Eğitim ve Sosyal Bilimlerde Yeni Araştırmalar

Editörler
Prof. Dr. Bülent OKUR
Dr. Gül Banu DAYANÇ KIYAT
Dr. Ebru ATEŞOK

Ankara 2019

Bu kitabın tüm hakları yazarına ve yayıncısına aittir.

Eğitim ve Sosyal Bilimlerde Yeni Araştırmalar

Editörler
Prof. Dr. Bülent OKUR
Dr. Gül Banu DAYANÇ KIYAT
Dr. Ebru ATEŞOK

ISBN: 978-605-7634-08-5

Genel Yayın Yönetmeni
Cuma AĞCA

Sayfa Düzeni / Kapak
Biçer YILDIRIM

Baskı & Cilt
Berikan Matbaacılık / Gersan-ANKARA
Matbaa Sertifika No: 13642

YAYINEVİ
BERİKAN YAYINEVİ
Kültür Mah. Kızılırmak Cad. Gonca Apt. No: 61/6
Çankaya-Kızılay/ANKARA
Tel: (0312) 232 62 18 Fax: (0312) 232 14 99

ÖNSÖZ

Merhaba;

Ülkemizde eğitim ve sosyal bilimler alanında yapılacak olan bilimsel çalışmalara katkı sunmak maksadıyla hazırlamış olduğumuz, "Eğitim ve Sosyal Bilimlerde Yeni Araştırmalar" başlıklı eserimizi siz değerli okurlarımızla buluşturmanın haklı gururunu ve sevincini yaşıyoruz.

Disiplinler arası bilimsel çalışmalara odaklandığımız eserimizin ilk bölümde; Sosyal, Beşeri ve İdari Bilimler alanları konu edilmiştir. Alt bölümleri ise örgüt içi iletişim, örgüt kültürü, finansal yönetim, kriz yönetimi, halkla ilişkiler, marka kimliği ve psikolojik sermaye gibi güncel birçok kavram incelenmiştir. Kitabımızın ikinci bölümü olan Eğitim Bilimleri başlığında ise öğrenme ve başarı kavramları incelenmiştir. Eserimizin son bölümü olan üçüncü bölümünde ise Güzel Sanatlar başlığında mimari tasarımdan, islam mimarisine, yavaş şehir kavramından, kültür ve sanat köylerine kadar birçok güncel kavram dair araştırmalar ve tespitler bulunmaktadır.

Kitabımızda bilimsel normlara ve bilim etiğine uygun, nitelikli ve özgün çalışmalar titizlikle değerlendirerek, yapılan çalışmaların daha geniş kitlelere duyurulması ve ulusal yayıncılığa katkı sağlanması hedeflenmiştir. Eserimizi tamamlarken gerçekleştirdiğimiz çalışma sürecinde, gerekli destekleri esirgemeyen herkese minnettarız.

Eserimizin hazırlık aşamasında görev alan bilim kurulumuza ve editörlerimize de teşekkürü borç biliriz. Son olarak ise özellikle yazarlarımıza, eserimizin ortaya çıkışında ki katkılarından dolayı şükranlarımızı ifade ederiz.

Prof.Dr. Bülent OKUR
Dr. Öğr. Üyesi Ebru ATEŞOK
Dr. Öğr. Üyesi Gül Banu DAYANÇ KIYAT
Dr. Esra KİRİK
15 MART 2019

Eksik Veri Atama Yöntemlerinin Bir Eğitim Veri Seti Üzerinde Karşılaştırılmalı İncelenmesi

