.3)$$

Şeklinde olup burada,

Öklid (d) : Her eksik veri içeren satır için öklid uzaklığı,

X_i : Tamamlanmış veri kümesi matrisi,

X_{ij} : i. durumun j. değişkeni,

Y_i : Tamamlanmamış veri kümesinin matrisi,

Y_{ij} : i. durumun j. değişkenidir.

3.1.5.2.5 Stokastik Regresyonla Değer Atama

Regresyonla değer atama tekniğinden farklı olarak stokastik regresyonla değer atama yönteminde (Stochastic Regression Imputation, SRI,) kayıp veri tahmini için oluşturulan doğrusal denkleme, normal dağılım gösteren bir hata terimi ilave edilerek, kayıp veri için analizi yapılmaktadır. Regresyon denklemi ile tahmin edilen değere normal dağılımından rastgele belirlenen bir değer ve standart hatanın mevcut regresyon denklemiyle çarpımından elde edilen hata terimi eklenir. Sonuç olarak bu yöntem ile kayıp verinin regresyonla atamasından kaynaklanan hata varyasının sıfır olması sorunu oluşmamaktadır. Regresyon atamasına göre eklenen hata terimi varyansı artırabilmekte ve sonuçlardaki yanlılığı azaltmaktabilmektedir (Enders, 2010).

3.1.6 Kayıp Veri İle Diğer Başetme Yöntemleri

3.1.6.1 Beklenti Maksimizasyonu Algoritması (EM)

Bu yöntem, kayıp veri parametrelerini belirleyen, yenilemeli algoritmik bir modellemedir. Beklenti Maksimizasyonu Algoritmasında, gözlemlenen veriler ile beklenen verilerin koşullu olasılık tahminleri amaçlanır. Kayıp verilerin yoğun olduğu durumlarda beklenti maksimizasyonu algoritma hızının yavaş olabileceği, bu yöntemin dezavantajları arasındadır.

EM algoritması bir nesnenin hangi kümeye ait olduğu ile ilgili kesin mesafe ölçütlerini kullanarak bulmayı tercih edebilmektedir. Bu yöntem regresyon atamasının tamamlama süreçli 2 aşamalı halidir. İlk olarak beklenen değerin bulunması (E adımı) ve Maksimizasyon (M adımı) gerçekleştirilir. E adımında kayba uğramamış verilerin parametrik değerlerine ait kestirimler kullanılarak kayıp veriyi en iyi temsil eden veri grubu tahmin edilir. M adımında ise tahmin sonucu elde edilen veriler, kayıp olan verilerin yerine konularak bütün veriler üzerinden maksimum olabilirlik hesaplanarak parametreler için yeni kestirimler elde edilmektedir (Sezgin ve Çelik, 2013).

Kayıp verinin maksimum ne olabileceği tahmin edilerek değerlendirme yapılır. Yapılan parametrik kayıp veri tahmini, şartlı beklenen değer üzerinden gerçekleşir. Kayıp veri yerine en çok olabilirlik tahminine göre yapılan değer atanır. Tekrarlanan algoritma ve tamamlanmış parametrik veri modeline göre kayıp veri, maksimum olabilirlik tahmini ile beraber değerlendirilebilmektedir.

İlk aşama da, regresyon tahimini yapılır. Hata terimi, mevcut veriler üzerinden bir modelleme yöntemi ile geliştirilir. Böylece; Kayıp veri içermeyen bir veri matrisi elde edilmiş olur. Buna göre bir hesaplama metodu belirlenir.

İkinci aşamada, ilk aşamaya göre atanan verilere uygun bir kayıp veri matrisi denklemi kurulur. Bu yöntem de, daha fazla bilgi ile daha iyi tahmin yapılır. Hesaplamalar, kayıp veri seti ile mevcut veriler arasında anlamlı bir fark kalıncaya kadar sürdürülür. Kovaryans matrisi, ortalama, korelasyon, t dağılımı, standart hata gibi istatistiksel hesaplamalar yapılır.

Mevcut veri grubunda Tamamen Rassal Kayıp (MCAR) mekanizmasını tespit edebilmek için ki-kare testi kullanılmaktadır. Hipoteze göre;

H_0: Tamamen Rastgele Kayıp veridir.

H_1: Tamamen Rastgele Kayıp veri değildir.

Test sonucunda H_0 hipotezi ret edilir ise ($p<0.05$) veri MCAR veri değildir. Yani MAR, MNAR olabilir (Little ve Rubin, 1987).

3.1.6.2 Karar Ağaçları

Karar ağaci için C.4.5 algoritması ile kayıp veriler için tahmini değer bulma işlemi gerçekleştirilmektedir. Burada kullanılan veri kümesi grubu T olsun. Bu kümenin kullanılan herhangi bir özelliği X olsun. Bu özellikler için bilgi kazanımı X.(T) olarak değerlendirilir. Olmayan veriler bulunurken X.(T) bu kümeden çıkarılır. Mevcut olan durum sayısı n olarak belirlenirse ve kayıp olan veriler b ile ifade edilirse n-b bize kayıp olarak veri grubunu sunmaktadır (Sezgin ve Çelik, 2013).

Sonra eksik olmayan değer, toplam değere oranlanarak F=n-b/n formülü bulunarak F=bilgi (T)-bilgiX (T) formulü ile bilgi kazancı elde edilebilmektedir. Bu işlem her kayıp veri grubu için hesaplanarak belirlenen verilerle bir tablo oluşturulur. Bu tablo bize kayıp verilerin tahmini için bir karar ağaci yapısını gösterir (Sezgin ve Çelik, 2013).

3.1.6.3 Markov Zincirleri Monte Carlo Yöntemi

Bu yöntem 3 aşamalı olarak değerlendirilmektedir. İlk olarak k adet veri seti simüle edilir. İkinci olarak kayıp veri içeren değer grupları için tam veri dağılımına bağlı tahminler yapılır ve son olarak bu iki veri grubu birleştirilerek veri setleri tam veri setleri haline getirilir (Hasan ve ark., 2017). Mevcut yöntem ile tahminlerdeki belirsizlikler ölçülebilmektedir. Yöntemin karmaşık yapısı ve analizlerin pahallı olması, yöntemi dezavantajlı duruma düşürmektedir. Ayrıca hesaplamalarda kullanılan kovaryans açık olarak belirtilmemektedir.

3.1.6.4 En Küçük Kareler Yaklaşımı

Bu yöntem parametrik olmayan bir yaklaşım sergileyip, ana faktörün iki ayrı yaklaşım ile belirlenmesine dayanır.

İlk olarak ele alınan Kayıp Verisiz Model yaklaşımında; Kayıp veri probleminin çözümü için temel bileşenler analiz edilir. Tek ve çok boyutlu uzaydaki veri gruplarının basitleştirilip tamamlanması sağlanabilmektedir. Bir çok değerlendirmede sonuçlarda anlamlı hatalar görülebilmektedir.

İkinci yaklaşım olarak tamamlanmış veri modelinde; kayıp veriler ad-hoc yani karşılıklı değer ile tamamlandıktan sonra karşılıklı olarak yer değiştirmesi ile bu yöntem uygulanmaktadır. Bu yöntem, ilk olarak değerlendirilen kayıp verisiz model varsayımında sonuca ulaşılamadığı durumlarda kullanılabilmektedır. Yöntemin dezavantajı verileri değerlendirirken yavaş çalışabilmesidir (Wasito, 2003).

3.1.6.5 Yapay Sinir Ağları

Yapay sinir ağları üzerinden kayıp veri analizi gerçekleştirilirken kendilerine örnekler halinde verilen kayıp verili bilgilerin örüntülerini kendi ve diğer bilgilerle ilişkilendirebildikleri, üzerinde çalışılan örneklemin hangi kümeye dahil olması hususunda faydalanılabilecek bir metod olmasının yanısıra kayıp verileri kayıpsız veri setleri haline getirmede başarılı oldukları yapılan çalışmalarda belirtilmektedir. Fakat yapay sinir ağları ile yapılan çalışmalar yeterli olmadığından bu yöntem hakkında soru işaretleri oluşturabilmektedir. Çünkü, yapay sinir ağları ile kayıp verilere çözüm üretilirken, çözümün neden ve nasıl yapıldığı ile ilgili bilgileri karşılayamadığı gözlemlenebilmektedir. Bu durum yapay sinir ağları ile elde edilen sonuçların güvenilirlik ve geçerliliklerini azaltabilmektedir (Öztemel, 2003)

3.1.6.6 Bayesci Veri Atama

Olasılık sınıflandırıcı bir yöntem olan Naive Bayes veri atama yönteminde her bir sınıf için olasılık hesabı gerçekleştirilerek, her bir örnek dahilinde en yüksek olasılık bulmaya çalışılmaktadır. Hesaplama hızının yüksek olması ve eksik verilere olan duyarsızlığı ile diğer yöntemlerden daha fazla ön plana çıkmaktadır. Bu yöntemde her bir veri grubu için olasılık hesaplaması yapılır. Hesaplaması kolay olan bu yöntemin kayıp verilere olan duyarlılık fazladır. Bayesci veri atamada 3 farklı yöntem izlenmektedir.

- Order Irrelevant Strategy (NBI-OI) yönteminde tamamlanacak özellikler tanımlanarak, kayıp olan veri kümesine değer ataması gerçekleştirilir. Tamamlanmış veri setinden elde edilen değerler, bir sonraki veri grubu için kullanılmaz (Sezgin ve Çelik, 2013).

- Order Relevant Strategy (NBI-OR) yönteminde kayıp veri kümesi için tamamlanma sırasına dikkat edilerek, değerleri tamamlamak için belirlenen özellikler

kullanılır. Tamamlanmış verilerin özelliklerin değeri, bir sonraki derlenen veri grubu için kullanılmaktadır (Sezgin ve Çelik, 2013).

- Hybrid Strategy (NBI-Hm) yönteminde İlk iki yöntem birleştirilerek bu teknik kullanılmaktadır. İlk olarak sıralı strateji kalan, kısmında da sırasız strateji kullanılmaktadır. Bu yöntem 2 ayrı metod altında sıralanarak uygulanır. İlk olarak kayıp veri kümesinin tamamlanacak özellikleri ve sırası belirlenir. Kayıp veri oranı (missing proportion) ve özelliğin önem faktörü (important factor) dikkate alınarak bir sıralama gerçekleştirilir. İkinci adımda tamamlanan kayıp veriler için her adımda sıralı olan strateji yöntemi dikkate alınarak kayıp veri ile değişen veri kümesi kullanılmaktadır (Sezgin ve Çelik, 2013).

Bu üç yöntem dikkate alınırken veri tamamlaması gerçekleştirilecek olan ilk özellik belirlenmeli ve bu özellik için kayıp verilerde tamamlama sıraları göz önünde bulundurulmalıdır.

3.1.6.7 Mahalanobis Uzaklığı Ataması

Gözlemlenen veriler arasındaki benzer veya benzer olmayan özelliklerin korelasyon katsayıları ve uzaklık ölçümleri bu veri atama yönteminde uygulanan çözümlemelerdir. Burda korelasyon veriler arasındaki benzerlik durumunu ifade ederken, verilerin benzer özellikler göstermediğini belirten kısmı uzaklıklar olarak belirtilmektedir.

Bu yöntemde kayıp veri grubunu gösteren değerlerin yerine, kendisine en yakın gözlem değerine sahip veri kümesinden değerler alınır ve kayıp olan veri seti tam bir veri seti olmak üzere tamamlanarak istatistiksel analizlere uygun hale getirilir (Çüm ve ark., 2018).

Kayıp veriler ile başetmede kulllanılan yöntemlerin avantaj ve dezavantajları Çizelge 3.1’ de sunulmuştur

Çizelge 3.1 Kayıp Veri Analiz Yöntemlerinin Avantaj ve Dezavantajları

Yöntemin Adı	Kullanım Şekli	Avantajları	Dezavantajları
Liste Durum Düzeyinde Veri Silme	*Veri setindeki satırların tamamen çıkarılmaktadır.	*Veri setleri üzerinde kullanımı kolaydır.	*Veri kaybının fazla olması *Analizleri değerlendirmede yanlılığın fazla olmasıdır.
Çiftler Düzeyinde Veri Silme	*Veri setinde sütunlar silinerek uygulanır	*Kullanımı kolaydır. *Veri kaybı azdır.	*Veri kaybı da göz önünde alınarak diğer yöntemlere göre etkili değildir.
Veri Atama Yöntemleri	*Veri setine uyumlu ortalama veya tahmini veri ataması şeklinde uygulanır.	*Kullanımı kolaydır. *İstatik programlarıyla Değerlendirme yapılır.	* Model uyumuna uygun elde edilen değer ortalama veya tahmini değerden yüksek olabilir.
Regresyon ataması	*Regresyon modeline uyumlu tahmini değer oluşturulur.	*Verilerin ortalama ve dağılım şekillerinin hesaplanmasında az veri ile çalışıldığından tassarruf sağlar.	*Serbestlik derecesi bozulduğundan yanlı sonuçların artmasına yol açabilir.
Ortalama Atama	*Mevcut verilerin ortalaması belirlenerek bu değer, kayıp olan verilerin yerine yazılır.	*Hesaplanması kolay bir yöntemdir. *Tamamen Rastsal Kayıp olan veri gruplarında anlamlı sonuçlar üretebilmektedir.	*Standart hatanın analize dahil edilmediği durumlarda ortaya konan çözümler yanlı ve taraflı olacaktır.
Cold Deck Veri Atama	*Varolan veri setinin ortalaması veya ortalamaya benzer özellikteki değerleri dış kaynaklı bir veri kümesinden tahmin yoluyla seçilerek kayıp değerler oluşturulur.	*Bu yöntemde sağlıklı kaynaklardan alınan veriler tarafsız sonuçlar oluşturmaya daha yakındır.	*Ortalama ve ortalamaya benzer tahmini değerler, sağlıklı kaynaklardan alınmazsa hata payı bu yöntemde yüksek olacaktır.
Hot Deck Veri Atama	*Aynı veri gruplarından aynı yoğunluktaki ortalama veya ortalamaya yakın özellikte olan değerler tahmini olarak seçilerek, kayıp değerler yerine atamaları gerçekleştirilir.	*Kendi veri grubunda seçilen ortalama ve ortalama benzeri değerler yanlı sonuçlardan uzak değerlendirmeler oluşturmaktadır.	*Örneklemdeki kayıp verilerin tespiti zordur. *Korelasyondaki çarpıtmalar ve kovaryans oluşumu yöntemin olumsuzluklarındandır.
Bayesci Veri Atama	*Her bir kayıp veri grubu için olasılık hesabı yapılır.	*Hesaplanması kolaydır.	*Eksik verilere olan hassasiyet fazladır.

