KAYIP VERİLER YERİNE YAKLAŞIK DEĞER ATAMAK İÇİN KULLANILAN GELİŞMİŞ YÖNTEMLERİN FARKLI KOŞULLAR ALTINDA KARŞILAŞTIRILMASI

A COMPARISON OF ADVANCED METHODS USED FOR MISSING DATA IMPUTATION UNDER DIFFERENT CONDITIONS

Sait ÇÜM* Elif Kübra DEMİR** Selahattin GELBAL*** Tarık Kışla****

Başvuru Tarihi:03.08.2017 Yayına Kabul Tarihi: 08.02.2018 DOI: 10.21764/ maeuefd.332605

Özet: Bu araştırmada, farklı oranlarda (%15 ve %25) ve yapılarda (TROK ve ROK) oluşturulan kayıp veriler yerine farklı yöntemlerle yaklaşık değer atanması sonucu elde edilen veri setlerinin tam veri setleriyle karşılaştırılarak incelenmesi amaçlanmıştır. Bu araştırma, PISA'ya (2012) Türkiye'den katılan 15 yaş grubundaki 4848 öğrenci arasından matematik özyeterliği anketine katılan ve eksiksiz bir şekilde yanıtlayan 3129 öğrencinin puanlarından oluşan veri seti üzerinde yürütülmüştür. Söz konusu veri seti içerisinden farklı yapılar oluşturulacak şekilde farklı oranlarda veri silinerek eksik veri setleri oluşturulmuştur. Bu eksik veri setleri BM, BVA, ESE, MUA, MZMC ve RA olmak üzere altı farklı gelişmiş değer atama yöntemiyle tamamlanmıştır. Söz konusu yöntemlerle yapılan yaklaşık değer atamaları sonucu elde edilen ölçek puanları ile tam veri ölçek puanları arasındaki korelasyon değerlerinin yüksek olduğu görülmüştür. Benzer şekilde farklı yöntemlerle tamamlanmış veri setlerinden elde edilen ölçek puanları arasındaki korelasyon değerleri de yüksek bulunmuştur. Tam veri seti ile tamamlanmış veri setlerinden hesaplanan ölçek puanları arası farkların mutlak değer toplamları ve ortalamaları göz önünde bulundurulduğunda belirlenen koşullar altında en iyi çalışan yaklaşık değer atama yöntemlerinin MZMC ve BM olduğu sonucuna ulaşılmıştır.

Anahtar Sözcükler: *Kayıp Veri, Yaklaşık Değer Atama Yöntemleri*

Abstract: In this study, it is aimed to comparatively research of data sets obtained imputation for missing values that is formed by different ratios (%15 and %25) and in different structures (MCAR and MAR) with different methods. This study has been conducted on data set formed by points of 3129 students who participated in mathematics self-efficacy survey and answered it completely among 4848 students -age group of 15- who participated in PISA 2012 from Turkey. Missing data sets have been constituted by deleting data in different ratios to be constitute different structures in the data set. These data sets have been completed by six different nearby value imputation including EM, BIM, PSM, MCMC, MDIM, and RIM. Obtained data sets have been compared with full data sets by scale points of students. In the scope of the research, correlation between obtained scale points and scale points of real data has been seen quite high. Similarly, when scale points is considered, correlation of missing data imputation methods with each other have also been found quite high. Considering the difference between the totals and avarages of student scores calculated from the full data set and imputed data sets EM and MCMC is founded that the best missing data imputation methods under all conditions.

Keywords: *Missing Data, Missing Values, Imputation Methods*

* MEB, Kırıkkale-Türkiye, e-posta: saitcum@hotmail.com ORCID ID: 0000-0002-0428-5088
** Arş. Gör., Ege Üniversitesi, Eğitim Fakültesi, İzmir-Türkiye, e-posta: elif.kubra.demir@ege.edu.tr ORCID ID: 0000-0002-3219-1644
*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, e-posta: sgelbal@gmail.com ORCID ID: 0000-0001-5181-7262
**** Doç. Dr., Ege Üniversitesi, Eğitim Fakültesi, İzmir-Türkiye, e-posta: tarik.kisla@ege.edu.tr ORCID ID: 0000-0001-9007-7455

Giriş

Araştırmacılar her ne kadar eksiksiz veri elde etme çabasında olsalar da üzerinde çalışılmak istenilen veriler bazı nedenlerle istenildiği gibi eksiksiz bir şekilde toplanamayabilir. Özellikle büyük gruplar üzerinde yürütülen çalışmalarda eksiksiz veri setlerinin elde edilmesi neredeyse olanaksızdır (Cool, 2000). Bu durum, araştırmaya katılan bireylerin bilinçli ya da bilinçsiz bir şekilde bazı soruları yanıtsız bırakması, belirlenen süre içerisinde bazı maddelere erişememesi veya çeşitli nedenlerle veri toplama sürecinin bazı aşamalarında bulunamaması gibi katılımcılardan kaynaklı sebeplerden dolayı oluşabilmektedir. Katılımcılardan kaynaklı olmayan veri kayıpları ise araştırmada kullanılan veri toplama aracının teknik özelliklerinin yetersizliği, aracın uygulanma koşullarının elverişli olmaması, araştırmacının veri girişi sürecinde dikkatsizliği ya da yorgunluğu gibi nedenlerle ortaya çıkabilmektedir. Veri setlerindeki bu eksiklikler kayıp veriler olarak adlandırılmaktadır. Kayıp veriler analizler için kullanılacak olan istatistiksel yöntemlerin hemen hemen hepsi için önemli bir sorun oluşturur çünkü tüm yöntemler veri setinin eksiksiz olduğu varsayımı altında geliştirilmiştir (Allison, 2003; Osborne, 2013; Pigott, 2001). Kayıp veriler dikkate alınmadan yapılacak olan analizler yanıltıcı sonuçlar verebilir. Ayrıca bu sonuçların araştırma evrenine genellenebilirliği noktasında hata miktarının artacağı da söylenebilir (Byrne, 2000). Bu bakımdan, toplanan veriler üzerinden yürütülecek olan ileri düzey analizlere geçilmeden önce kayıp verilerin miktarı ve yapısıyla ilgili incelemelerin yapılması ve karşılaşılabilecek sorunların bertaraf edilmesi için gerekli önlemlerin alınması araştırmacılar için bir zorunluluk olarak ortaya çıkmaktadır.