Arş. Gör. Tugay KARADAĞ[1]
Yıldız Teknik Üniversitesi

Dr. Öğr. Üyesi Elif TUNA[2]
Yıldız Teknik Üniversitesi

Özet

Bu çalışmada yaygın kullanılan eksik veri atama yöntemlerinin performansları karşılaştırılmıştır. Eksik veri atama yöntemlerinin performanslarını karşılaştırmak için bir eğitim araştırması olan TIMSS veri setinden yararlanılmıştır. Burada eksiksiz bir TIMSS veri setinden, öğrencilerin matematik alanında kazandıkları bilgi ve becerilerin değerlendirilmesi sonucu yayınlanan ve eksik veri bulunmayan 5 değişkenli ve 3848 gözleme sahip bir veri seti üzerinde çalışılmıştır. Eksiksiz bu veri setinden 5 farklı eksik veri yüzdesi aralığına sahip veri setleri oluşturulmuştur. Bu veri setlerinin her bir değişkeni tamamen rastlantısal olan eksik veri (TROK) yapısındadır. TROK yapıda olan bu veri setlerinin değişkenleri sırasıyla %0-5, %5-10, %10-15, %15-20, %20-25 arasında rastgele yüzdeye sahip eksik verili bir yapıya dönüştürülmüştür. Çalışmada kullanılan eksik veri atama yöntemleri Beklenti Maksimizasyonu (EM), Markov Zinciri Monte Carlo (MCMC), Regresyon Atama Yöntemi, Yakın Noktalar Ortalama Ataması, Yakın Noktalar Medyan Ataması, Doğrusal İnterpolasyon Ataması, Seri Ortalamaları Ataması ve Bir Noktanın Doğrusal Eğimi Ataması şeklindedir. 5 farklı yüzde aralığında eksik verili hale getirilmiş veri setlerine, bahsedilen 8 atama yöntemi uygulanmıştır. Ardından eksik veri bulunmayan orijinal veri setindeki değerler Ortalama Hata Kareler (MSE), Ortalama Mutlak Hata (MAE) ve orijinal veri setiyle atanmış veri setlerinin korelasyon katsayıları olmak üzere 3 farklı sınama yöntemiyle karşılaştırılmıştır. Atamalar sonucunda, eksik veri yüzdesi arttıkça atama yapılan veri seti ile orijinal veri setinin korelasyon katsayısının düzenli bir şekilde azaldığı tespit edilmiştir. Fakat MSE ve MAE değerleri için aynı durum söz konusu değildir. %10-15 arası eksik veri bulunan veri setinde yöntemler en iyi MSE ve MAE değerlerini verirken, %0-5 arası eksik veri bulunan veri setinde yöntemler en kötü MSE ve MAE değerlerini vermişlerdir. 3 farklı sınama çeşidine göre en iyi sonucu EM, en iyi ikinci sonucu MCMC yöntemi ve en iyi üçüncü sonucu Regresyon Atama yöntemi vermiştir.

Anahtar Kelimeler: Beklenti Maksimizasyonu, Eksik Veri Atama Yöntemleri, Regresyon, Markov Zinciri Monte Carlo

Comparative Analysis of Missing Data Imputation Methods on a Training Data Set

Abstract

In this study, performances of widely used missing data imputation methods were compared. TIMSS data set, which is an educational research, was used to compare the performance of missing data imputation methods. Here, a TIMSS data set with 5 variables and 3848 observations in which there is no missing value was studied. Data sets which has 5 different missing percentage interval were generated from this complete data set. All variables are Missing Completely at Random (MCAR). All data sets are converted to a missing data sets with a random percentage with %0-5, %5-10, %10-15, %15-20, %20-25 missing value, respectively. The methods used in this study are Expectation Maximization, Markov Chain Monte Carlo, Regression Imputation, Mean of Nearby Points, Median of Nearby Points, Linear Interpolation, Series Mean and Linear Trend at a Point. The methods were applied to the missing data sets and after the methods were compared with 3 test methods which are Mean Squared Error (MSE), Mean Absolute Error (MAE) and the correlations of the data sets. Finally, it was detected that Expection Maximization method is the best among the methods with three test results. At the same time, Markov Chain Monte Carlo method gave the second best results among all methods and Regression Imputation method took the third place according the three test methods. The performances of other five methods vary from test to test and they all have close performance scores.