Çizelge 3.1 Kayıp Veri Analiz Yöntemlerinin Avantaj ve Dezavantajları (Devamı)

Yöntemin Adı	Kullanım Şekli	Avantajları	Dezavantajları
Tekli Veri Atama	*Kayıp olan veri yerine örneklemi teşkil edecek tahmini bir değer atanır.	*Hesaplaması kolaydır.	*Mevcut veri grubunda özel düzelmelere ihtiyaç duyar.
Çoklu Veri atama	*Kayıp olan iki veya daha fazla değerin yerine tahmin usulüne dayalı veri atanmasıdır.	*Standart hata veri atamaya dahil edilir.	*Analizler iş gücü zaman kaybi fazladir.
Son Gözlem Değerini İleri Taşıma	*Kayıp veriden sonra gelen değer kayıp veri yerine ikame edilir.	*Uygulanması kolaydır.	TROK ve ROK mekanizmasinda taraflı sonuçlar verebilir.
Eldeki Tüm Gözlem Değerlerinin Kullanılması Yöntemi	*Eldeki verilerle basit istatistik yöntemlerine göre, bir kayıp veri analizi yapilir.	*İstatistik paket programlarında kullanıma uygundur.	*Kayıp veriler TROK yapısında olması gerekmektedir.
Eldeki Tam Gözlem Değerlerinin Kullanılması Yöntemi	*Eldeki verilerden kayıp veriler örneklemden grubundan çıkarılarak tam veri grubu ile analiz gerçekleştirilir.	*İstatistik paket programlarında kullanıma uygundur.	*Kayıp veriler TROK yapısında olması gerekmektedir.
Kayıp Gözlem Değeri ile Tam Gözlem Değerinin Yer Değiştirmesi Yöntemi	*Kayıp veriye benzer özellikler gösteren örnekte yer bulamamış verilerle yer değiştirmesidir.	* Hesaplanması kolaydır.	*Yer değiştirilen tam gözlem değerinin örneklemi düşük temsiliyeti söz konusu olduğunda taraflı sonuçlar elde edilebilmektedir.
Tekli Atama Yöntemi	***Kayıp veriler gözlenmemiş verilere gerçek değere yakın değer tahmini ile veri elde edilmesidir.	***Hesaplaması kolay bir yöntemdir.	***Veriler hesaplanırken özel bir düzenlemeye gerek duymasıdır.
Mahalanobis Uzaklığı Ataması	*Kayıp veri grubuna yakın ve benzer özelliğe sahip veri gruplarından veriler ile tamamlanır.	*Kayıp verileri tam veriler haline getirir.	*Veri kaynağı doğru seçilmezse anlamlı sonuçlar elde edilemeyebilir.

Çizelge 3.1 Kayıp Veri Analiz Yöntemlerinin Avantaj ve Dezavantajları (Devamı)

En Küçük Kareler Yaklaşımı	*Verilerin temel bileşenleri üzerinden analiz edilerek tamamlanması sağlanır. *Tamamlanan verilerde karşılıklı olarak yer değiştirilir.	*Kayıp veri içermeyen veri setleri oluşturulur.	*Veriler değerlendirilirken yöntemin yavaş çalışmasıdır.
Yapay Sinir Ağları	*Yapay sinir ağları üzerinden istatistiki analiz gerçekleştirilmektedir.	*Kayıp verili veri setinde çalişabilir.	*Elde edilen bilgilerin geçerliliği ve güvenilirliği az olabilmektedir.
Karar Ağaçları	*Karar ağaçları C4.5 algoritması üzerinden istatistiksel analiz gerçekleştirilmektedir.	*Tahmin yoluyla kayıp veri elde edilir.	*Kayıp veri tahminleri doğru gerçekleştirilmezse, elde edilen verilerin örneklemi temsiliyeti düşük olabilir.
Stokastik Regressyonla Veri Atama	*Kayıp veri için oluşturulan doğrusal denkleme hata terimi ilave edilerek regresyon veri ataması yapılabilmektedir.	*Sonuçlar için yanlılık azalabilmektedir.	*İlave edilecek hata teriminin normal dağılım göstermesi gerekebilmektedir.
Markov Zincirleri Monte Carlo Yöntemi	*Simüle edilen veriler üzerinden kayıp veri seti için tahmin yapılarak tam bir veri seti elde edilebilmektedir.	*Tahminlerdeki belirsizlikler ortadan kalkabilmektedir.	*Karmaşık ve pahallı bir yöntemdir.
Beklenti Maksimizasyonu Algorotması	*Gözlenen veriler ile beklenen veriler arasında bir olasılık tahmini gerçekleştirilir. *Bu tahmine göre bir regresyon modeli geliştirilerek veri ataması yapılır.	*Gözlemlenen verilerle bir değerlendirme yapılır. *Veri atamada rassal hata terimleride dikkate alınır.	*Tarafsız değerlendirmeden uzak bir yöntemdir.

4. UYGULAMA

4.1 SPSS Programı ile Uygulama

Bu bölümde SPSS v26 (IBM Corporation, NY, USA) programında kayıp veriyi manipüle etmek amacıyla sunulan menülerin incelenmesi amaçlanmıştır. Bu amaçla SPSS programının Transform menüsünde yer alan Compute Variable ile N(50,5) dağılımdan tesadüf sayıları üretilmiş ve hesaplamalar bu veriler üzerinde yapılmıştır.

4.1.1 Uygulama 1

SPSS programının Transform menüsünde yer alan Compute Variable ile N(50,5) dağılımdan üretilen n=20 olacak şekilde tesadüf sayıları içeren bir değişken oluşturulmuştur. "Veriler_1" olarak isimlendirilen bu değişkene ait tesadüfi olarak seçilen 9. ve 14. sıradaki 2 adet gözlem değeri (örneklem genişliğinin %10'u kadar) silinmiş ve "Veriler_2" olarak adlandırılan kayıp veri seti elde edilmiştir. Veri setlerine ait tanıtıcı istatistik değerlerini içeren SPSS çıktısı Şekil 4.1'de verilmiştir.

Statistics

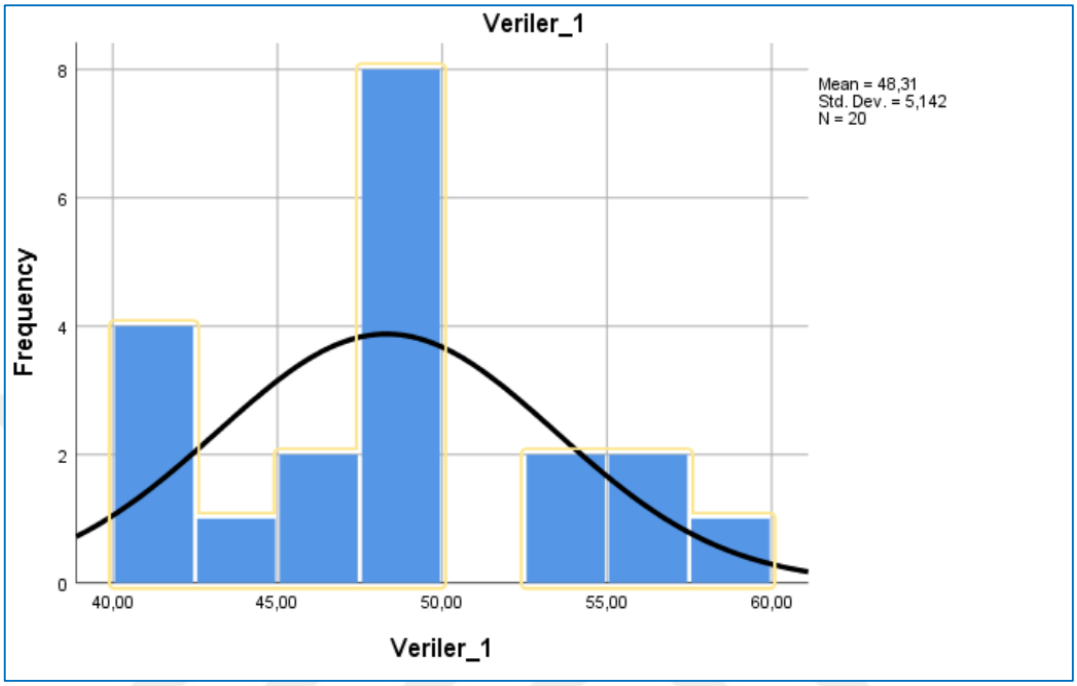
		Veriler_1	Veriler_2
N	Valid	20	18
	Missing	0	2
Mean		48,3091	48,4153
Std. Error of Mean		1,14990	1,27130
Median		48,3558	48,3558
Mode		40,72[a]	40,72[a]
Std. Deviation		5,14249	5,39366
Variance		26,445	29,092
Range		17,45	17,45
Minimum		40,72	40,72
Maximum		58,17	58,17
Sum		966,18	871,47

a. Multiple modes exist. The smallest value is shown

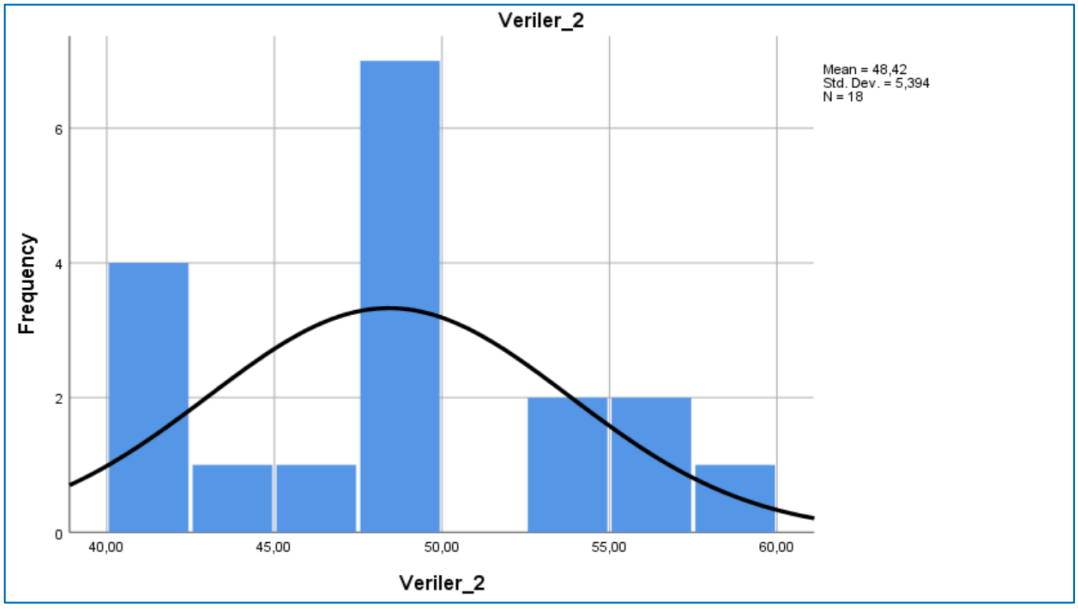
Şekil 4.1 SPSS Programında Veri Setlerine Ait Tanıtıcı İstatistikler

Şekil 4.1 incelendiğinde tam veri seti Veriler_1 ile iki adet kayıp veri içeren Veriler_2 arasında aritmetik ortalama, ortanca değer, tepe değeri gibi merkezi eğilim ölçüleri bakımından farklılık gözlenmezken, standart sapma, varyans ve standart hata gibi değişim ölçülerinde farklılık olduğu görülmektedir. Örneklem genişliğinde meydana gelen azalma ile serbestlik derecesi küçülmüş ve dolayısıyla kayıp veri içeren Veriler_2 ye ait değişim ölçülerinin Veriler_1'den daha yüksek çıkmasına sebep olmuştur. Kayıp verilerin dolaylı olarak değişim ölçülerinde meydana getirdiği artışın

istatistik testler üzerinde olumsuz etkileri bilinmektedir. Kayıp veri içeren ve içermeyen very setlerinin dağılım şekilleri Şekil 4.2 ve 4.3'de verilen histogram grafiklerinde görülmektedir.



Şekil 4.2 SPSS Programında Veriler_1 için Histogram Grafiği

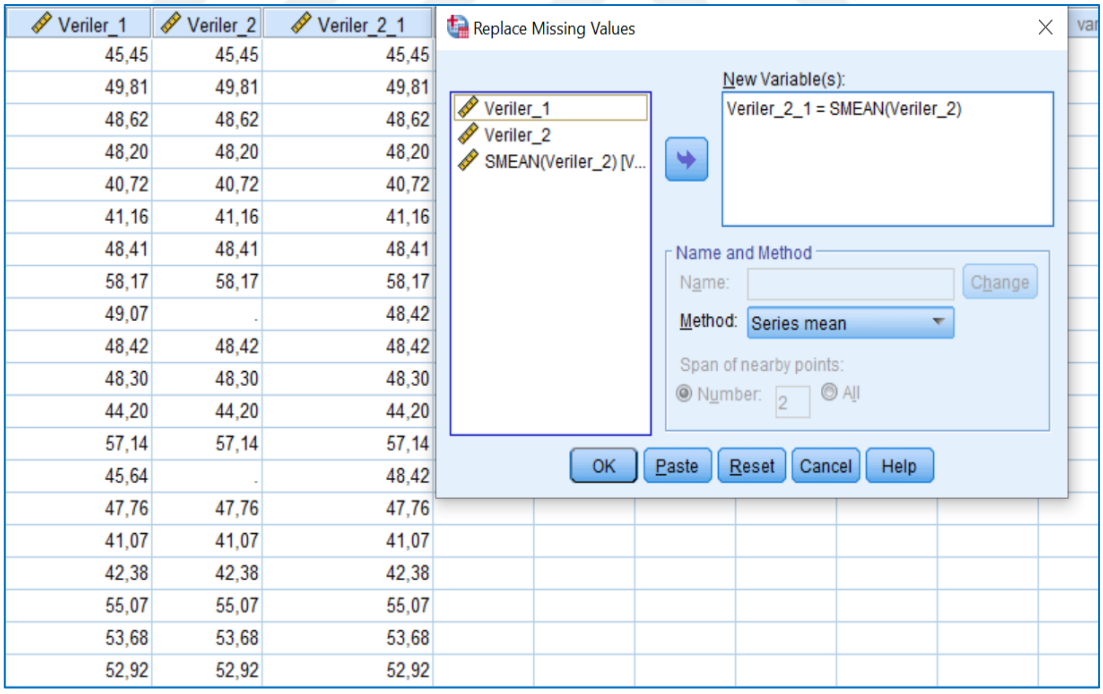


Şekil 4.3 SPSS Programında Veriler_2 için Histogram Grafiği

Şekil 4.2 ve Şekil 4.3 incelendiğinde kayıp verilerin normal dağılım şekilde meydana getirdiği sapmalar gözrülmektedir. Kayıp veri oranı arttıkça yapılacak olan normal dağılım kontrolü tetslerinde H_0 hipotezinin ret edilme olasıklıklarında artış gözlenecektir.