Kayıp veriler değişkenlere bağlı olarak üç farklı yapısal özellik gösterebilir. Bunlar (Allison, 2001):

Tamamıyla Rastlantısal Olan Kayıp Veriler (TROK):

Bu türden kayıp veriler, değişkenin düşük ya da yüksek değerler almasıyla ilişkili olmayan ve başka bir değişkenin etkisiyle ortaya çıkmayan veri seti içerisine tamamen rastlantısal olarak dağılmış kayıplar olarak ifade edilebilir. Örneğin, bireylerin aylık gelir miktarlarından oluşan bir değişkende gözlemlenen kayıpların yüksek ya da düşük gelire sahip olan bireylerde farklılaşmaması ya da başka bir değişkenin (cinsiyet, yaş vb.) kayıplar üzerinde bir etkisinin olmaması durumunda kayıp verilerin tamamıyla rastlantısal dağıldığı (TROK) yorumu yapılabilir.

Rastlantısal Olan Kayıp Veriler (ROK):

Değişkende yer alan kayıpların başka değişkenlerle ilişkili olduğu fakat değişkenin kendisindeki değişimle ilişkili olmadığı anlamına gelen kayıp verilerdir. Benzer şekilde, aylık gelir değişkenindeki kayıpların başka bir değişkenle (cinsiyet, yaş vb.) ilişkili olması fakat değişkenin kendi değerlerindeki değişimden etkilenmemesi durumu örnek verilebilir. Erkek bireylerin aylık gelirleri hakkında bilgi vermekten kaçınması fakat cinsiyet değişkeni kontrol edildiğinde gelirin düşük ya da yüksek olmasının veri kaybı olasılığını etkilememesi durumunda kayıp verilerin rastlantısal dağıldığı (ROK) söylenebilir.

İhmal Edilemez Kayıp Veriler (İEK):

Bu tür yapılarda, değişkendeki veri kaybı olasılığı hem diğer değişkenlerle ilişkili hem de değişkenin kendi değerleriyle ilişkilidir. Aynı örnek üzerinden gidilirse, bireylerin aylık gelirlerine ilişkin bir değişkendeki kayıplar başka değişkenlerden (cinsiyet, yaş vb.) etkilenebildiği gibi bu değişkenler kontrol altında tutulduğunda kendi değerlerindeki değişimden de etkilenmektedir. Erkeklerin aylık gelirleri hakkında bilgi saklamaya kadınlardan daha eğilimli olması ve bu değişkenin kontrol edilmesi durumunda da aylık geliri yüksek olan bireylerin bu konuda bilgi saklamaya daha eğilimli olması bu tür bir kayıp veri dağılımının oluşmasına neden olabilir. İhmal edilemez kayıp verilerle başa çıkmak diğer kayıp verilere göre daha zor ve uzmanlık gerektiren bir iştir.

Kayıp verilerin varlığında araştırmacılar, (1) veriye yeni gözlemlerin eklenmesi, (2) kayıp verili gözlemlerin veri setinden çıkartılması, (3) kayıp verilere ilişkin kestirimlerin yapılması ve elde edilen yaklaşık değerlerin kayıp veriler yerine kullanılması yöntemlerinden birini kullanarak olası sorunlara yönelik önlem alabilirler. Veriye yeni gözlemler ekleme sürecinin zaman ve emek maliyeti ortaya çıkaracağı göz önünde bulundurulmalıdır. Eksik verili gözlemlerin veri setinden çıkartılması ise gözlem sayısında ciddi bir azalmaya yol açabilir ve yeterli sayıda oluşturulmuş bir örneklem yetersiz sayıdaki bir örnekleme dönüşebilir. Bu durum yapılacak olan istatistiksel analizlerin gücünün azalmasına neden olacaktır (Roth, 1994; Alpar, 2011). Üstelik kayıp veri içeren gözlemlerin veri setinden çıkartılabilmesi için kayıp verilerin tamamen rastlantısal olarak dağılıyor olması gerekmektedir. Kayıpların analize dâhil edilen başka değişkenlerle ilişkili olduğu durumlarda yapılacak olan silme işlemi önemli bir yanlılığa yol açabilir (Osborne, 2013;Schafer, 1999; Tabachnick ve Fidell, 1996). Bu bağlamda değerlendirildiğinde, kayıp veriler yerine yaklaşık değer atama yöntemleri, araştırmacıların hem zamandan ve emekten tasarruf edecekleri hem de topladıkları verileri koruyabilmelerini sağlayacak bir yol olarak ortaya çıkmaktadır.

Kayıp veriler yerine yaklaşık değer atamada kullanılan birçok yöntem bulunmaktadır. Bu yöntemler arasında bulunan Ortalama Atama (Mean Substitution), Yakın Noktalar Medyan Ataması (Median of Nearby Points), Doğrusal Değerleme (Lineer Interpolation) gibi yöntemler basit atamaya dayalı yöntemler olarak adlandırılmaktadır. Ayrıca bu yöntemler arasında daha gelişmiş yöntemler olarak nitelendirilen en çok olabilirlik kestirimine dayalı Beklenti Maksimizasyonu Algoritması (Expectation Maximization Algorithm) ve çoklu atamaya dayalı yöntemler olarak adlandırılan Eğilim Skorları Eşleştirmesi (Propensity Score Matching), Markov Zincirleri Monte Carlo (Marcov Chain Monte Carlo) gibi yöntemler de bulunmaktadır. Ayrıca son yıllarda Bayesci Veri Atama (Bayesian Imputation), Stokastik Regresyon Ataması (Stochastic Regression Imputation), Mahalanobis Uzaklığı Ataması (Mahalanobis Distance Imputation), K-Ortalama Kümeleme Ataması (K-Means Clustering Imputation) gibi farklı gelişmiş yöntemlerin de kullanıldığı görülmektedir. Kayıp veri miktarının az sayıda ve tamamen rastlantısal olarak dağıldığı (TROK) durumlarda basit atamaya dayalı yöntemler yeterli olabilir fakat aksi durumlarda gelişmiş yöntemlerin daha güvenilir sonuçlar vereceği belirtilmektedir (Osborne, 2013; Schafer, 1999). Buna karşın araştırmacılar, çeşitli nedenlerle (yaygın kullanılan istatistik programlarında yer almaması, uygulama ve yorumlamada zorluklarla karşılaşılacağı düşüncesi, alan bilgisi eksikliği vb.) gelişmiş yaklaşık değer atama yöntemlerini kullanmaktan kaçınmaktadırlar. Kayıp verilerle başa çıkma konusunda henüz hangi yöntemin hangi koşullar altında daha etkili sonuçlar verdiği konusunda yeterli bilimsel birikimin sağlanmamış olması da söz konusu durumun nedenlerinden biri olarak gösterilebilir. Alanyazında kayıp veriler yerine yaklaşık değer atama yöntemlerinin performanslarının karşılaştırıldığı çok sayıda araştırma yer almaktadır (Akbaş ve Tavşancıl, 2015; Allison, 2003; Altaş ve Kaspar, 2012; Çokluk ve Kayrı, 2011; Demir, 2013; Engels ve Diehr, 2003; Koçak ve Çokluk Bökeoğlu, 2017; Köse, 2014; Köse ve Öztemur, 2014; Musil, Warner, Yobas ve Jones, 2002; Saunders, Morrow-Howell, Spitznagel, Proctor ve Pescarino, 2006; Shrive, Stuart, Quan ve Ghali, 2006; Takahashi, 2017). Söz konusu araştırmaların bir kısmında yöntemler arası karşılaştırma yapılırken gelişmiş yaklaşık değer atama yöntemlerinin de ele alındığı fakat araştırmalara ağırlıklı olarak basit atamaya dayalı yöntemlerin konu edildiği belirlenmiştir. Alanyazına gelişmiş yaklaşık değer atama yöntemleriyle ilgili katkı sağlamak ve az bilinen bazı yöntemlerin kayıp verilerle başa çıkma konusundaki performanslarını tartışmaya açmak amacıyla bu araştırmada aşağıdaki yöntemler karşılaştırılmıştır.