Keywords: Expextation Maximization, Missing Value Imputation Methods, Regression, Markov Chain Monte Carlo

1. GİRİŞ

İstatistik biliminin önemi günümüz dünyasında teknolojinin giderek gelişmesiyle ve yaygın kullanımıyla daha da çok artmıştır. Sosyal bilimlerden tıbba, bilgisayar bilimlerinden tarihe kadar geniş bir kullanım alanına sahip olan istatistiği veri olmadan düşünmek olanaksızdır. Türk Dil Kurumu'na göre verinin tanımları

- Bir araştırmanın, bir tartışmanın, bir muhakemenin temeli olan ana öge, muta, done
- Gözlem ve deneye dayalı araştırmanın sonuçları
- Bilgi, data
- Olgu, kavram veya komutların, iletişim, yorum ve işlem için elverişli biçimli gösterimi

şeklinde belirtilmiştir. Bilgiye ulaşmaya çalışmak bütün bilim dallarının ortak özelliğidir fakat çok fazla durumda istatistik bilimi olmadan bu bilgiyi kullanabilmek, yorumlayabilmek olanaksızdır. Burada önemli olan husus, işlenmesi ve yorumlanabilmesi için verinin analize hazır hale getirilmesidir. Gerçek dünya problemle-

rinde, özellikle sosyal bilimlerde ve anket çalışmalarında yapılan herhangi bir çalışmanın neticeye ulaşabilmesi için elde edilen veri seti eksiksiz olmalıdır. Bu duruma ulaşmanın zor olması, yani eksik veri bulunmayan veri setleri elde etmenin zorluğundan dolayı 1970'li yıllarda başlayan eksik veri atama teknikleri, günümüzde yapılan istatistiksel çalışmalarda kullanılmaktadır.

İstatistiksel analiz yöntemlerinin çoğu 1900'lü yılların başlarında geliştirilmiştir ve bu yöntemler eksik veri içermeyen veriler dikkate alınarak yapılmıştır. Bu yöntemlerin tam bilgi gerektirmesinden dolayı eksik verilerle ilgili herhangi bir çözüm arayışına girilmemiştir. 1970'lerin sonlarında eksik veri sorununun gündeme getirilmesine kadar eksik veri setleri istatistiksel analiz yöntemlerinde herhangi bir problem teşkil etmemekteydi. (Demir ve Parlak, 2012:230).

Eksik verilerin üstündeki zorluklar araştırmacıları tam verilerle birlikte çalışmaya itmektedir. Bazı çalışmalarda da eksiksiz veri seti elde etmenin olanaksız olduğu açıktır. Bu durumda araştırmacılar çoğunlukla eksik veriyi ihmal edilebilir kabul etmekte ve ilgili verileri analiz dışı bırakma eğiliminde olmaktadır. (Demir& Parlak, 2012:230). Özellikle büyük gruplar üzerinde yapılan çalışmalarda eksik olmayan veri seti elde etmek neredeyse imkânsızdır. (Cool, 2000:1).

Bütün istatistiksel yöntemler veri setinin eksiksiz olduğu varsayımı altında geliştiğinden dolayı eksik veriler analizler için yararlanılacak olan istatistiksel analiz yöntemlerinin neredeyse bütünü için sorun oluşturmaktadır. (Allison, 2001:72; Osborne, 2013; Pigott, 2001:353).

Eksik veriler göz önünde bulundurulmadan herhangi bir analiz yanıltıcı sonuçlar verebilir. Ayrıca bu yanıltıcı sonuçların araştırma evrenine genellenebilirliği noktasında hata miktarının artmasının olası olduğu da söylenebilir. (Byrne, 2016). Bu sebeple analize geçilmeden önce eksik verinin miktarı ve yapısıyla ilgili bütün saptamaların yapılması ve karşılaşılabilecek sorunların önlenmesi maksadıyla araştırmacıların önlem alması bir zorunluluk olarak ortaya çıkmaktadır. (Çüm vd., 2018:230).

Eksik değerlerin ele alınmasına yönelik alternatif yöntemler literatürde mevcuttur.

Bu yöntemler aşağıdaki gibi ifade edilebilir (Çokluk ve Kayrı, 2011:289):

1) Eksik veri yerine bir veya birden fazla değer atanması ve bu verilerin analiz dışı bırakılması,

2) Eksik veri içeren gözlem veya değişkenlerin silinmesi,

3) Eksik verilere ilişkin tahminler yapma, yaklaşık bir değer atama (imputation)

Çalışmamızda eksik değerlere ilişkin kestirme yöntemlerinden bahsedileceğinden dolayı buradan sonra üçüncü madde olan "Eksik verilere ilişkin tahminler

yapma, yaklaşık bir değer atama (imputation)" kısmına ağırlık verilecektir. Eksik değerlere müdahale etmenin bilimsel bir yolu, eksik değerlere ilişkin kestirimler yapmak ve bu değerleri temel analizler sırasında kullanmaktır. Ancak kestirim ya da yaklaşık değer atama işlemi sadece nicel değişkenler için yapılabilecek bir işlemdir. Bu kestirimleri yapmanın en yaygın üç yöntemi "geçmiş bilgileri kullanmak", "ortalama değer atamak" ve "regresyondur" (Çokluk ve Kayrı, 2011:289; Mertler ve Vannatta, 2005; Tabachnick ve Fidell, 1996).