SPSS v26 programında kayıp veriyi yerine koymak için Transform menüsünde bulunan "Replace Missing Values" ile Seriler Ortalaması (Series Mean), Yakın Noktaların Ortalaması (Mean of Nearby Points), Yakın Noktaların Ortancası (Median of Nearby Points), Doğrusal Değer Kestirimi (Linear Interpolation), Noktanın Doğrusal Eğimi (Linear Trend of Point) yöntemleri uygulanabilmektedir. Veriler_2 için bu yöntemler sırası uygulanmış ve kayıp olan iki adet veri yerine konulmuştur.

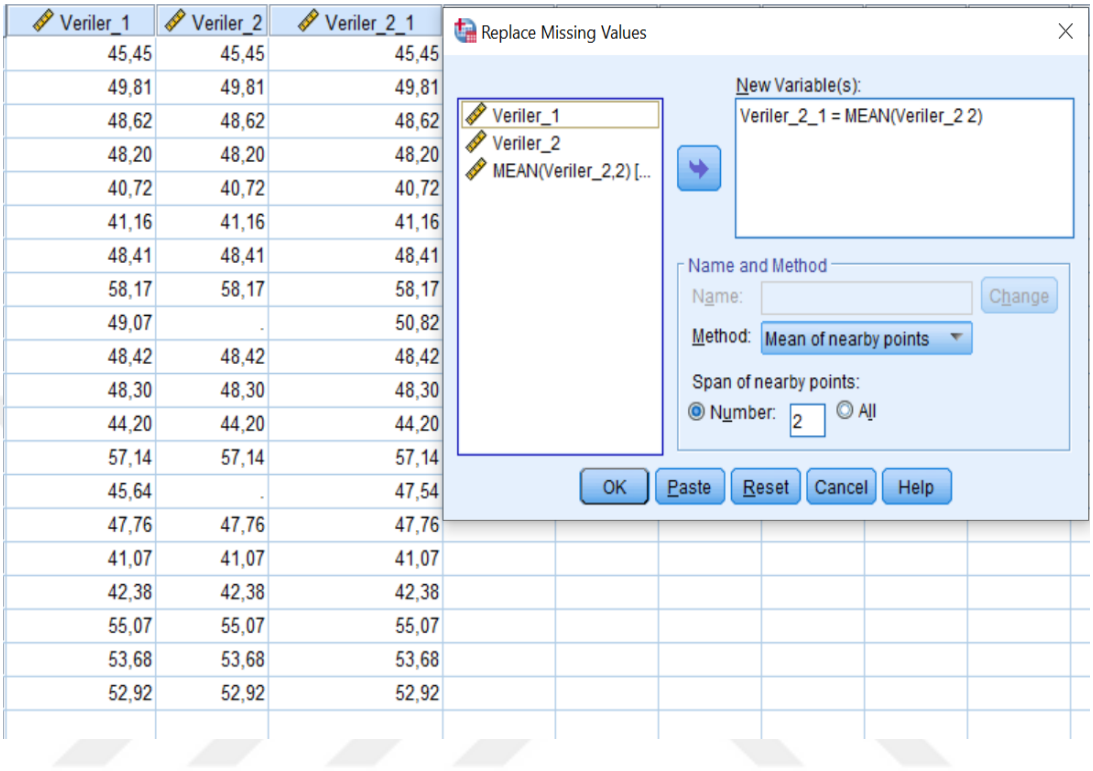
SPSS programında Seriler Ortalaması (Series Mean) Veri Atama Penceresi Şekil 4.4'de gösterilmiştir. Seriler ortalaması tüm veri ünitelerine ait değişkenlerle ilişkili ortalamadır. SPSS programı için kurulu değer (default) olarak yer almaktadır (Çokluk ve Kayrı, 2011). Şekil 4.4'de görüldüğü üzere bu yöntem ile yerine konulan kayıp veriler 48.42'dir.



Şekil 4.4 SPSS Programında Seriler Ortalaması (Series Mean) Veri Atama Penceresi

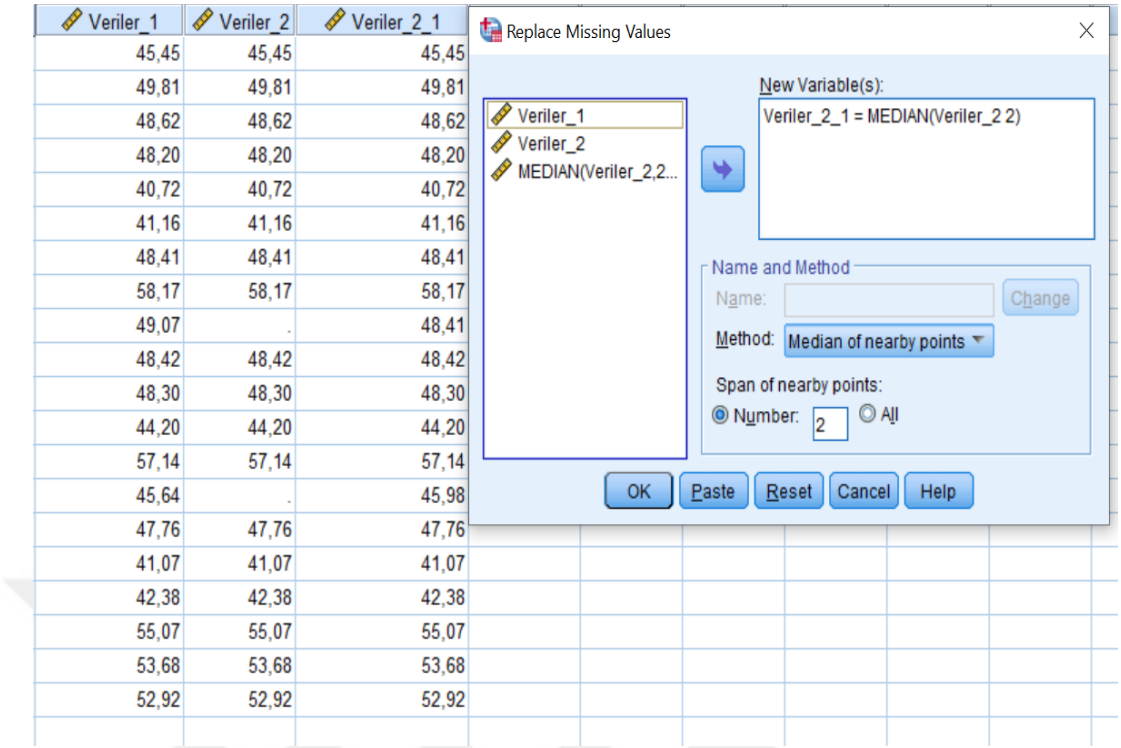
Seriler ortalaması atamasında ortalama değer kayıp verili satırlara işlenerek analitik değerlendirmelere uygun hale getirilmiş olup kayıp veri seti tam bir veri seti haline getirilmiştir.

SPSS programında Yakın Noktaların Ortalaması Veri Atama Penceresi Şekil 4.5'de gösterilmiştir.



Şekil 4.5 SPSS Programında Yakın Noktaların Ortalaması Veri Atama Penceresi

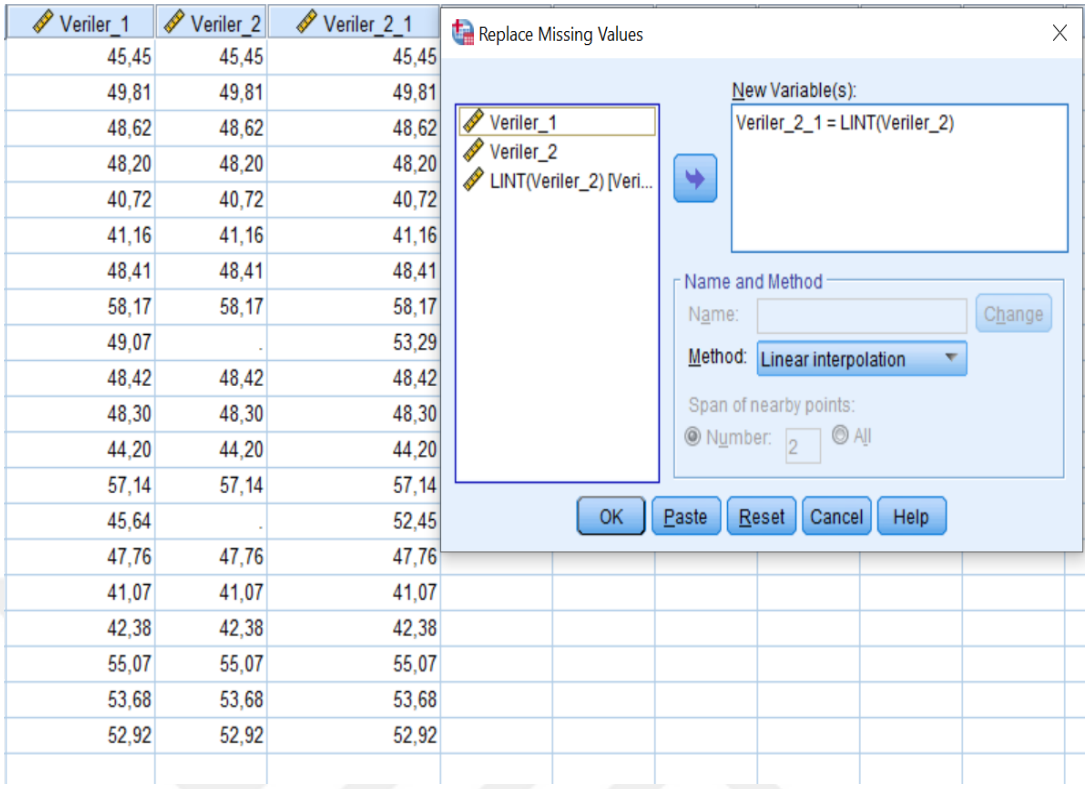
Kayıp veri ile ilişkili yakın değerlerin ortalaması alınarak gerçekleştirilen bu yöntemin uygulaması yakın noktaların uzaklığı (span of nearby points) bölümüne kayıp very sayısı yazılarak yapılabilmektedir. Kayıp olan verilerin altındaki ve üstündeki tam olan gözlem değerlerinden yararlanılarak aritmetik ortalama hesaplanarak, bu değerlerin ataması gerçekleştirilmektedir (Çokluk ve Kayrı, 2011). Yakın noktaların ortalaması ataması işlemi gerçekleştirildikten sonra elde edilen ortalama değer örneklemi temsiliyeti yüksekse, ortalama değer kayıp verili kısımlara işlenerek analiz işlemine devam edilir. Şekil 4.5'de görüldüğü üzere bu yöntem ile yerine konulan kayıp very değerleri sırasıyla 50.82 ve 47.54 olmuştur.



Şekil 4.6 SPSS Programında Yakın Noktaların Ortancası Veri Atama Penceresi

Şekil 4.6 'da kayıp olan 2 değer için yakın noktaların ortancası ataması ile gerçekleştirilecek yöntemle analizlere uygun hale getirilen veriler için kayıp veriye yakın olan değerlerin ortancası alınarak veri ataması sağlanabilmektedir. Kayıp veri için çevreleyen değerlerin sayısı, araştırıcılar tarafından belirlenebilmektedir. Kayıp verilerin altındaki ve üstündeki tam gözlem değeri kullanılarak ortanca değer hesaplanır. Kayıp veriler yerine bu değerlerin ataması gerçekleştirilebilmektedir (Çokluk ve Kayrı, 2011). SPSS yakın noktaların ortancası yöntemiyle kayıp verilerin yerine sırasıyla 48.41 ve 45.98 değerlerini atamıştır.

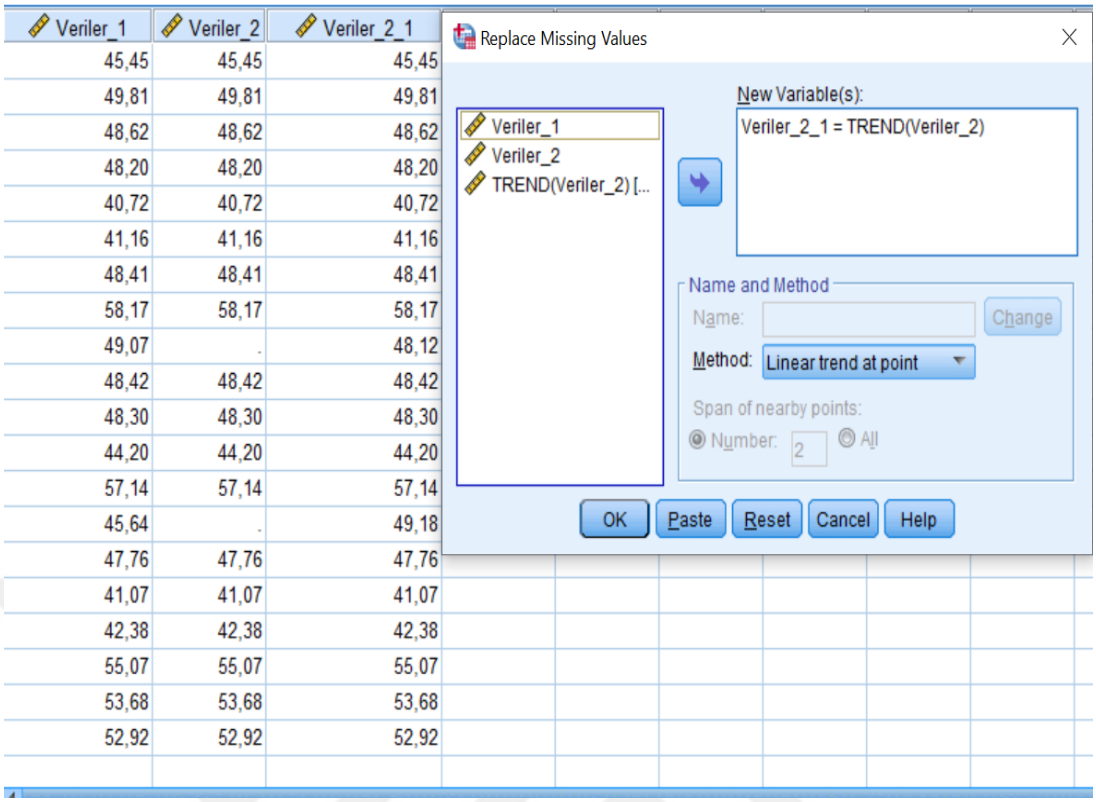
Yakın noktaların ortacası ataması gerçekleştirilirken noktaların değerleri yakın noktaların mesafesi (span of nearby points) seçeneğiyle belirlenerek 2 adet yakın nokta seçilmiş olup ataması gerçekleştirilen verilerle tam veri setleri oluşturulabilmektedir.