Bayesci Veri Atama:

Bayes teoremi, rastsal bir sürece bağlı olarak ortaya çıkan rastgele bir X olayı ile diğer bir rastgele Y olayı için koşullu olasılıklar ve marjinal olasılıklar arasındaki ilişkiyi tanımlamaktadır. Bayesci Veri Atama yaklaşımı, gözlenebilen verilerden (önsel bilgi) olabilirlik fonksiyonu yoluyla sonsal olasılıkların tahmin edildiği ve elde edilen sonsal bilgilere dayalı olarak kayıp verilerin tamamlandığı bir yöntem olarak ifade edilebilir.

Eğilim Skorları Eşleştirmesi (ESE):

Lojistik regresyon yardımıyla tahmin edilen eğilim (propensity) skorları eşleştirilmiş setlerin ve tabakaların oluşturulması için bir düzenektir ve ortak değişken bilgilerinin özet ölçüsünü içermektedir. Eşit veya birbirine yakın eğilim skoruna sahip birimlerin, ortak değişkenlerine bağlı olarak, aynı (veya benzer) dağılıma sahip olma eğiliminde olacağı varsayılır (Kaspar, 2011). ESE ile yaklaşık değer atama mantığı, ortak değişkenlere sahip birimleri eşleştirmek ve kayıpsız birimlerden kayıplı birimlere veri atamak üzerine kuruludur.

Beklenti Maksimizasyonu Algoritması:

Yaklaşık değer atama süreci beklenti adımı (expectation step) ve maksimizasyon adımından (maximization step) oluşan iki aşamalı iteratif bir yöntemdir. Beklenti adımında kayıp veriler yerine regresyon tahminleriyle yaklaşık değerler atanır. Maksimizasyon adımında ise tamamlanmış olan veri üzerinden tahminler yenilenir. Bu süreç log-olabilirlik değerinin en yüksek noktaya çıktığı ve beklenen değerler arasındaki farkların önemsizleştiği noktaya kadar devam eder (Hedderley ve Wakeling, 1995).

Mahalanobis Uzaklığı Ataması:

Gözlemler arasındaki benzerlik ve benzemezlik korelasyon katsayıları ya da uzaklık ölçüleriyle incelenebilir. Korelasyon katsayıları iki gözlem arasındaki benzerliği ifade ederken; uzaklık ölçüleri ise benzemezliği ifade etmektedir. MUA yöntemine dayalı olarak kayıp veri içeren gözlemler kendisine en yakın (shortest mahalanobis distance) gözlemlerden oluşturulmuş bir kümeden çekilen verilerle tamamlanır.

Markov Zincirleri Monte Carlo Yöntemi:

Üç aşamalı bir çoklu yaklaşık değer atama yöntemidir. Birinci aşamada k adet veri seti simüle edilir. İkinci aşamada her bir veri seti üzerinden tam veri setinin dağılımına ilişkin tahminler yapılır ve son aşamada üretilen bilgiler birleştirilerek veri seti tamamlanır (Hasan, Ahmad, Osman, Sapri ve Othman, 2017).

Regresyon Ataması:

Bu yöntemde veri seti, kayıp veri içeren gözlemlerin yordanan; kayıp veri içermeyen gözlemlerin yordayıcı konumda olduğu regresyon denklemleriyle tamamlanmaktadır (Enders, 2010).

Little ve Rubin (1987)'e göre kayıp veriler yerine bilinçsizce atanan değerler var olan sorunları ortadan kaldırmadığı gibi ortaya çözümü daha güç olan yeni sorunlar çıkarmaktadır. Sözgelimi, kayıp verilerin TROK yapıda olmadığı durumlarda basit atamaya dayalı yöntemlerle yapılacak bir müdahale sonrası kayıp verilerden kaynaklı olarak meydana gelen yanlılık giderilememiş olabilir. Bu anlamda, kayıp verilerin gelişigüzel bir yöntemle tamamlanarak analizlere devam edilmesi gibi bir yaklaşımın hatalı sonuçların raporlanmasına ve sunulan bilimsel bilgilerin güvenirliği noktasında şüphelerin ortaya çıkmasına yol açacağı düşünülebilir. Yapılan tartışmalar doğrultusunda bu araştırmada, Tamamıyla Rastlantısal Olan Kayıp (TROK) ve Rastlantısal Olan Kayıp (ROK) yapılarında ve %15 ile %25 oranlarında oluşturulan kayıp veriler yerine farklı gelişmiş yöntemlerle yaklaşık değer atanması sonucu elde edilen veri setlerinin tam (gerçek) veri setiyle karşılaştırılarak incelenmesi amaçlanmıştır. Bu genel amaç doğrultusunda aşağıdaki sorulara yanıt aranmıştır.

1. Tamamıyla rastlantısal olan kayıp veri (TROK) yapısında %15 ve %25 kayıp veri içeren veri setine farklı yöntemlerle yaklaşık değer atanması sonucu,
    a. Elde edilen test puanlarının birbirleri ve tam veri seti test puanları ile korelasyonları ne düzeydedir?
    b. Elde edilen test puanları ile tam veri seti test puanları arasındaki farkların mutlak değer toplamları ve ortalamaları nasıl sıralanmaktadır?
    c. Elde edilen test puanlarının ortalamaları ile tam veri test puanları ortalaması arasındaki farklar istatistiksel olarak anlamlı mıdır?