Araştırmacı belirli bir alanda uzun zamandan beri çalışıyorsa, ele alınan evren veya değişkenler hakkında bilgi birikimi fazla ise geçmiş bilgileri kullanmak yöntemini kullanabilmektedir. Burada araştırmacı, tecrübelerine dayanarak eksik değerleri yeni değerlerle değiştirebilmektedir. (Mertler ve Vannatta, 2005). Bir diğer yöntem olan ortalama değer atama, eksik verili çalışmaların büyük bir kısmında kullanılmaktadır. Fakat bu yöntemin kullanımı eksik verinin miktarının ve değişkenlerin yapısının uygunluğu durumunda önerilmektedir. Ayrıca bir değişkenin değeri, diğer bir değişkenin değerinin artması gibi durumlarda artışa geçiyorsa ortalama değer atama yöntemi yanıltıcı sonuçlar verebilmektedir. Üçüncü yöntem ise regresyon yaklaşımıdır. Bir veya birkaç değişken bağımsız değişken olarak belirlenir ve eksik veri ataması yapılacak olan gözleme ait değişken ile regresyon modeli oluşturulup tahminler yapılır. Fakat kestirim için kurulacak olan modelde bağımsız değişken ya da değişkenlerin hangi kıstasa göre belirleneceği yine ayrı bir problemdir. Eksik değer ataması yapılacak olan değişken hariç diğer değişkenler bütünüyle bağımsız değişken olarak alındığı takdirde çok fazla bilgi kaybının bulunduğu bir atama söz konusu olabilir. Her değişkende ayrı ayrı eksik verinin bulunabileceği düşünüldüğünde regresyon tahmini yapılacağı zaman bu eksik verilere ait gözlemler devre dışı kalacağından, regresyon tahmininde kullanılacak olan gözlem sayısında bir hayli azalma görülecektir. Bunun önüne geçebilmek için de diğerlerine nazaran daha az eksik gözleme sahip değişken ya da değişkenlerle kestirim yapmak uygun olabilir.

2. EKSİK VERİLERİN YAPISAL ÖZELLİKLERİ

Eksik veriler değişkenlere bağlı olarak üç farklı yapısal özellik gösterebilir. Bunlar (Çüm vd., 2018:230; Allison, 2001:72):

Tamamıyla Rastlantısal Olan Eksik Veriler (TROK): A ve B olarak iki değişken düşünelim. İlgili gözlemlerin verdikleri yanıtların olasılığı bu iki değişkenden bağımsız ise eksik veri Tamamıyla Rastlantısal Eksik Veri olarak düşünülür. Bir başka ifade ile verinin eksik olma nedeni, veri setinde yer alan herhangi bir değişkenden kaynaklı değildir. Örnek olarak, yaş ve cinsiyet birbirlerinden bağımsız değişkenlerdir. Bu durumda A veya B değişkeninin eksik verili bir değişkenler olması TROK özelliğini sağlamaktadır denilebilir. TROK koşulunun sağlandığı ya da sağlanmadığı

yanıt verenler ile yanıt vermeyenler arasındaki gözlenen verilerin dağılımlarının karşılaştırılması ile belirlenebilir.

Rastlantısal Olan Eksik Veriler (ROK): A ve B değişkenleri arasından B'nin yanıt olasılığının A'ya bağlı fakat A'nın yanıt olasılığının B'ye bağlı olmaması durumu rastlantısal olan eksik veri olarak adlandırılır. Örnek olarak, sigara içme durumu ile yaş değişkeni birbirinden bağımsız değişkenler olarak adlandırılabilse de ebeveynden çekindiklerinden dolayı yaşı küçük olan bireyler sigara içmelerine rağmen sigara içmediklerini beyan edebilirler.