Şekil 4.7 SPSS Programında Doğrusal Değer Kestirimi Veri Atama Penceresi

Şekil 4.7'de görüldüğü üzere SPSS Doğrusal Değer Kestirimi ile kayıp verilerin yerine sırasıyla 53.29 ve 52.45 değerlerini atamıştır. Bu yöntemde kayıp veriden önceki son tam gözlem değeri ve kayıp veriden sonraki ilk tam gözlem değerinin kayıp olan veriler yerine atamasının gerçekleştirilebilmesidir. Eğer seride bulunan ilk ve son gözlem eksik ise kayıp verinin yerine herhangi bir veri atamasının gerçekleştirilmesi söz konusu olmayabilmektedir (Çokluk ve Kayrı, 2011). Kayıp veriler için doğrusal değer kestirimi işlemi gerçekleştirilirken kayıp veri öncesindeki ilk gözlem değerinden ve kayıp veriden sonra gelen gözlem değeri arasındaki fark bulunur. Bu fark kayıp veri sayısına bölünerek kayıp veriden sonra gelen ilk gözlem değerine ilave edilir. Elde edilen sonuç kayıp veri için oluşturulmuş bir kestirim değeri olarak tam veri setine yazılır ve tamamlanmış veri seti istatistiksel analizlere uygun hale getirilerek değerlendirilmeye alınabilmektedir.

Şekil 4.8'de kayıp veri için bir regresyon tahmin denklemi oluşturularak her kayıp veri için bir değer ataması ile elde edilen tam bir veri seti gösterilmektedir. Şekil 4.8'de görüldüğü üzere SPSS Noktanın Doğrusal Eğimi ile kayıp verilerin yerine sırasıyla 48.12 ve 49.18 değerlerini atamıştır.



Şekil 4.8 SPSS Programında Noktanın Doğrusal Eğimi Veri Atama Penceresi

Kayıp veri, mevcut örneklemin (örneğin değerler ilk denekten, son deneğe doğru yükselme eğilimi gösteriyorsa) gösterdiği eğilim (trend) ile uyumlu ya da tutarlı olarak belirlenebilmektedir. Mevcut veri serilerinin 1'den n'e kadar ölçeklendirildiği bir indeks değişkeninde kayıp verilere öngörülen değerler atanabilmektedir (Çokluk ve Kayrı, 2011).

SPSS menüsünde yer alan tüm kayıp değer atama yöntemleri ile elde edilen tam veri setlerine ait tanıtıcı istatistik değerleri Şekil 4.9'da verilmiştir. Tanıtıcı istatistik değerleri karşılaştırıldığında; kayıp veri içermeyen Veriler_1'e en yakın sonuçlar Seriler Ortalaması (SMEAN) ile elde edilmiştir. Elde edilen tanıtıcı istatistikler karşılaştırıldığında; standart hatanın en yüksek kayıp verinin dikkate alınmadığı durumda (Veriler_2) en düşük ise Seriler Ortalaması (SMEAN) ile elde edilmiştir. Kayıp veriyi yerine koyan yöntemleri içerisinde en yüksek standart hata Doğrusal Değer Kestirimi (LINT) ile ortaya çıkmıştır.

Statistics

		SMEAN (Veriler_2)	MEAN (Veriler_2, 2)	MEDIAN (Veriler_2, 2)	LINT (Veriler_2)	TREND (Veriler_2)	Veriler_1	Veriler_2
N	Valid	20	20	20	20	20	20	18
	Missing	0	0	0	0	0	0	2
Mean		48,4153	48,4921	48,2935	48,8608	48,4387	48,3091	48,4153
Std. Error of Mean		1,14082	1,14823	1,14729	1,18171	1,14157	1,14990	1,27130
Median		48,4144	48,3558	48,3558	48,4144	48,3558	48,3558	48,3558
Mode		48,42	40,72[a]	40,72[a]	40,72[a]	40,72[a]	40,72[a]	40,72[a]
Std. Deviation		5,10189	5,13503	5,13085	5,28476	5,10526	5,14249	5,39366
Variance		26,029	26,368	26,326	27,929	26,064	26,445	29,092
Range		17,45	17,45	17,45	17,45	17,45	17,45	17,45
Minimum		40,72	40,72	40,72	40,72	40,72	40,72	40,72
Maximum		58,17	58,17	58,17	58,17	58,17	58,17	58,17
Sum		968,31	969,84	965,87	977,22	968,77	966,18	871,47

a. Multiple modes exist. The smallest value is shown

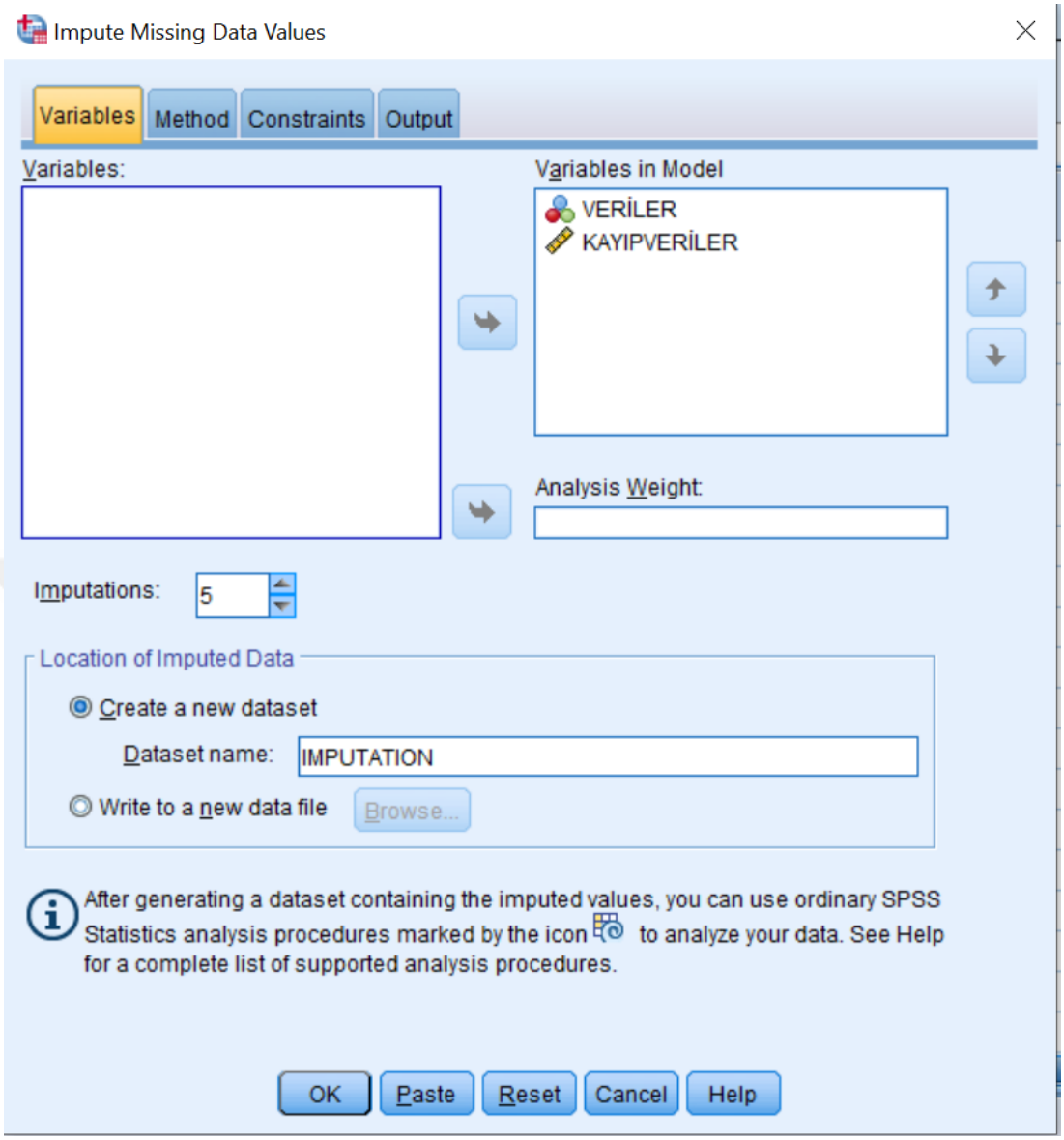
Şekil 4.9 SPSS Programında Kayıp Veri Atama Yöntemlerine Göre Tanıtıcı İstatistik Değerleri

4.1.2 Uygulama 2

Uygulama 1'de hazırlanan Veriler_1 ve Veriler_2 isimli veri seti kullanılmış (Şekil 4.10) ve SPSS programının Analyze menüsünde yeralan "Impute Missing Data Values" ile çoklu atama (Multiple Imputation) gerçekleştirilmiştir.

Veriler_1	Veriler_2
45,45	45,45
49,81	49,81
48,62	48,62
48,20	48,20
40,72	40,72
41,16	41,16
48,41	48,41
58,17	58,17
49,07	.
48,42	48,42
48,30	48,30
44,20	44,20
57,14	57,14
45,64	.
47,76	47,76
41,07	41,07
42,38	42,38
55,07	55,07
53,68	53,68
52,92	52,92

Şekil 4.10 SPSS Programında Veri Penceresi



Şekil 4.11 SPSS Programında Çoklu Veri Atama Penceresi

Şekil 4.11'de SPSS programı çoklu veri atama (multiple imputation) seçeneğiyle yapılabilen Kayıp Veri Atama (Imput Missing Data Values) bölümden kayıp olan 2 verinin yeribne konulması için 5 çoklu veri atama işlemi gerçekleştirilmiştir.

Imputation_	VAR00001	Veriler_1	Veriler_2
1	1,00	45,45	45,45
1	2,00	49,81	49,81
1	3,00	48,62	48,62
1	4,00	48,20	48,20
1	5,00	40,72	40,72
1	6,00	41,16	41,16
1	7,00	48,41	48,41
1	8,00	58,17	58,17
1	9,00	49,07	56,82
1	10,00	48,42	48,42
1	11,00	48,30	48,30
1	12,00	44,20	44,20
1	13,00	57,14	57,14
1	14,00	45,64	51,51
1	15,00	47,76	47,76
1	16,00	41,07	41,07
1	17,00	42,38	42,38
1	18,00	55,07	55,07
1	19,00	53,68	53,68
1	20,00	52,92	52,92

Şekil 4.12 SPSS Programında Kayıp Veri Atama 1.Sonuç Penceresi

Şekil 4.12' de çoklu veri atama (multiple imputation) ile oluşturulan 1. veri atama ile çoklu veriler için atanan değerlerde olasılıklar gözetilerek kayıp veri kümesine yakın ve benzer özellikteki değerlerden veri atamaları gerçekleştirilmiş olup analizlere uygun tam veri setleri oluşturulmuştur. 1. atama sonucunda kayıp verilerin yerine 56.82 ve 51.51 değerlerinin atandığı görülmektedir.

Imputation_	VAR00001	Veriler_1	Veriler_2
2	1,00	45,45	45,45
2	2,00	49,81	49,81
2	3,00	48,62	48,62
2	4,00	48,20	48,20
2	5,00	40,72	40,72
2	6,00	41,16	41,16
2	7,00	48,41	48,41
2	8,00	58,17	58,17
2	9,00	49,07	45,53
2	10,00	48,42	48,42
2	11,00	48,30	48,30
2	12,00	44,20	44,20
2	13,00	57,14	57,14
2	14,00	45,64	54,91
2	15,00	47,76	47,76
2	16,00	41,07	41,07
2	17,00	42,38	42,38
2	18,00	55,07	55,07
2	19,00	53,68	53,68
2	20,00	52,92	52,92

Şekil 4.13 SPSS Programında Kayıp Veri Atama 2. Sonuç Penceresi

Şekil 4.13'de kayıp veriler için 2. veri atama işleminden sonra elde edilen veriler görülmektedir. 2. atama sonucunda kayıp verilerin yerine 45.53 ve 54.91 değerlerinin atandığı görülmektedir.

Imputation_	VAR00001	Veriler_1	Veriler_2
3	1,00	45,45	45,45
3	2,00	49,81	49,81
3	3,00	48,62	48,62
3	4,00	48,20	48,20
3	5,00	40,72	40,72
3	6,00	41,16	41,16
3	7,00	48,41	48,41
3	8,00	58,17	58,17
3	9,00	49,07	53,64
3	10,00	48,42	48,42
3	11,00	48,30	48,30
3	12,00	44,20	44,20
3	13,00	57,14	57,14
3	14,00	45,64	59,82
3	15,00	47,76	47,76
3	16,00	41,07	41,07
3	17,00	42,38	42,38
3	18,00	55,07	55,07
3	19,00	53,68	53,68
3	20,00	52,92	52,92

Şekil 4.14 SPSS Programında Kayıp Veri Atama 3. Sonuç Penceresi

Şekil 4.14'de görüldüğü gibi 3. atama sonucunda kayıp verilerin yerine 53.64 ve 59.82 değerlerinin atandığı görülmektedir. Her kayıp veri grubu için değer ataması gerçekleştirilirken tamamen rastgele kayıp içeren değerler belirlenerek çoklu veri atama metodu ile gerçekleştirilmiştir.

Imputation_	VAR00001	Veriler_1	Veriler_2
4	1,00	45,45	45,45
4	2,00	49,81	49,81
4	3,00	48,62	48,62
4	4,00	48,20	48,20
4	5,00	40,72	40,72
4	6,00	41,16	41,16
4	7,00	48,41	48,41
4	8,00	58,17	58,17
4	9,00	49,07	45,61
4	10,00	48,42	48,42
4	11,00	48,30	48,30
4	12,00	44,20	44,20
4	13,00	57,14	57,14
4	14,00	45,64	43,56
4	15,00	47,76	47,76
4	16,00	41,07	41,07
4	17,00	42,38	42,38
4	18,00	55,07	55,07
4	19,00	53,68	53,68
4	20,00	52,92	52,92

Şekil 4.15 SPSS Programında Kayıp Veri Atama 4. Sonuç Penceresi

Şekil 4.15'de 4. atama sonucunda kayıp verilerin yerine 45.61 ve 43.56 değerlerinin atandığı görülmektedir. Tahmin usulüne göre veri atama seçeneği için kayıp veri ile ataması gerçekleştirilen değerler arasındaki fark benzer özelliklerde ataması gerçekleştirilen verilere göre daha az görülebilmektedir.