2. Rastlantısal olan kayıp veri (ROK) yapısında erkekler bazında %25 kayıp veri içeren veri setine farklı yöntemlerle yaklaşık değer atanması sonucu,
    d. Elde edilen test puanlarının birbirleri ve tam veri seti test puanları ile korelasyonları ne düzeydedir?
    e. Elde edilen test puanları ile tam veri seti test puanları arasındaki farkların mutlak değer toplamları ve ortalamaları nasıl sıralanmaktadır?
    f. Elde edilen test puanlarının ortalamaları ile tam veri test puanları ortalaması arasındaki farklar istatistiksel olarak anlamlı mıdır?

Yöntem

Bu araştırma, bilgi üretmeye yönelik kuramsal bir çalışma özelliği taşıması bakımından temel araştırma niteliğindedir.

Araştırma Verileri

Araştırmada, bir test uygulamasına katılan bireyler üzerinden elde edilen veriler aracılığıyla kuramsal bir inceleme yapılması durumu söz konusudur dolayısıyla sonuçların herhangi bir bireyler evrenine genellenmesi amaçlanmamaktadır. Araştırma, PISA'ya (2012) Türkiye'den katılan 15 yaş grubundaki 4848 öğrenci arasından matematik öz-yeterliği ölçeğine katılan ve eksiksiz bir şekilde yanıtlayan 3129 öğrencinin puanlarından oluşan veri seti üzerinden yürütülmüştür. Çalışmada kullanılan PISA (2012) matematik öz-yeterliği anketi 1-4 arası puanlanan sekiz maddeden oluşmaktadır.

Verilerin Analizi

Araştırmada, öğrencilerin yanıt örüntüleri arasından tamamıyla rastlantısal olarak (TROK) dağılacak şekilde %15 ve %25 oranlarında veri silinerek iki farklı eksik veri seti elde edilmiş ayrıca rastlantısal olarak (ROK) dağılacak şekilde erkek öğrencilerin (cinsiyet değişkeni etkisi) yanıtları içerisinden %25 (bu kayıpların toplam veriye oranı yaklaşık %13 olarak belirlenmiştir) oranında veri silinerek toplamda üç farklı eksik veri seti elde edilmiştir. Kayıp veri dağılımlarının rastlantısallığının kontrol edilmesi amacıyla oluşturulan veri setleri Little'ın MCAR testine tabi tutulmuştur. Test sonuçlarından (TROK, $p > .05$; ROK, $p < .05$) istenen örüntülerin elde edildiğine dair kanıt sağlanmıştır. Araştırmada eşit oranda TROK ve ROK yapıda kayıp veri üzerinde çalışılmamıştır. Bunun nedeni, araştırmacıların yaklaşık değer atama yöntemlerinin bu iki örüntü (yapı) arasındaki işleyiş farklılıklarını incelemeyi değil, yöntemlerin farklı kayıp veri örüntüleri altında birbirleri ile karşılaştırmayı amaçlamalarındandır.

Söz konusu eksiltilmiş veri setlerinde kayıp veriler yerine farklı gelişmiş yaklaşık değer atama yöntemleriyle atamalar yapılmış (Beklenti Maksimizasyonu (BM), Regresyon Ataması (RA), Bayesci Veri Atama (BVA), Markov Zincirleri Monte Carlo(MZMC), Mahalanobis Uzaklığı Ataması (MUA), Eğilim Skorları Eşleştirmesi (ESE)) ve her bir yöntemle tamamlanan veri setlerinden hesaplanan ölçek puanları tam (gerçek) veri setinden hesaplanan ölçek puanları ile karşılaştırılmıştır. Hangi yöntemin tam veri seti değerlerine daha yakın değerler kestirdiğinin incelenebilmesi için ölçek puanları arasındaki farkların mutlak değer toplamları sıfıra yakınlık

dereceleri bakımından karşılaştırılmıştır. Kayıp veriler yerine yaklaşık değer atama yöntemleriyle tamamlanan veri setlerinden elde edilen ölçek puanlarının ortalaması ile tam veri setinden hesaplanan ölçek puanları ortalaması arasındaki farkın istatistiksel olarak manidar olup olmadığının incelenebilmesi amacıyla t-testi yapılmıştır. Araştırma verilerinin düzenlenmesi, analizi ve kayıp veriler yerine yaklaşık değer atanması sürecinde SPSS V22, LISREL 8.80, AMOS V22 ve SOLAS V5 paket programları kullanılmıştır.

Bulgular

Tamamıyla rastlantısal olan kayıp veri (TROK) yapısında ve %15 kayıp veri içeren veri setine BVA, BM, MUA, MZMC, ESE ve RA yöntemleri ile veri ataması yapılmıştır. Yaklaşık değer atamaları sonucu elde edilen ölçek puanları ve tam veri setinden hesaplanan gerçek ölçek puanlarına ait betimsel istatistikler Tablo 1'de sunulmuştur.

Tablo 1
TROK %15 Kayıp Veri Durumunda Betimsel İstatistiklere İlişkin Bulgular

	Ölçek Puanları Toplamı	$\bar{X}$	S	Minimum Puan	Maksimum Puan	Ranj	Çarpıklık	Basıklık
Tam V.	49289.00	15.75	4.61	8.00	32.00	24.00	0.42	0.17
BVA	49271.87	15.75	4.61	6.52	33.11	26.59	0.41	0.12
BM	49225.60	15.73	4.54	7.91	32.07	24.16	0.42	0.14
MUA	49255.19	15.74	4.60	5.62	32.82	27.20	0.38	0.09
MZMC	49197.00	15.72	4.53	8.00	32.00	24.00	0.41	0.16
ESE	49256.23	15.74	4.54	5.61	32.85	27.24	0.38	0.13
RA	49285.52	15.75	4.63	5.19	33.69	28.50	0.40	0.12

Tablo 1 incelendiğinde, tam veri seti ölçek puanlarının 8.00 ile 32.00 arasında değiştiği, puanların aritmetik ortalamasının 15.75, standart sapmasının ise 4.61 olarak hesaplandığı görülmektedir. Ayrıca tam veri seti ölçek puanları dağılımının çarpıklık değeri 0.42, basıklık değeri ise 0.17 değerini almıştır. Eksiltilmiş veri setine farklı yöntemlerle yapılan atamalar sonucu elde edilen ölçek puanlarının betimsel istatistiklerinin tam veri seti ölçek puanları istatistiklerine yakın değerler aldığı belirlenmiştir. Bununla birlikte aritmetik ortalama ve standart sapma istatistikleri öncelikli olarak dikkate alındığında BVA sonucu elde edilen istatistiklerin tam veri setine en yakın değerleri aldığı sonucuna ulaşılmıştır. En düşük ve en yüksek puanlar dikkate alındığında MZMC sonucu elde edilen puanların tam veri seti ile aynı olduğu, çarpıklık ve basıklık değerlerinin ise tam veri değerlerine çok yakın olduğu belirlenmiştir.

Her bir yöntem için veri ataması sonucunda elde edilen ölçek puanları ile tam veri seti ölçek puanları arasındaki ilişki düzeyleri Pearson Momentler Çarpımı Korelasyon Katsayısı Tekniği ile elde edilmiş ve bulgular incelenmiştir. Elde edilen bulgular Tablo 2'de sunulmuştur.

Tablo 2
TROK %15 Kayıp Veri Durumunda Puanlar Arasındaki Korelasyonlara İlişkin Bulgular

	Tam Veri	BVA	BM	MUA	MZMC	ESE	RA
Tam Veri	1.00						
BVA	0.97	1.00					
BM	0.98	0.98	1.00				
MUA	0.97	0.97	0.98	1.00			
MZMC	0.98	0.98	0.99	0.98	1.00		
ESE	0.96	0.96	0.98	0.99	0.98	1.00	
RA	0.97	0.97	0.99	0.97	0.98	0.97	1.00

Tablo 2'de verilen değerler incelendiğinde tam veri seti ile yaklaşık değer atama yöntemleri ile tamamlanmış veri setlerinden hesaplanan ölçek puanları arasındaki ilişki düzeylerinin pozitif yönde ve yüksek (0.96 ile 0.98 değerleri arasında) düzeyde olduğu belirlenmiştir. Benzer şekilde yaklaşık değer atama yöntemleri sonucu elde edilen ölçek puanlarının birbirleri ile ilişkilerinin de pozitif yönde ve yüksek (0.97 ile 0.99 değerleri arasında) düzeyde olduğu belirlenmiştir.

Yöntemlerin tam veri setine ne derece yakın değerler kestirdiğinin incelenebilmesi için tam veri seti ile yaklaşık değer atanmış veri setlerinden hesaplanan ölçek puanları arasındaki farkların mutlak değer toplamları ve ortalamaları karşılaştırılmıştır. Elde edilen bulgular en küçük farkın elde edildiği yöntemden en büyüğe doğru sıralanarak Tablo 3'te sunulmuştur.

Tablo 3
TROK %15 Kayıp Veri Durumunda Puanlar Arasındaki Farklara İlişkin Bulgular

	Farkların Mutlak Değerlerinin Toplamı	$\bar{X}$*
BM	1680.44	0.54
MZMC	1694.00	0.54
RA	2297.47	0.73
MUA	2365.91	0.76
BVA	2370.50	0.76
ESE	2455.73	0.78

**ortalamalar virgülden sonraki ikinci basamağa yuvarlanmıştır.*

Tablo 3 incelendiğinde, tam veri seti ölçek puanlarından en az farkla kestirim yapan yöntemlerin BM ve MZMC olduğu her iki yöntemde de farkların aritmetik ortalamasının 0.54 olarak hesaplandığı görülmektedir. Bu bulgu, söz konusu yöntemlerin TROK yapısında %15 oranında kayıp veri olması durumunda her bir öğrenci için ortalama 0.54 puanlık bir sapma (32 puanlık bir ölçekte) ile gerçeğe yakın puanlar kestirdiğini göstermektedir. Tam verideki ölçek puanları ile en fazla farkın meydana geldiği atama yöntemi ise ESE yöntemi olarak belirlenmiştir.

Son olarak, yaklaşık değer atama yöntemleri ile elde edilen ölçek puanlarının ortalamaları ile tam veri seti ölçek puanlarının ortalamaları arasındaki farkların anlamlılığı t-testleri ile incelenmiş ve ortalamalar arasında istatistiksel olarak anlamlı bir fark olmadığı sonucuna ulaşılmıştır.

Tamamıyla rastlantısal olan kayıp veri (TROK) yapısında ve %25 kayıp veri içeren veri setine BVA, BM, MUA, MZMC, ESE ve RA yöntemleri ile veri ataması yapılmıştır. Yaklaşık değer atamaları sonucu elde edilen ölçek puanları ve tam veri seti ölçek puanlarına ait betimsel istatistikler Tablo 4'te verilmiştir.

Tablo 4
TROK %25 Kayıp Veri Durumunda Betimsel İstatistiklere İlişkin Bulgular

	Ölçek Puanları Toplamı	$\bar{X}$	S	Minimum Puan	Maksimum Puan	Ranj	Çarpıklık	Basıklık
Tam Veri	49289.00	15.75	4.61	8.00	32.00	24.00	0.42	0.17
BVA	49191.81	15.72	4.59	3.18	33.40	30.22	0.37	0.03
BM	48188.58	15.40	4.45	5.51	32.09	26.58	0.42	0.06
MUA	49262.85	15.74	4.56	3.89	33.71	29.83	0.37	0.13
MZMC	49188.00	15.72	4.54	8.00	32.00	24.00	0.39	0.08
ESE	49279.01	15.75	4.48	3.89	33.71	29.83	0.38	0.19
RA	49233.68	15.73	4.57	5.72	33.78	28.07	0.39	0.02

Tablo 4 incelendiğinde, TROK yapıda %25 oranında eksiltilmiş veri setine farklı yöntemlerle yapılan atamalar sonucu elde edilen ölçek puanlarının betimsel istatistiklerinin tam veri seti ölçek puanları istatistiklerine yakın değerler aldığı belirlenmiştir.

Eksiltilmiş veri setlerine yaklaşık değerler atanması sonucu elde edilen ölçek puanları ile tam veri seti ölçek puanları arasında korelasyonlar incelenmiş ve elde edilen bulgular Tablo 5'te sunulmuştur.

Tablo 5

TROK %25 Kayıp Veri Durumunda Puanlar Arası Korelasyonlara İlişkin Bulgular

	Tam Veri	BVA	BM	MUA	MZMC	ESE	RA
Tam Veri	1.00						
BVA	.94	1.00					
BM	.96	.96	1.00				
MUA	.94	.94	.96	1.00			
MZMC	.97	.96	.98	.96	1.00		
ESE	.94	.93	.96	.99	.96	1.00	
RA	.95	.95	.96	.95	.97	.94	1.00

Tablo 5 incelendiğinde, tam veri seti ile yaklaşık değer atanmış veri setlerinden elde edilen ölçek puanları arasındaki korelasyonların pozitif yönde ve yüksek düzeyde olduğu (0.94 ile 0.97 değerleri arasında) belirlenmiş ve tam veri seti ile en yüksek korelasyona sahip yöntemin 0.97 değeri ile MZMC olduğu görülmüştür. Yaklaşık değer atama yöntemleri ile elde edilen ölçek puanlarının da birbirleri ile pozitif yüksek ilişkili olduğu (0.94 ile 0.99 değerleri arasında) belirlenmiştir.