Imputation_	VAR00001	Veriler_1	Veriler_2
5	1,00	45,45	45,45
5	2,00	49,81	49,81
5	3,00	48,62	48,62
5	4,00	48,20	48,20
5	5,00	40,72	40,72
5	6,00	41,16	41,16
5	7,00	48,41	48,41
5	8,00	58,17	58,17
5	9,00	49,07	49,31
5	10,00	48,42	48,42
5	11,00	48,30	48,30
5	12,00	44,20	44,20
5	13,00	57,14	57,14
5	14,00	45,64	54,40
5	15,00	47,76	47,76
5	16,00	41,07	41,07
5	17,00	42,38	42,38
5	18,00	55,07	55,07
5	19,00	53,68	53,68
5	20,00	52,92	52,92

Şekil 4.16 SPSS Programında Kayıp Veri Atama 5. Sonuç Penceresi

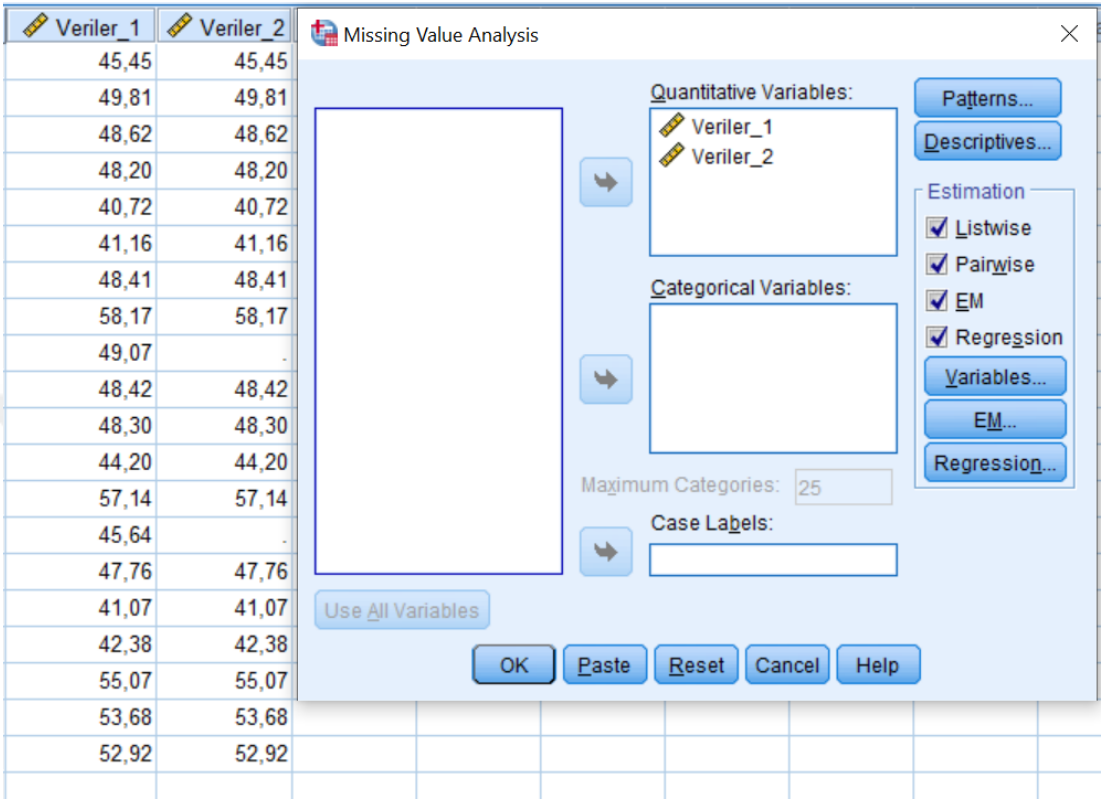
Şekil 4.16'da 5. atama sonucunda kayıp verilerin yerine 49.31 ve 54.40 değerlerinin atandığı görülmektedir.

Kayıp veriler çoklu atama yöntemiyle kayıpsız veri setleri haline getirilirken, örneklemi temsil kabiliyeti yüksek, sonuçlar için anlamlı değerler oluşturan en uygun veri veya veriler seçilmelidir.

4.1.3 Uygulama 3

Uygulama 3 için Uygulama 1'de hazırlanmış veri seti kullanışmıştır. "Veriler_1" ve "Veriler_2" olarak adlandırılan tam ve kayıp verili veri setleri SPSS programında Analyze menüsünde yer alan "Missing Value Analysis" ile analiz edilmiş ve SPSS penceresi Şekil 4.17' de gösterilmiştir. Programda kayıp veri analizinde tahmin yöntemleri olarak liste durum düzeyinde veri silme (listwise), çiftler düzeyinde veri silme (pairwise), beklenti makisimazasyonu (expectation maximizasyon) ve regresyon (regression) seçenekleri bulunmaktadır. Bu analiz ile kayıp verilerin kayıp

veri mekanizmasında dahil oldukları tamamen rassal kayıp özelliği gösterip göstermediği de belirlenmektedir. Kayıp veri analizi gerçekleştirilirken tanıtıcı istatistik değerleride elde edilebilmektedir.



Şekil 4.17 SPSS Programı Kayıp Veri Analiz (Missing Value Analysis) Penceresi

Summary of Estimated Means

	Veriler_1	Veriler_2
Listwise	48,4153	48,4153
All Values	48,3091	48,4153
EM	48,3091	48,3091
Regression	48,3091	48,2148

Şekil 4.18 SPSS Programı Kayıp Veri Analiz (Missing Value Analysis) Çıktısı

Şekil 4.18' de Kayıp Veri Analiz (Missing Value Analysis) yöntemlerinden liste durum düzeyinde veri silme (listwise), çiftler düzeyinde veri silme (pairwise), beklenti makisimazasyonu (expectation maximizasyon) ve regresyon (regression) seçenekleri ile analiz gerçekleştirilen verilerin ortalaması elde edilmiş olup, beklenti

makisimazasyonu (expectation maximizasyon) ve regresyon (regression) yöntemi ile tahmini ortalamalar elde edilmiştir.

Univariate Statistics

	N	Mean	Std. Deviation	Missing Count	Missing Percent	No. of Extremes[a] Low	No. of Extremes[a] High
Veriler_1	20	48,3091	5,14249	0	,0	0	0
Veriler_2	18	48,4153	5,39366	2	10,0	0	0

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

Şekil 4.19 SPSS Programı Kayıp Veri Analiz (Missing Value Analysis) Çıktısı

Şekil 4.18'de elde Şekil 4.18' de Kayıp Veri Analiz (Missing Value Analysis) yöntemlerinden liste durum düzeyinde veri silme (listwise), çiftler düzeyinde veri silme (pairwise), beklenti makisimazasyonu (expectation maximizasyon) ve regresyon (regression) seçenekleri elde edilen veriler için tanıtıcı istatistik sonuçları elde edilmiştir.

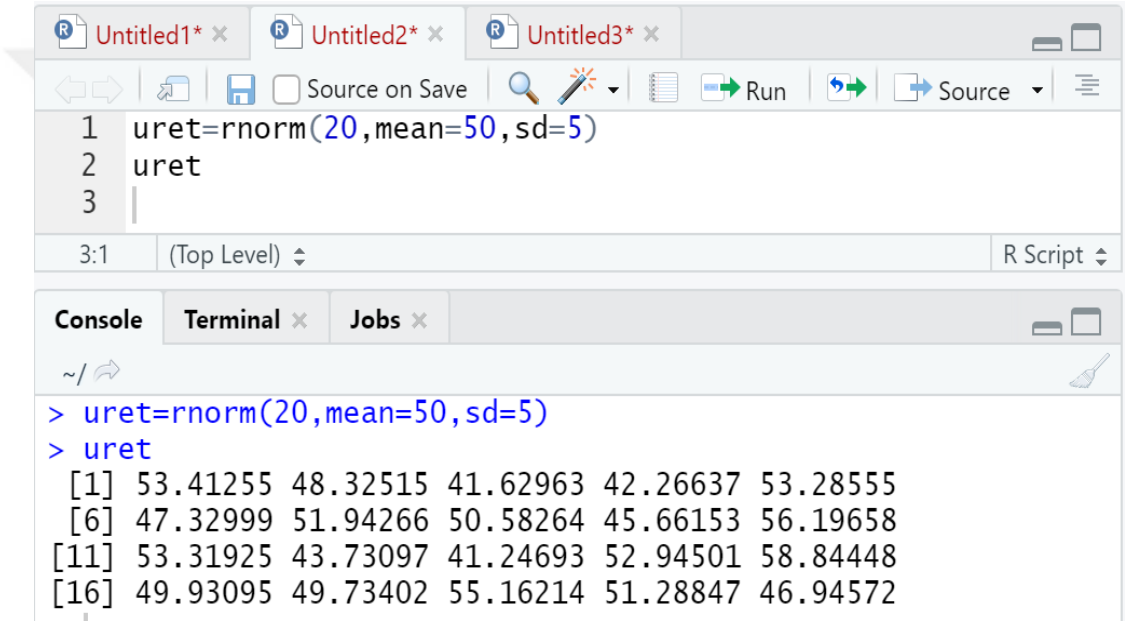
EM Correlations[a]		
	Veriler_1	Veriler_2
Veriler_1	1	
Veriler_2	1,000	1
a. Little's MCAR test: Chi-Square = ,069, DF = 1, Sig. = ,793		

Şekil 4.20 SPSS Programı Kayıp Veri MCAR Testi Çıktısı

Şekil 4.20 Kayıp Veri Analiz (Missing Value Analysis) seçeneğinde beklenti makisimazasyonu (expectation maximizasyon) algoritması ile gerçekleştirilen kayıp veri mekanizmasını belirlemek için yapılan analizde p=0,793 değeri verilerin %5'den fazla olan kayıplık durumu için TROK mekanizmasına dahil olan veri yapısına sahip olduğunu ngöstermektedir.

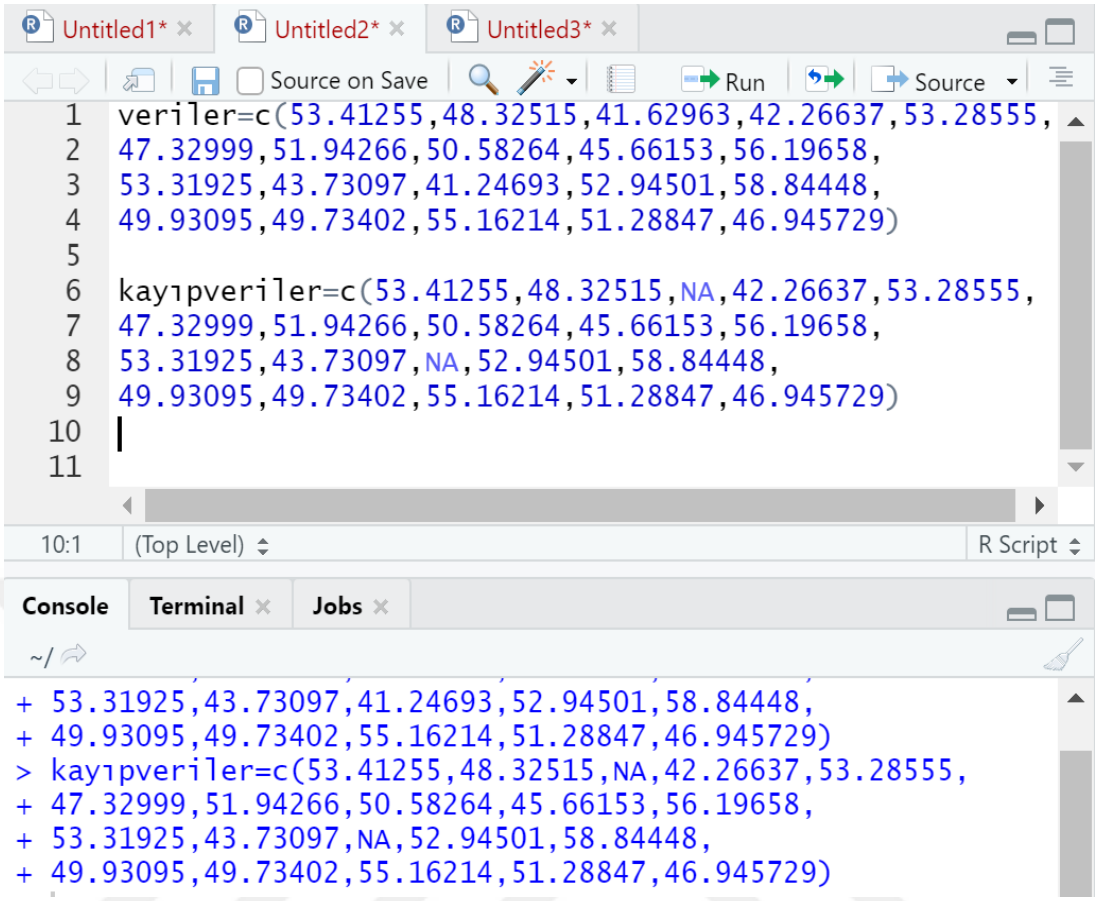
4.3 R Programı ile Uygulama

Kayıp veri analizleri için R programında üretilen veriler içerisinde 3. ve 12. numaralı verinin silinmesi ile yeniden bir veri seti oluşturulmuştur. Bu verilere R istatistik programı ile ortalama atama, yakın değerlerin ortancası ataması ve veri kayıp verileri veri setinden çıkarma seçenekleri uygulanarak kayıpsız veri setleri elde edilmesi amaçlanmaktadır. Ayrıca veriler içerisinde kayıp verinin yerinin tespiti, kayıp verisiz veri setleri oluşturma gibi komut seçenekleri ile gösterim sağlanmıştır. Kayıp olan veriler NA (Not Available) komutu ile gösterilmektedir. Simülasyon ile veri üretirilirken paremetrelerden ortalama 50, standart hata 5 kabul edilmiştir.