TROK %25 kayıp veri durumunda yaklaşık değer atama yöntemlerinin tam veri setine ne derece yakın değerler kestirdiğinin incelenebilmesi için tam veri seti ile yaklaşık değer atanmış veri setlerinden hesaplanan ölçek puanları arasındaki farkların mutlak değer toplamları ve ortalamaları karşılaştırılmıştır. Elde edilen bulgular en küçük farkın elde edildiği yöntemden en büyüğe doğru sıralanarak Tablo 6.'da sunulmuştur.

Tablo 6
TROK %25 Kayıp Veri Durumunda Puanlar Arası Farklara İlişkin Bulgular

	Farkların Mutlak Değerlerinin Toplamı	$\bar{X}$*
MZMC	2465.00	0.79
BM	2729.83	0.87
RA	3286.76	1.05
BVA	3525.25	1.13
MUA	3568.39	1.14
ESE	3619.22	1.16

**ortalamalar virgülden sonraki ikinci basamağa yuvarlanmıştır.*

Tablo 6. incelendiğinde, toplam puanlar bakımından tam veri seti ile en az farkın MZMC yöntemiyle yapılan atama sonucundan elde edildiği belirlenmiştir. Söz konusu yöntem birey bazında ortalama 0.79 puanlık bir sapmayla gerçek sonuçlara en yakın değerlerin kestirilmesini sağlamıştır. Bulgulara göre tam veriden en uzak sonuçlar ESE yöntemi ile yapılan atamalardan elde edilmiştir.

Son olarak, yaklaşık değer atama yöntemleri ile elde edilen ölçek puanlarının ortalamaları ile tam veri seti ölçek puanlarının ortalamaları arasındaki farkların anlamlılığı t-testleri ile incelenmiş ve ortalamalar arasında istatistiksel olarak anlamlı bir fark olmadığı sonucuna ulaşılmıştır.

Tam veri setinden erkekler bazında %25 oranında eksiltilmiş rastlantısal olan kayıp veriler (ROK) yerine BVA, BM, MUA, MZMC, ESE ve RA yöntemleri ile yaklaşık değerler atanması sonucu elde edilen veri setlerinden hesaplanan ölçek puanları ile tam veri seti ölçek puanlarına ilişkin betimsel istatistikler Tablo 7'de sunulmuştur.

Tablo 7
Erkekler Bazında ROK%25 Kayıp Veri Durumunda Betimsel İstatistiklere İlişkin Bulgular

	Toplam Puan	$\bar{X}$	S	Minimum Puan	Maksimum Puan	Ranj	Çarpıklık	Basıklık
Tam V.	49289.00	15.75	4.61	8.00	32.00	24.00	0.42	0.17
BVA	49104.28	15.69	4.61	5.79	33.06	27.26	0.42	0.16
BM	49186.23	15.72	4.56	7.89	32.00	24.11	0.42	0.17
MUA	49003.70	15.66	4.64	5.30	33.93	28.63	0.43	0.18
MZMC	49160.00	15.71	4.57	8.00	32.00	24.00	0.41	0.18
ESE	49200.00	15.72	4.57	5.77	32.01	26.24	0.41	0.16
RA	49187.86	15.72	4.63	6.92	32.54	25.62	0.43	0.16

Tablo 7 incelendiğinde, ROK yapıda %25 oranında eksiltilmiş veri setine farklı yöntemlerle yapılan atamalar sonucu elde edilen ölçek puanlarının betimsel istatistiklerinin tam veri seti ölçek puanları istatistiklerine yakın değerler aldığı belirlenmiştir. Eksiltilmiş veri setlerine yaklaşık değerler atanması sonucu elde edilen ölçek puanları ile tam veri seti ölçek puanları arasındaki korelasyonlar incelenmiş ve elde edilen bulgular Tablo 8'de sunulmuştur.

Tablo 8
Erkekler Bazında ROK%25 Kayıp Veri Durumunda Atamasında Korelasyonel Değerler

	Tam Veri	BVA	BM	MUA	MZMC	ESE	RA
Tam Veri	1.00						
BVA	.97	1.00					
BM	.99	.98	1.00				

MUA	.97	.97	.98	1.00			
MZMC	.98	.98	.99	.98	1.00		
ESE	.97	.97	.98	.97	.98	1.00	
RA	.97	.97	.99	.97	.99	.97	1.00

Tablo 8'deverilen değerler incelendiğinde, tam veri seti ile yaklaşık değer atama yöntemleri ile tamamlanan veri setlerinden elde edilen ölçek puanları arasındaki ilişki düzeylerinin pozitif yönde ve yüksek (0.97 ile 0.99 değerleri arasında) olduğu belirlenmiştir. Benzer şekilde farklı yöntemlerle tamamlanan veri setlerinden elde edilen ölçek puanlarının birbirleri ile ilişkilerinin de pozitif yönde ve yüksek (0.97 ile 0.99 değerleri arasında) düzeyde olduğu belirlenmiştir.

Erkekler bazında ROK %25 kayıp veri durumunda yaklaşık değer atama yöntemlerinin tam veri setine ne derece yakın değerler kestirdiğinin incelenebilmesi için tam veri seti ile yaklaşık değer atanmış veri setlerinden hesaplanan ölçek puanları arasındaki farkların mutlak değer toplamları ve ortalamaları karşılaştırılmıştır. Elde edilen bulgular en küçük farkın elde edildiği yöntemden en büyüğe doğru sıralanarak Tablo 9'da sunulmuştur.

Tablo 9
Erkekler Bazında ROK %25 Kayıp Veri Durumunda Puanlar Arası Farklara İlişkin Bulgular

	Farkların Mutlak Değerlerinin Toplamı	$\bar{X}$*
BM	1199.68	0.38
MZMC	1251.00	0.40
RA	1663.49	0.53
ESE	1760.89	0.56
BVA	1762.61	0.56
MUA	1768.79	0.57

**ortalamalar virgülden sonraki ikinci basamağa yuvarlanmıştır.*

Tablo 9 incelendiğinde, tam veri ölçek puanlarından en az farkla kestirim yapan yöntemin BM olduğu belirlenmiştir. Söz konusu yöntem birey bazında ortalama 0.38 puanlık bir sapmayla gerçek sonuçlara en yakın değerlerin kestirilmesini sağlamıştır. Birey bazında ortalama 0.40 puanlık bir sapmayla gerçek puanlara en yakın ikinci kestirim MZMC yöntemi ile elde edilmiştir. Tabloya göre tam veriden en uzak sonuçlar MUA yöntemi ile yapılan atamalardan elde edilmiştir.

Son olarak, farklı yaklaşık değer atama yöntemleri ile tamamlanan veri setlerinden elde edilen ölçek puanlarının ortalamaları ile tam veri seti ölçek puanlarının ortalaması arasındaki

farkların anlamlılığı t-testleri ile incelenmiş ve ortalamalar arasında istatistiksel olarak anlamlı bir fark olmadığı sonucuna ulaşılmıştır.

Tartışma, Sonuç ve Öneriler

Araştırma kapsamında ele alınan, kayıp veriler yerine yaklaşık değer atama yöntemlerinden BM, BVA, ESE, MUA, MZMC ve RA yöntemleri, TROK %15 ve %25, ROK erkekler bazında %25 kayıp veri koşulları altında karşılaştırılmıştır.

Belirlenen koşulların tümünde,

- Söz konusu yöntemlerle tamamlanan veri setlerinden elde edilen ölçek puanları ile tam veri ölçek puanları arasındaki korelasyon değerlerinin yüksek düzeyde ve pozitif yönlü olduğu görülmüştür.
- Tamamlanan veri setlerinden elde edilen ölçek puanlarının ortalamaları ile tam veri seti ölçek puanlarının ortalaması arasında istatistiksel olarak anlamlı bir fark olmadığı belirlenmiştir.

Elde edilen bulgulardan hareketle, araştırma kapsamında incelenen tüm yöntemlerin tam veriye yakın değerler ürettiği ve kayıp verilerle başa çıkma noktasında başarılı sonuçlar ortaya koyduğu öne sürülebilir. Bu sonuç, kayıp veri sorunu ile başa çıkma noktasında yöntemler arasında çok büyük farklılıklar olmadığı yönünde Nartgün'ün (2015) ulaştığı sonuçlarla benzerlik göstermektedir. Bununla birlikte, tam veri seti ile tamamlanmış veri setlerinden hesaplanan ölçek puanları arası farkların mutlak değer toplamları ve ortalamaları göz önünde bulundurulduğunda her üç koşul altında da en iyi çalışan yaklaşık değer atama yöntemlerinin MZMC (Markov Zincirleri Monte Carlo) ve BM (Beklenti Maksimizasyonu) olduğu sonucuna ulaşılmıştır. Kayıp verilerle başa çıkma söz konusu olduğunda kullanılabilecek çok sayıda yöntem mevcuttur. Bu yöntemlerden hangisinin daha iyi çalıştığı noktasında yapılacak bir karşılaştırma için pek çok koşul ve ölçüt belirlenebilir. Bu nedenle kayıp veriler üzerine yapılan araştırmaların sonuçları arasında bağlantılar kurmak zorlayıcı olduğu gibi bazı durumlarda yanıltıcı da olabilir. Sözgelimi, bu araştırmada karşılaştırılan yöntemlerin doğrudan ham puanlar üzerinden karşılaştırıldığı benzer bir çalışmaya rastlanılmamıştır. Buna karşın, sonuçlar arasında tam bir paralellikten söz edilemese de BM ve MZMC yöntemlerinin diğer yöntemlerle farklı ölçütler dikkate alınarak karşılaştırıldığı ve yüksek performans gösterdiği için önerildiği araştırmaların (Akbaş ve Tavşancıl, 2015; Aljuaid ve Sasi, 2016; Demir, 2013; Enders, 2004; Lin, 2008; Mao ve Lin, 2008; Musil, Warner, Yobas ve Jones; 2002; Ni ve Leonard, 2005; Şahin Kürşad ve Nartgün, 2015) sonuçlarıyla bu araştırmanın

sonuçlarının birbirini desteklediği söylenebilir. Bu bakımdan, araştırmacılara TROK ve ROK yapısındaki kayıp veriler yerine BM veya MZMC yöntemleri ile yaklaşık değer atanması yapmaları önerilmektedir.

Kaynakça

Akbaş, U. ve Tavşancıl, E. (2015). Farklı örneklem büyüklüklerinde ve kayıp veri örüntülerinde ölçeklerin psikometrik özelliklerinin kayıp veri baş etme teknikleri ile incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi,* 6(1), 38-57.

Aljuaid, T. ve Sasi, S. (2016). Proper imputation techniques for missing values in data sets. *2016 International Conference on Data Science and Engineering (ICDSE),* Cochin, 23-25 Ağustos 2016, s. 1-5.

Allison, P.D. (2001).*Missing Data.* CA: Sage University Paper.

Allison. P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology.* 4(1), 545-557.

Alpar. R. (2011). *Çok değişkenli istatistiksel yöntemler.* Ankara: Detay Yayıncılık.

Altaş, D. ve Kaspar, E.Ç. (2012). Propensity skor ve hot-deck veri atama yöntemlerinin karşılaştırılması. *Trakya Üniversitesi İktisadi ve İdari Bilimler Fakültesi E-Dergi, 1(1), 26-41.*

Byrne, B. (2000). *Structural equation modelingwith AMOS: Basic concepts, applications, andprogramming*. Mahwah, NJ: Lawrence Erlbaum.

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

Çokluk, Ö. ve Kayrı, 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. ve Gelbal, S. (2015). Kayıp veriler yerine yaklaşık değer atamada kullanılan farklı yöntemlerin model veri uyumu üzerine etkisi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi,* 35, 87-111.

Demir, E. (2013). Kayıp verilerin varlığında çoktan seçmeli testlerde madde ve test parametrelerinin kestirilmesi: SBS örneği. *Eğitim Bilimleri Araştırmaları Dergisi,* 3(2), 47-68.

Enders, C. K. (2004). The impact of missing data on sample reliability estimates: implications for reliability reporting practices. *Educational and Psychological Measurement,* 64(3), 419-436.

Enders, C. K. (2010). *Applied missing data analysis.* NY: The Guilford Press.

Engels, J. M. ve Diehr, P. (2003). Imputation of missing longitudinal data: A comparison of methods. *Journal of Clinical Epidemiology, 56(1), 968-976.*

Hasan, H., Ahmad, S., Osman, B. M., Sapri, S., ve 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.

Hedderley, D. ve Wakeling, I. (1995). A comparison of imputation techniques for internal preferencemapping using Monte Carlo simulation. *Food Quality and Preference, 6, 281-297.*

Kaspar, E.Ç. (2011). *Kayıp veriler ve kayıp veriler için bir çoklu veri atama yöntemi: Propensity skor* (yayımlanmamış doktora tezi). Marmara Üniversitesi, İstanbul.

Koçak, D. ve Çokluk Bökeoğlu, Ö. (2017). Kayıp veriyle baş etme yöntemlerinin model veri uyumu ve madde model uyumuna etkisi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 8(2), 200-223.*

Köse, A. (2014). The effect of missing data handling methods on goodness of fit indices in confirmatory factor analysis. *Educational Research and Reviews, 9(8), 208-215.*

Köse, İ. A. ve Öztemur, B. (2014). Kayıp veri ele alma yöntemlerinin t-testi ve ANOVA parametreleri üzerine etkisinin incelenmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 14(1), 400-412.*

Lin, T.H. (2008). A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. *Quality & quantity*. 2010, 44 (2): 277-287.