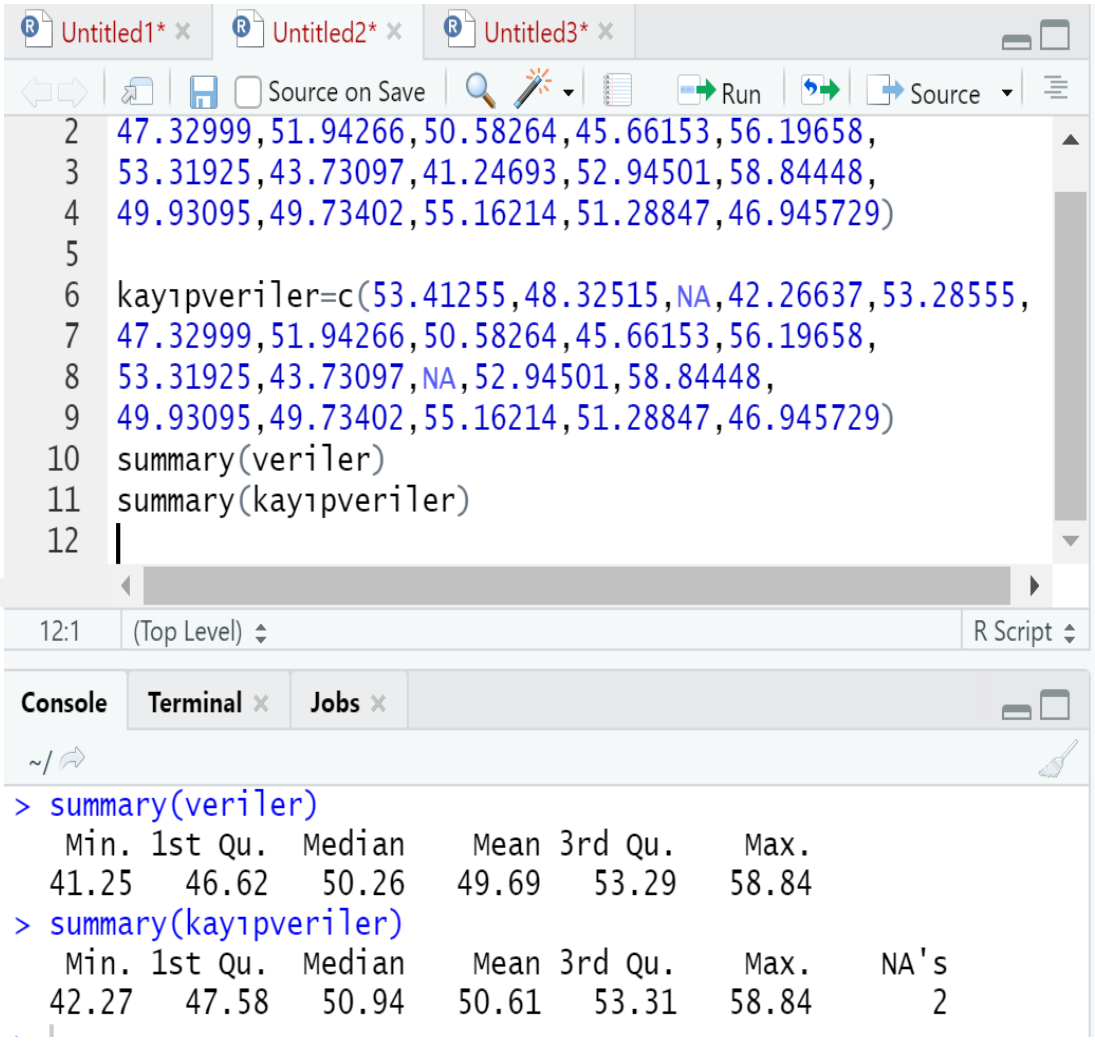
Şekil 4.21 R Programında Normal Dağılımdan Veri Üretilmesi

Şekil 4.21'de n=20 olarak üretilen veri setinde kayıp veri analizlerinin yapılabilmesi için 3. ve 12. veriler silinmiş ve Şekil 4.20'de görüldüğü gibi kayıp veri seti analize uygun hale getirilmiştir. R programı kayıp verileri NA (not available) ile belirtmektedir. R programı ile ortalama atama, yakın değerlerin ortancası ataması ve veri kayıp verileri veri setinden çıkarma seçenekleri uygulanarak kayıpsız veri setleri elde edilmesi amaçlanmaktadır. Ayrıca veriler içerisinde kayıp verinin yerinin tespiti, kayıp verisiz veri setleri oluşturma gibi komut seçenekleri ile gösterim sağlanmıştır.



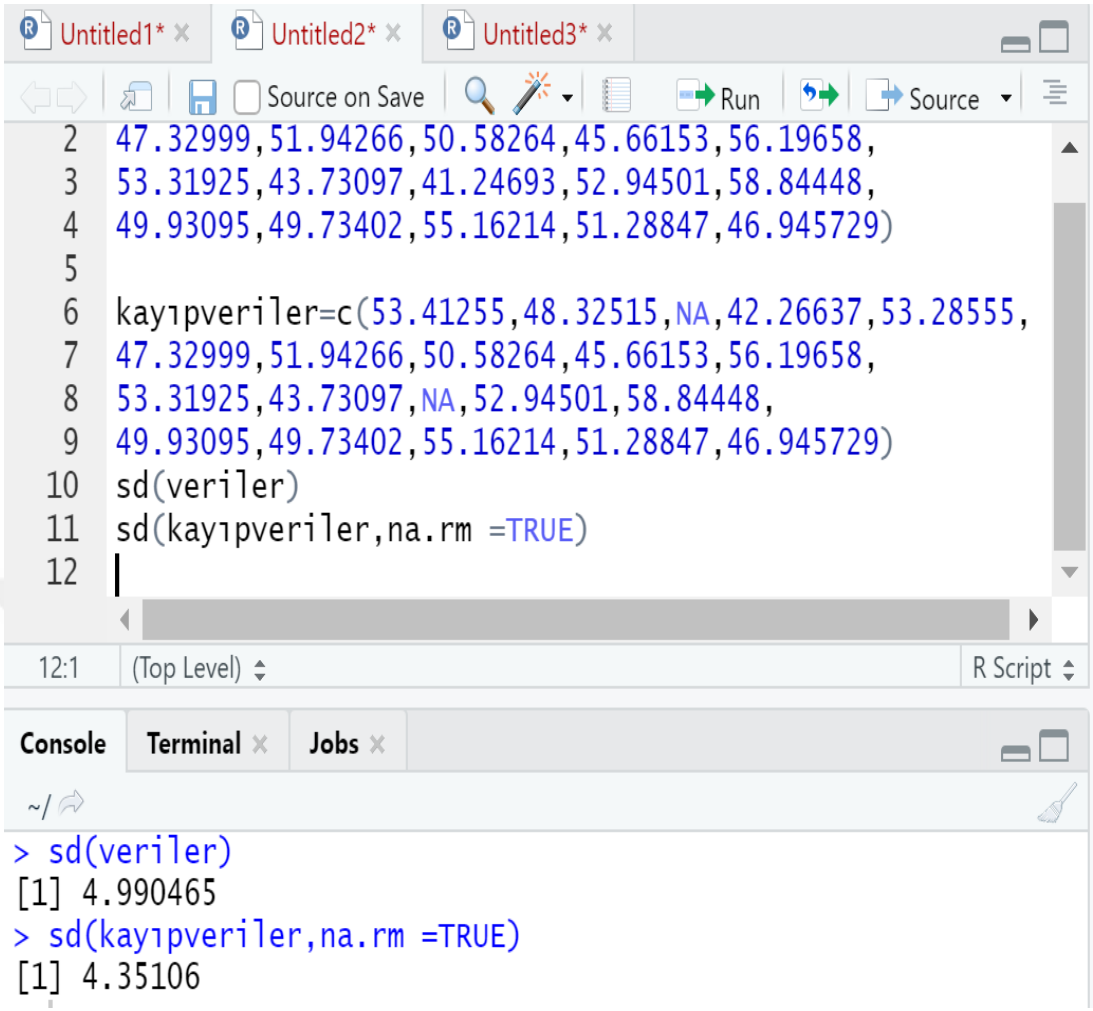
Şekil 4.22 R Programında Tam ve Kayıp Verilerin Tanımlanması

Şekil 4.22 için kayıp veri setleri analiz konsoluna aktarıldıktan sonra kayıp veri analiz seçenekleri için uygulanacak komutlar yazılarak analizler gerçekleştirilmektedir. Analizlere uygun kayıp verili set oluşturmada kayıp veri oranı ve verilerin yapısı dikkate alınarak bir mevcut durum ve kayıp verileri alanların sayısı dikkate alınarak kayıp veri analizleri gerçekleştirilmektedir.



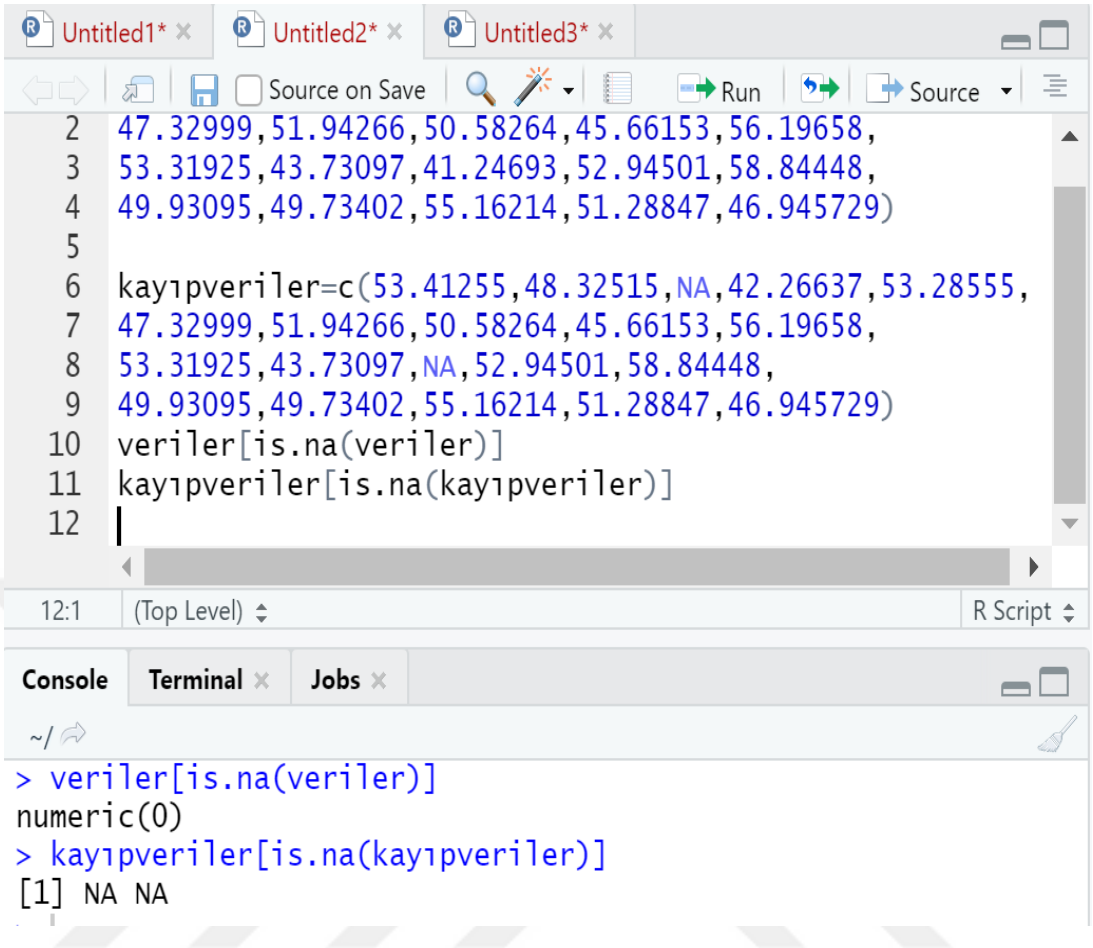
Şekil 4.23 R Programında Verilere Ait Tanıtıcı İstatistik Değerlerinin Hesaplaması

Şekil 4.23'da R programında kayıp verili gözlemler ve kayıp verisiz gözlemler için iki adet veri özeti alanı oluşturulmuş ve mean (ortalama), ortanca (median), minimum ve maksimum gibi tanıtıcı istatistik değerleri hesaplanmıştır. Kayıp verili veri setinde analizler sonucu 2 adet kayıp veri tespit edilmiş olup, mevcut değil (not available (NA) seçeneği ile istatistiksel analiz penceresine işlenmiştir.



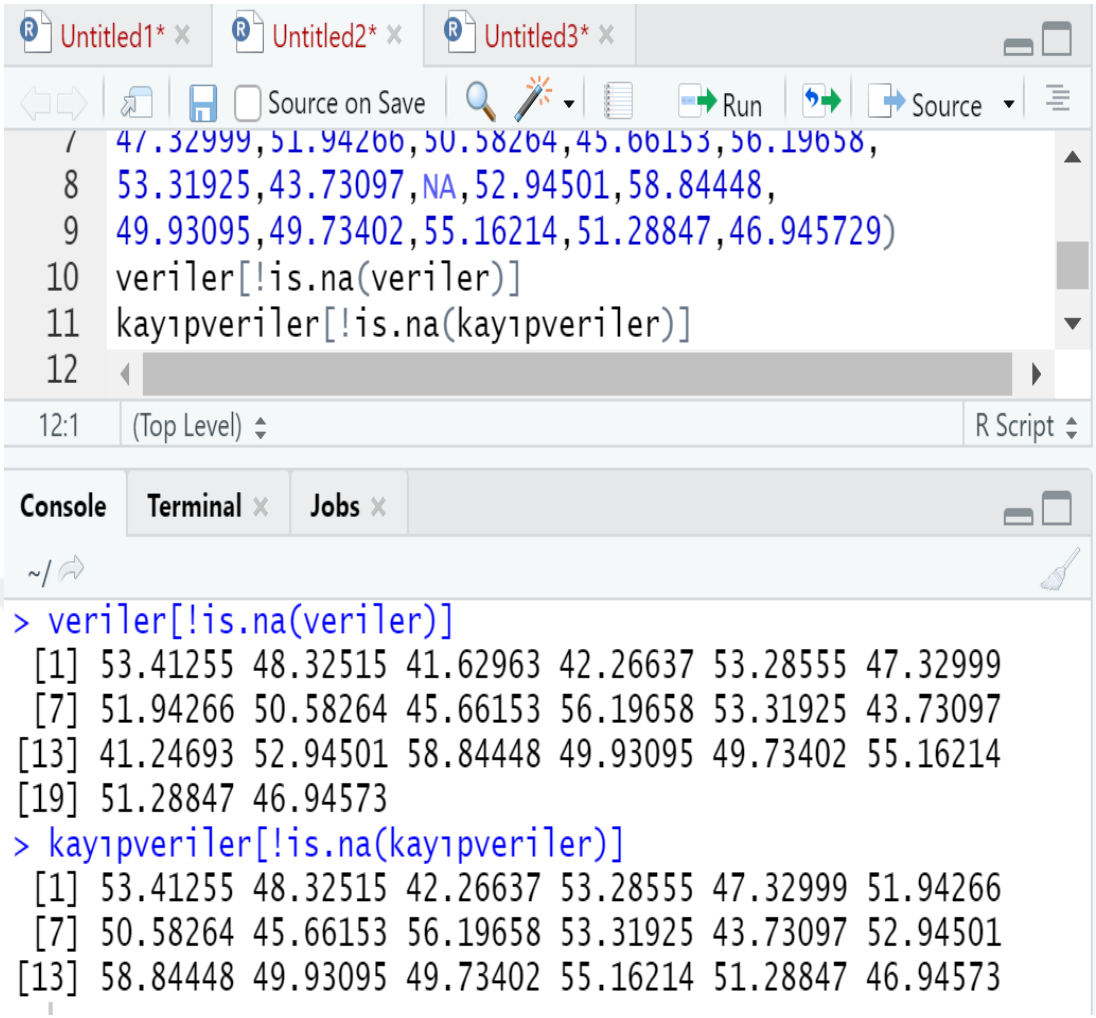
Şekil 4.24 R Programında Verilere Ait Standart Sapma Değerlerinin Hesaplanması

Şekil 4.24'de Standart Sapma değerleri kayıp verisiz ve kayıp verili veri setinde analiz edilerek kayıp veri olarak atamasını sağlayacak komut dizilimi oluşturulmuştur.



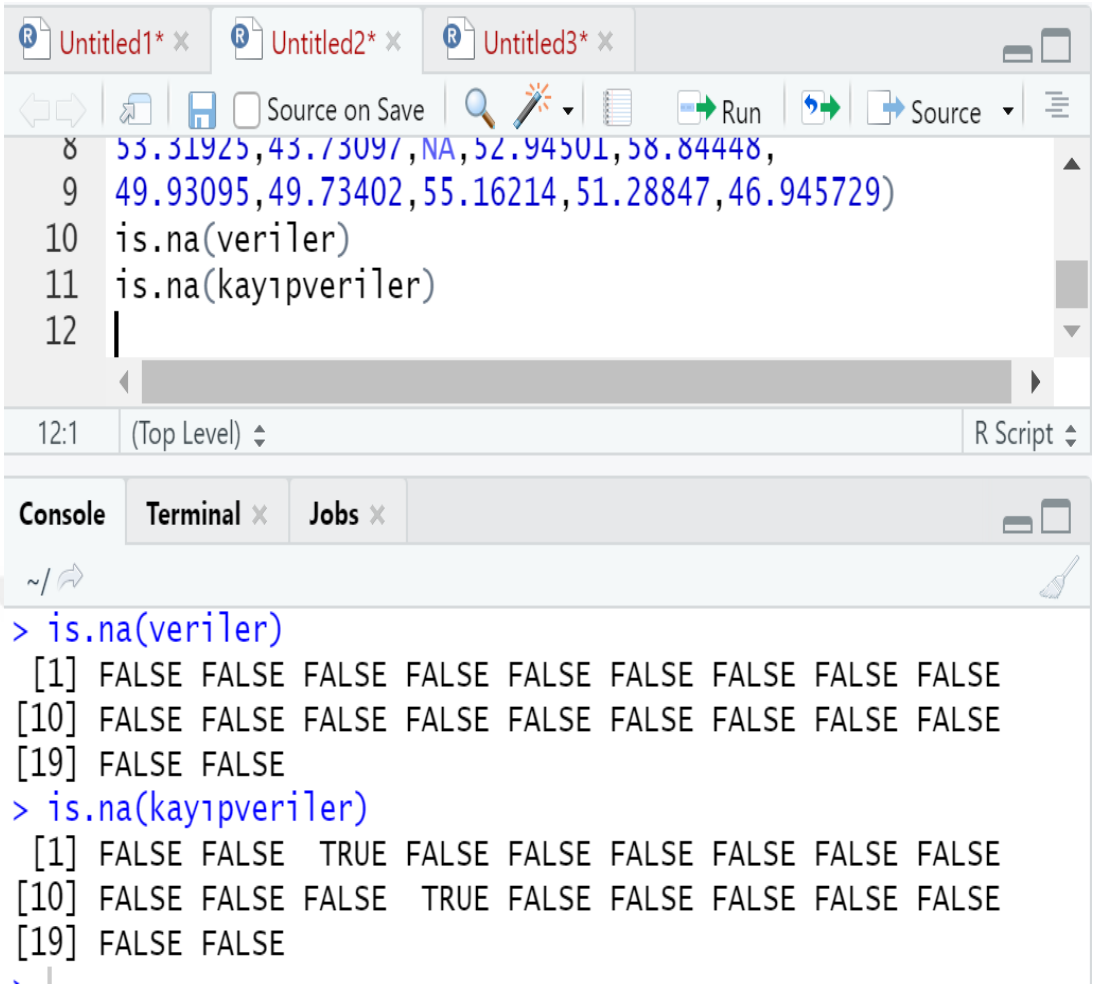
Şekil 4.25 R Programında Verilerin Kayıp Veri Sayısı Penceresi

Şekil 4.25'de kayıp veri oranları her veri grubu için ayrı gösterilmiş olup kayıp veri içermeyen veriler grubu için 0, kayıp veri grubu için 2 adet veri olarak komutlar ile veri analiz sonuç penceresinde belirtilmiştir.



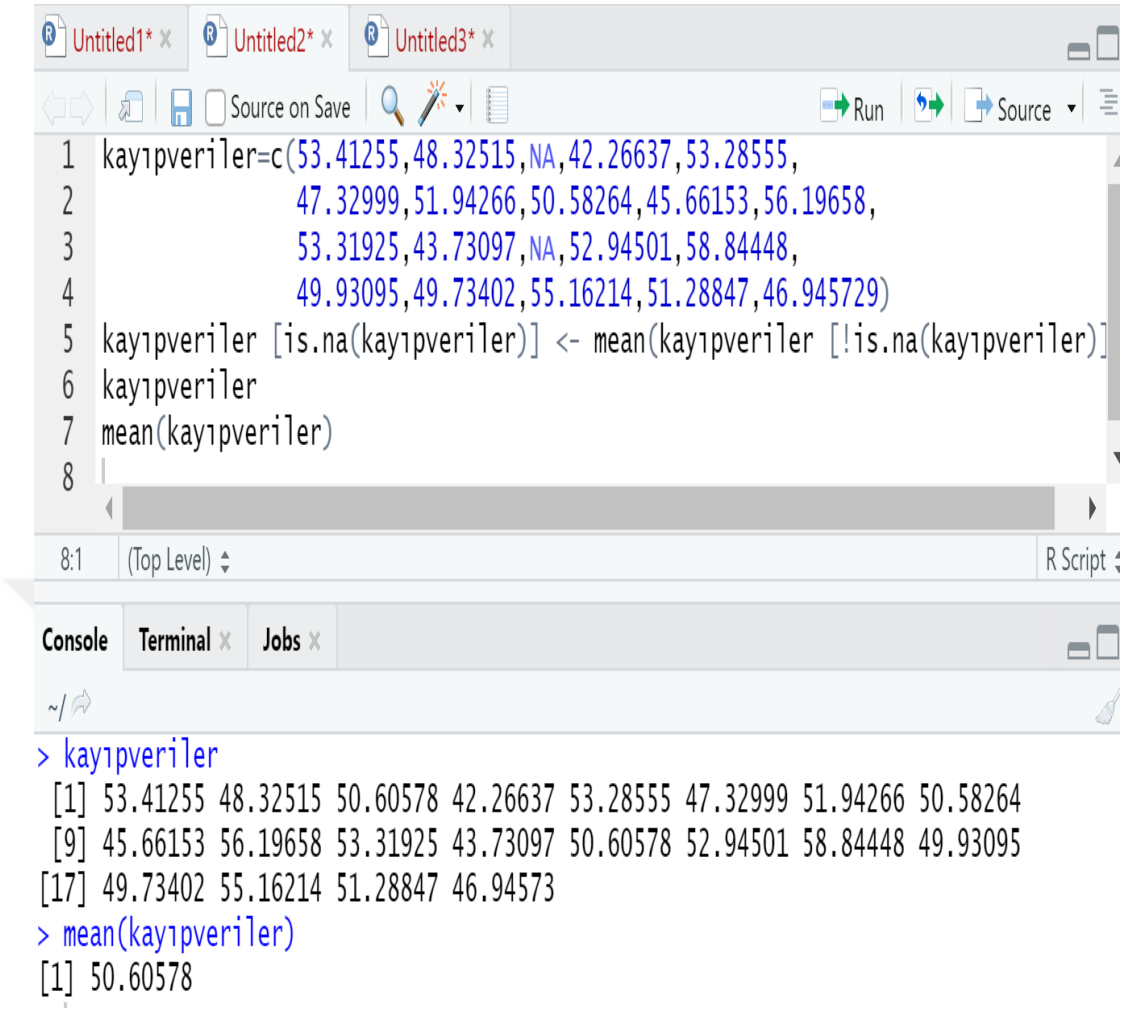
Şekil 4.26 R Programında Kayıp Veri Atama Penceresi

Şekil 4.26' de kayıpsız veri seti komutu ile "Veriler" isimli ver grubu, kayıp veri seti için ise "Kayıpveriler" başlıklı n=18 bir veri seti oluşturulmuştur. Tam veri setleri oluşturulurken veri komutlar ile analiz konsoluna aktarılır ve kayıp veri veya veriye ait gözlem değeri silinerek devam edilir.



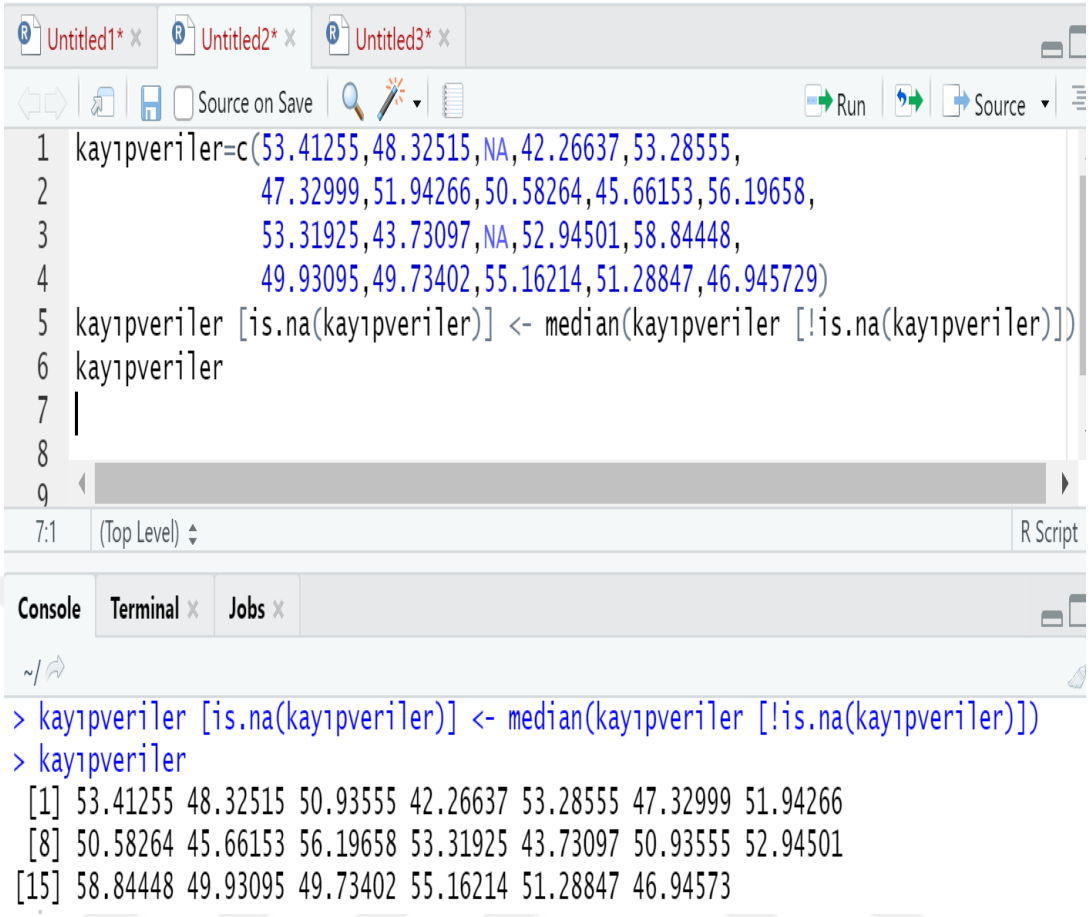
Şekil 4.27 R Programında Kayıp Veriler Yanlış (False) ve Doğru (True) Penceresi

R programında NA olarak tanımlanan değerler "True" kayıp olmayanlar ise "False" olarak gösterilmektedir. Uygulama veri setimize ait True ve False veriler Şekil 4.27'de görülmektedir. Uzunlamasına veri gruplarında bu şekilde kayıp verilerin yerleri tespit edilebilmektedir. Gözlem değerlerinde kayıp verisiz alanlar Yanlış (False) seçeneği ile belirtilebilmektedir.



Şekil 4.28 R Programında Kayıp Veriler İçin Ortalama Atama Penceresi

R Programında Kayıp Veriler İçin Ortalama Atama için yürütülen işlemlere ait program penceresi Şekil 4.28' de gösterilmiştir. Pencere de görülen R komut dizilimi ile Ortalama değer (Mean) veri setinde kayıp verili alana atanabilmekte ve tam bir veri seti haline dönönüştürülüp analizlere uygun hale getirilebilmektedir. R program veri setimizde kayıp olan veriler için 50.60578 değerini atamıştır.



Şekil 4.29 R Programında Kayıp Veriler için Ortanca Değer Ataması Penceresi

R Programında Kayıp Veriler için Ortanca Değer Atamasına ait işlemler penceresi Şekil 4.29' de gösterilmiştir. R komut dizilimi ile Ortanca değer (Median) veri setinde kayıp verili bölümlere veri ataması gerçekleştirilebilmiş ve kayıp verilerin yerine 50.93555 değerini atanmıştır. Böylece R programı ile istatistiksel analizler için kullanılabilecek tam veri seti elde edilebilmektedir.

5. SONUÇ ve ÖNERİLER

Farklı sebepler ile ortaya çıkabilecek kayıp veriler özellikle istatistik analizlerin kullanılabilmesini kısıtlamakta ve/veya sonuçlarını değiştirebilmektedir. Bir sorun olarak araştırıcıların karşısına çıkan kayıp veriler ile başetmek için çeşitli yaklaşımlar ileri sürülmüş ve kullanılması Kabul görmüş yada görmemiş bir çok yöntem geliştirilmiştir. Bu yaklaşımlardan biri olan veri silme yöntemlerinin kullanılması beraberinde bazı olumsuzlukları getirmektedir. Özellikle deney ünitesi sayısının azalacak olması beraberinde istatistiksel hataların artmasına ve kullanılacak testlerin gücünün azalmasına neden olabilir. Bu sebeple kayıp verilerin tahmin edilerek yerine atanması gerekliliği ortaya çıkmaktadır. Bu amaçla önerilen çok sayıda yaklaşım veya geliştirilen yöntem bulunmaktadır. Örneğin, regresyon atama, tekli atama, çoklu atama gibi tahmin odaklı bir model oluşturmada kayıp veri setine uygun bir denklem oluşturularak veriler belirlenebilmektedir. Bunlar arasından hangisini kullanılacağı kayıp verinin oluşum mekanizması, veri tipi, verilerin analizinde kulanılacak istatistik yöntem vb. durumlara göre değişkenlik göstermektedir.

Araştırma gruplarında eksik olan bölüme karşılık gelen veri grubu silinerek kayıpsız bir örnek oluşturulmaktadır. Bu iki yöntemle kayıp veri ile ilgili çözümlemeler geliştirilmektedir. Ortalama değer atama da elde edilen kayıp veri tahmini olarak dikkate alınarak eksik olan, belirli bir zaman dilimine yayılmış veri gruplarına uygun veri atama gerçekleştirilirilebilmektedir. Kayıp veri atamalarda veri seti bir bütün olarak incelendiğinde ortalama, standart hata, varyasyon gibi istatatiksel analizler sırasıyla gerçekleştirilerek en uygun kayıp veri grubu oluşturulmaktadır. Mevcut yöntemler dikkate alındığında hangi yöntemin hangi veri grubu için uygun olduğu ile ilgili değerlendirmeler mevcut olmaktadır. Bu değerlendirmeleri destekleyen veri eksiltme ve atamalar doğru sağlanamazsa farklı yöntemlerde farklı sonuçlar ortaya çıkmaktadır.