Little. R. ve Rubin. D. (1987). *Statistical analysis with missing data.* New York: Wiley.

Mao, Q. Ve Li, X. (2005). Markov Chain Monte Carlo Method of multiple imputation for longitudinal data with missing values in the survey of maternal and children health. *Sichuan da xue xue bao. Yi xue ban (Journal of Sichuan University)* Medical science edition, 36(3), 422–425.

Musil, C,M., Warner, C, B., Yobas, P,K. ve Jones, L,S. (2002). A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research, 24(7), 815-829.*

Nartgün, Z. (2015). Kayıp veri sorununun çözümünde kullanılan farklı yöntemlerin farklı kayıp veri koşulları altında ölçeklerin psikometrik nitelikleri ve ölçme sonuçları

bağlamında karşılaştırılması. *International Online Journal of Education Sciences,* 7(4), 252-265.

Ni, D. ve Leonard, JD. (2005) Markov chain monte carlo multiple ımputation for ıncomplete ıts data using bayesian networks. *84. Annual Meeting of the Transportation Research Board*, 12 Temmuz, 2005.

Osborne. J. W. (2013). *Best practices in data cleaning.* California: Sage Publication. Inc.

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

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

Saunders, J. A., Morrow-Howell, N., Spitznagel, P. D., Proctor, E.K. ve Pescarino, R. (2006). Imputing missing data: A comparison of methods for social work researchers. *Social Work Research, 30(1),*

Şahin Kürşad, M., ve 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.

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

Shrive, F. M., Stuart, H., Quan, H. ve Ghali, W. A. (2006). Dealing with missing data in a multi-question depression scale: A comparison of imputation methods. *BMC Medical Research Methodology, 57(6), 110.*

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

Takahashi, M. (2017). Statistical inference in missing data by MCMC and non-MCMC multiple imputation algorithms: Assessing the effects of between-imputation iterations. *Data Science Journal, 37(16), 1-17.*

Extended Abstract

Missing data imputation methods instead of missing data emerges as a way for researchers to save both time and labor. Although there are many techniques used for imputation of missing values, it is difficult to decide which technique should be preferred in which cases. It is aimed to compare the data sets which have been completed with approximate value to the complete (real) dataset and thus determine which method works better under the determined conditions. With this aim, we used different methods (Expectation Maximization (EM), Regression Imputation (RIM), Bayesian Imputation (BIM), Markov Chains Monte Carlo (MCMC), Mahalanobis Distance Imputation (MDI), Propensity Score Matching (PSM) instead of the

missing data that have different rate (25% and 15%) and structures (MCAR and MAR) in this study. The group of research data, PISA (2012) was conducted on the data set consisting of the scores of 3129 learners participating in the mathematics self-efficacy scale and fully answered among the 4848 students in the 15-year-old group participating from Turkey. The scale consists of 8 items scored between 1-4. In the study, two different missing data sets were obtained by deleting 15% and 25% of the data from the response patterns of the students completely at random (MCAR) and also three different missing data sets were obtained by deleting 25% (The total loss ratio of these losses is determined about 13%) of the replies of male students (gender change effect) in a missing at random (MAR). The data sets that created to control the randomness of the missing data distributions were subjected to Little's MCAR Test. The evidence was obtained that the desired patterns were obtained from the test results (MCAR, $p>.05$; MAR, $p<.05$). Missing values were not studied with the same way in MCAR and MAR structure in the study. The reason for this is that missing value imputation methods in the research are not intended to examine the functional differences between these two patterns (structures) but to compare the methods with each other under different loss data patterns.

The missing data (MCAR) structure, which is completely random, and the reduced data set which containing 15% lost data are completed with BIM, EM, MDI, MCMC, PSM and RIM methods,

- The descriptive statistics obtained from the whole of the completed data sets were found to be close to the descriptive statistics of the complete (real) data set.
- The relation levels between scale scores obtained from data completed and complete (real) data set scale scores were examined and it was determined that the relationship levels among all scores were positive and high (between 0.96 and 0.98) for each method.
- The significance of the differences between the mean of the scale scores obtained by the completed data sets and the mean of the scale scores of the complete data set was examined with t-tests and it was concluded that there was no statistically significant difference between the averages.
- Absolute value sums and mean values of the differences between the scale scores calculated from the real data set and the approximate value data sets were compared and

it was determined that EM (1) and MCMC (2) were the least predictive methods from the complete data set scores.

The missing data (MCAR) structure, which is completely random and a reduced data set containing 25% lost data were Completed with BIM, EM, MDI, MCMC, PSM and RIM methods,

• The descriptive statistics obtained from the whole of the completed data sets were found to be close to the descriptive statistics of the complete (real) data set.

• The relation levels between scale scores obtained from data completed and complete data set scale scores were examined and it was determined that the relationship levels among all scores were positive and high (between 0.94 and 0.97) for each method.

• The significance of the differences between the mean of the scale scores obtained by the completed data sets and the mean of the scale scores of the comlete data set was examined with t-tests and it was concluded that there was no statistically significant difference between the averages.

• It was determined that the differences between the scale scores calculated from the complete data set and completed data sets were compared with the absolute value sums and the mean values were MCMC (1) and EM (2), which were the least predicted from the complete data set scores.

The data set, which contains missing data in reduced MAR structure based on 25% in males from complete data set, was completed with BIM, EM, MDI, MCMC, PSM and RI methods,

• The descriptive statistics obtained from the whole of the completed data sets were found to be close to the descriptive statistics of the complete (real) data set.

• The relation levels between scale scores obtained from completed and complete data set scale scores were examined and it was determined that the relationship levels among all scores were positive and high (between 0.97 and 0.99) for each method.

• The significance of the differences between the mean of the scale scores obtained by completed data sets and the mean of the scale scores of the complete data set was examined with t-tests and it was concluded that there was no statistically significant difference between the averages.

• Absolute value sums and mean values of the differences between the scale scores calculated from the complete data set and the completed data sets were compared and it

was determined that EM (1) and MCMC (2) were the least predictive methods from the complete data set scores.

According to findings, it can be asserted that all the methods investigated within the scope of the study produced close values to real values and they have successful results were coped with missing data. Nevertheless, when the absolute value totals and the average values of the differences between the scale scores calculated from the complete data set and the completed data sets are considered, it is concluded that MCMC and EM are the best missing data imputation methods under all three conditions.