Analitik değerlendirmelere göre kayıp veri için tahmin, silme ve karar verme gibi kriterler gözlenen veri sayısı ve verilerin özelliklerine göre değişim göstermektedir. Verilerin birbiri ile olan ilişkileri veri sayısı ve toplam eksik veri sayısı değerlendirilerek kayıp veri analizinde doğru yöntem seçilebilmektedir. Bu yöntemler beraber değerlendirildiğinde bir yöntemin diğerine göre herhangi bir üstünlüğü bulunmamaktadır.

Veri dağılımına göre kayıp verisi az olan kayıp verili örneklemlerde yerine koyma yöntemi kullanılmaktadır. Kayıp verinin yoğun olduğu veri setlerinde yerine koyma yöntemi varyans değerinin artmasına verilerin homojen bir şekilde dağılmamasına yol açar. Veri silme yöntemide tamamen rassal olan kayıp mekanizmasında kullanılabilir bir yöntemdir. Bunun dışında kalan diğer kayıp veri mekanizmalarında varyansı artırıcı bir durum söz konusudur.

Kayıp veri setleri mevcut veri setlerine benzer özellik gösteriyorsa beklenti maksimizasyonu yöntemine göre değer belirlenebilmektedir.Bu yöntem birbiri ile bağlantısı olmayan verilerde çalışmayan bir yöntem olabilmektedir. Hot Deck atama yönteminde kayıp veriler arasındaki mesafe belirlenerek boş olan kısımlara veriler işlendiğinden hata payının yüksek çıkma olasılığı fazladır. Kayıp veri gruplarının az olduğu durumlarda hata payı da düşük olmaktadır.

Son gözlem değerini ileri taşımada kayıp veri aralığında birbirine eş değerler olan kısımlarda kullanımı yanlı sonuç vermeyi engeller ve hata payını düşük tutar. Bayes'çi veri atama yöntemi tahmini bir veri modeli üzerinden oluşturulan çok sayıda kayıp veri grubunda kullanıldığında daha tarafsız sonuçlar ortaya koyabilmektedir. Bu yöntemde tüm veriler olasılıklar dahilinde kullanılmaktadır. Regresyon atamasında korelasyon yüksek tutulduğu kayıp verili alanlar seçilip bir regresyon modeli türetilir. Birbiri ile ilişkili olan veri gruplarında sonuçlar hata payını düşük tutar.

Kayıp veri miktarı, verilerin özelliklerine uygun tahmin yöntemiyle değer elde etmede önem teşkil etmektedir. Regresyon ataması gerçekleştirilirken veri gruplarına bakılıp önce korelasyon değeri yüksek olan iki adet alan seçilip ona göre bir regresyon modeli gerçekleştirilebilmektedir. Beklenti maksimizasyonunda maksimumum benzerlik prensibine bağlı çalışan bir yöntem olduğunda verilerin tamamının kullanılması gerekmektedir. Bu prensibe göre kayıp gözlem değerine sahip veri grubuna değer aralığı az olan ve büyük veri grubuna sahip benzer özellikli değerlerin atanması sonuçların tarafsız ve doğru olmasını sağlamaktadır.

Hot Deck değer atamada veriler arasındaki mesafe değerlendirilerek sabit sayılar kayıp verili alanlara eklendiğinde hata oranı yüksek olan bir algoritma olarak şekillenmektedir.Bu yöntem kayıp verinin yoğun olmadığı durumlarda kullanımı kolay olup hata oranını fazla etkilememektedir. Karar ağacı ile analizde kayıp verilerin fazla olması anlamlı sonuçlar oluşturmada yetersiz kalacaktır. Karar ağacını oluşturan

yapı karmaşıklaştıkça iki ayrı değerlendirme kümesinden biri olan eğitim kümesinin doğruluğu artmakta, diğer bir küme olan sağlama kümesinin doğruluğu azalmaktadır. Yapay sinir ağları ile veri atama yöntemi diğer yöntemlere göre anlamlı sonuçlar oluşturmada yeni nir yöntem olması sebebiyle yetersiz kalabilmektedir.

Sonuç olarak, kayıp veri sorunu özellikle bazı istatistik analiz yöntemlerinde testin gücünü etkilemesi sebebiyle baş edilmesi gereken önemli bir sorundur. Kayıp veri sorunu ile başetmede veri tipinin uygunluğu esas alınmalıdır. Kayıp veri başetme yöntemleri arasındaki farklılığı kayıp verinin özellikleri ve kayıp verinin miktarı belirlemektedir.

6. KAYNAKLAR

Afifi, AA. & Elashoff, RM. (1966). Missing observations in multivariate statistics I. Review of the literature. *Journal of the American Statistical Association*, 61(315), 595-604.

Akbaş, U. ve Koğar, H. (2020). Nicel Araştırmalarda Kayıp Veriler ve Uç Değerler: Çözüm Önerileri ve SPSS Uygulamaları. Pegem Akademi.

Arıkan, ÇA. & Soysal S. (2018). *Güvenirlik katsayılarının kayıp veri atama yöntemlerine göre incelenmesi* Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 33(2), 316-336.

Alkan, N. (2012). Kayıp verili COX regresyon yöntemine bayesci bir yaklaşım. Doktora Tezi, Ondokuz Mayıs Üniversitesi, Fen Bilimleri Enstitüsü, Samsun.

Allison, PD. (2001). Missing data, sage university papers series on quantitativ eapplications in the social sciences, ThousandsOaks, CA, Sage.

Allison, PD. (2002). *Missing Data.* Thousands Oaks: Sage Publication.

Allison, PD. (2003). Missing Data Techniques for Structural Equation Modeling. *Journal of Abnormal Psychology*. *112*(4), 545-557.

Allison, PD. (2009). Missing data (Sage University Paper Series on Quantitative Applications in the Social Sciences, 72-89). London: Sage Publication.

Alpar, R. (2003). Uygulamalı çok değişkenli istatistiksel yöntemlere giriş-1, Nobel Kitabevi.

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

Bal, C. (2003). Çok gruplu veri setlerinde eksik gözlem sorununun çözümlenmesi ve sağlık Alanında Bir Uygulama, Yayınlanmamış Doktora Tezi, Eskişehir Osmangazi Üniversitesi, Sağlık Bilimleri Enstittisü.

Baygül, A. (2007). Kayıp veri analizinde sıklıkla kullanılan etkin yöntemlerin değerlendirilmesi, Yüksek Lisans Tezi, İstanbul Üniversitesi Sağlık Bilimleri Enstitüsü, İstanbul.

Bayhan, A. (2018). Farklı koşullardaki kayıp veri oranının iç tutarlığa etkisi. Yüksek Lisans Tezi, Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Byrne, BM. (2000). Structural equation modeling with AMOS; basic concepts, applications, and programming. Testing for invariant factorial structure of a theoretical construct (First-order CFA model).

Carpita, M. & Manisera, M. (2011). On the imputation of missing data in surveys with Likert-type scales. *Journal of Classification,* 28(1), 93-112.

Cool, AL. (2000). A review of methods for dealing with missing data (rapor). *Annual Meeting of the Southwest Educational Resarch Association*. Dallas.

Çokluk, Ö. & Kayri, M. (2011). Kayıp Değerlere Yaklaşık Değer Atama Yöntemlerinin Ölçme Araçlarının Geçerlik ve Güvenirliği Üzerindeki Etkisi. *Kuram ve Uygulamada Eğitim Bilimleri*. *11* (1), 289-309.

Çüm, S., Demir, EK., Gelbal, S., & Kışla, T. (2018). Kayıp veriler yerine yaklaşık değer atamak için kullanılan gelişmiş yöntemlerin farklı koşullar altında karşılaştırılması. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, (45), 230-249.

De Luca, G. & Peracchi, F. (2007). A sample selection model for unit and item nonresponse in cross-sectional surveys. *CEIS Tor Vergata-Research Paper Series,* 33(99), 1-45

Demir, E. (2013). Kayıp verilerin varlığında çoktan seçmeli testlerde madde ve test parametrelerinin kestirilmesi. *Eğitim Bilimleri Araştırma Dergisi* (3), 47-68.

Dempster, AP., Laird, NM. & Rubin, DB. (1977). Maximum likelihood from incomplete data via the EM Algorithm, *Journal of the Royal Statistical Society,* 39, 1-38.

Diggle, P. & Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *43*(1), 49-73.

Eberle, W. & Toutenburg, H. (1999). Handling of Missing Values in Statistical Software Packages for Windows. *Paper: Institut für Statistik*, Sonderforschungsbereich 386, s.170.

Enders, CK. (2011). Analyzing longitudinal data with missing values. *Rehabilitation psychology*, *56*(4), 267.

Enders, CK. (2010). Applied missing data analysis. New York: Guilford Press

Enders, CK. & Bandalos, DL. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural equation modeling*, *8*(3), 430-457.

Fox-Waslylyshyn, SM. & El-Masri, MM. (2005). Focus on research methods: handling missing data in self-report measures. Research in Nursing&Health, 28, 488-495.

Garson, D. (2015). Missing values analysis and imputation methods. USA: Statistical Publishing Associates.

Groves, RM. (2006). Nonresponse rates and nonresponse bias in house hold surveys. Public Opinion.

Hasan, H., Ahmad, S., Osman, BM., Sapri, S. & Othman, N. (2017). A comparison of model-based imputation methods for handling missing predictor values in a linear regression model: A simulation study. AIP Konferansı, 8-9 Ağustos 2017, s. 60003

Heerwegh, D. (2005). Web surveris. Explaining and reducing unit nonresponse, Item nonresponse, item nonresponse and patial-nonresponse. (Doktora tezi, Katholike Universiteit Leuven Faculteit Sociale Wetenschappen)

Heyting, A., Tolboom, JTBM. & Essers, JGA. (1992). Statistical handling of drop-outs in longitudinal clinical trials. *Statistics in medicine*, 11(16), 2043-2061.

Horton, NJ. & Kleinman, KP. (2007). Much ado about nothing: a comparison of missing data methods. *American Statistical Association,* 61, 79-90.

Howell, DC. (2007). The Treatment of Missing Data. W. Outhwaite, & S. P. Turner içinde, The SAGE handbook of social science methodology (s. 208-224). Los Angeles: Sage Publications.

Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. Quality&Quantity, 34, 331-351.

Kaspar, ÇE. (2011). Kayıp veriler ve kayıp veriler için bir çoklu veri atama yöntemi: propensity skor, Doktora Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü. Ekonometri Anabilim Dalı. İstatistik Bilim Dalı. İstanbul

Kürşad, MŞ. & Nartgün, Z. (2015). Kayıp veri sorununun çözümünde kullanılan farklı yöntemlerin ölçeklerin geçerlik ve güvenirliği bağlamında karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(2), 254-267.

Little, RJA. & Rubin, DB. (2002). *Statistical analysis with missing data*, Second Edition, Wiley, New York.

Little, RJA. (1998). A test of missing completely at random for multivariate veri with missing values. Journal of the American Statistical Association 38, 11981202

Little, RJA. & Rubin, DB.(1987). Statistical analysis with missing data. New York: Wiley

Longford, N. (2005). Missing data and small-area estimation: Modern analytical equipment for the survey statistician. Springer Science & Business Media.

Mertler, CA. & Vannatta, RA. (2005). Advanced and multivariate statistical methods: Practical application and interpretation. Glendale, CA: Pyrczak Publishing.

Molenbergh G. & Verbeke G. (2005).Models for Discrete Longitudinal Data, Springer, New York.

Oğuzlar, A. (2001). Alan araştırmalarında kayıp değer problemi ve çözüm öneriler. 5. Ulusal Ekonometri ve İstatistik Sempozyumu, Adana: Çukurova Üniversitesi, 20-22 Eylül 2001, s.1-28.

Osborne, JW. (2013). Best practices in data cleaning. California: Sage Publication, Inc.

Öztemel, E. (2003). Yapay Sinir Ağları. Papatya Yayıncılık, İstanbul.

Öztemur, B. (2014). Kayıp veri yöntemlerinin farklı değişkenler altında varyans analizi (t-testi, anova) parametreleri üzerine etkisinin incelenmesi. (Yayımlanmamış Yüksek Lisans Tezi). Abant İzzet Baysal Üniversitesi, Eğitim Bilimleri Enstitüsü, Bolu.

Peng, CYJ., Harwell, M., Liou, SM. & Ehman, LH. (2006). Advances in Missing Data Methods and Implications for Educational Research. S. S. Sawilowsky içinde, *Real Data Analysis* (s. 31-78). New York.

Pigott, TD. (2001). A review of methods for missing data. *Educational Resarch and Evaluation*, 7(1), 353-383.

Roth, PL. (1994). Missing data: A conceptual review for applied psychologists. Personnel Psychology, 3(1), 537-560

Rubin, DB. (1976). Inference and missing data. Biometrika, 581-592.

Schafer, JL. (1997). Analysis of incomplete multivariate data. New York: Chapman&Hall/Crc.

Schafer, JL. (1999). Multiple imputation: a primer. *Statistical Methods on Medical Resarch*, 8(1), 3-15.

Schafer, JL. & Graham, JW. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.

Sezgin, Çelik. (2013). Akademik Bilişim 2013-XV. Akademik Bilişim Konferansı Bildirileri. 23-25 Ocak 2013-Akdeniz Üniversitesi, Antalya

Sıddıkoğlu, D. (2019). Kayıp veri problemi ve farklı değer atama yaklaşımlarının whodas-2.0 ölçeğinin psikometrik özellikleri üzerindeki etkisinin incelenmesi, Doktora tezi, Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara.

Sinharay, S., Stern, HS. & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods* (6), 317-329.

Soley Bori M. (2013). Dealing with missing data: key assumptions and methods for applied analysis, Technical report No. 4.

Soysal, S. & Akın Arıkan Ç. (2017). Kayıp veri atama yöntemlerinin faktörleştirme teknikleri üzerindeki etkisi. Pegem İndeks.

Şeker, ŞE & Eşmekaya, E. (2017). Eksik Verilerin Tamamlanması (Imputation), YBS Ansiklopedi, 4(3), 10-17.

Tabachnick, BG. & Fidell, LS. (2014). Using multivariate statistics. USA: Pearson Education Limited

Tabachnick, B. & Fidell, L. (1996). Using multivariate statistics (3th ed.). New York: Herper Collins College Publishers

Wasito, I. (2003). Least squares algorithms with nearest neighbour techniques for ımputing missing data values. Doktora Tezi, University of London, 9-28.

Yazıcı, F. (2005). EM algoritması ve uzantıları, Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Fatih Kale
Doğum Yeri	Ordu
Doğum Tarihi	17.10.1982
Uyruğu	☒T.C. ☐Diğer:
Telefon	531 629 45 37
E-Posta Adresi	fatih.kale8682@gmail.com

Eğitim Bilgileri	
Lisans	
Üniversite	Ondokuz Mayıs Üniversitesi
Fakülte	Ziraat Fakültesi
Bölümü	Zootekni Bölümü
Mezuniyet Yılı	02.02.2006