Yani sigara içme durumuyla yaş değişkeni arasında bir ilişki olduğu düşünülebilir. Bu durumda ortaya çıkan eksik veri yapısı rastlantısal olan eksik veriler olarak isimlendirilebilir. İhmal Edilemez Eksik Veriler (İEK): TROK ya da ROK yapıda olmayan eksik veriler ihmal edilemez eksik veriler olarak adlandırılmaktadır. Burada değişkenlerdeki eksik yanıt olasılığı A ve B gibi değişkenlerin ikisi ile de ilişkili ise eksik veri ihmal edilemez eksik veri olarak düşünülebilir. İhmal edilemez eksik veriler ile çalışmak diğer yapıdaki eksik verilere göre daha zor ve uzmanlık gereği barındıran bir işlemdir. Eksik verilerin TROK olduğu ve veri setindeki değişkenlerin eksik veri yüzdesi ideal bir seviyede olduğu durumda bütün atama yöntemleri makul sonuçlar verebilir fakat araştırmacı hangi yöntemi neye dayanarak kullandığını bildiği sürece daha akla yatkın ve güvenilir sonuçlar elde edebilecektir. Bu çalışmada veri setinin sadece TROK olduğu durumda kullanılabilecek yöntemler, sadece İEK olduğu durumlarda kullanılabilecek yöntemler birlikte incelenmiş ve yöntemlerin avantajları dezavantajları bu çalışmanın konusu olan veri seti ile karşılaştırılmıştır. Veri setindeki eksik verilerin bulunma durumu değişiklik gösterdikçe, daha önce güvenilir olarak kabul edilen bir yöntem diğer veri setinde güvenilir sonuçlar üretmeyebilir.

3. EKSİK VERİ ATAMA YÖNTEMLERİ

Bu çalışmaya dâhil edilen atama yöntemleri ve kısa açıklamaları aşağıdaki gibi ifade edilebilir:

Beklenti Maksimizasyonu (EM): EM algoritması iki adımdan oluşmaktadır (Akbaş ve Tavşancıl:38, 2015; Enders, 2010),

1) Eksik değerlerin, gözlenen değerler üzerinden kurulan bir regresyon denklemiyle kestirildiği beklenti adımı,

2) Birinci adım sonrasında ulaşılan eksiksiz veri seti üzerinden regresyon denkleminin yeniden kurulduğu maksimizasyon adımı iteratif bir süreçte birbirini izlemektedir.

Markov Zincirleri Monte Carlo (MCMC) Yöntemi: MCMC yöntemi, birkaç adımı izleyerek eksik verilerle Bayesci çıkarımı temel alır:

1) Atama adımı. Bu adımda ortalamalar ve kovaryans matrisi tahmin edilir. Sonrasında her gözlem için eksik değerler simüle edilir.

2) Sonsal (Posterior) Adım (P- Step). Birinci adımdan ortalama vektörü ve kovaryans matrisi simüle edilir. Bu sürecin sonucuna sonsal dağılım denir (Suhaimi vd., 2017:552).

Regresyon Atama: Bu teknikte eksik veriler, eksik veri içermeyen diğer değişkenler kullanılarak kurulan bir regresyon denklemine göre tahmin edilmektedir. Regresyon denklemi, eksik veri içeren değişkenin yordanan, eksik veri içermeyen diğer değişkenlerin ise yordayıcı konumunda olacağı şekilde kurulur. Yordanan değişkende eksik olan gözlemler, diğer değişkenlere ait değerlerin bu denklemde yerine konulmasıyla tahmin edilmekte ve eksiksiz bir veri setine ulaşılmaktadır (Akbaş ve Tavşancıl, 2015:38; Enders, 2010).

Seri Ortalamaları Ataması: Bu atama yöntemi diğer yöntemler arasında en çok kullanılan yöntemdir. Atama yapılacak olan değişkenin eksik olmayan değerlerinin ortalamasının eksik olan gözlemlere atanması işlemidir (Alpar, 2011:139). Sadece TROK yapıda eksik gözlemler için kullanıma uygun olduğu söylense de çok fazla eksik veri bulunan değişkenler için yanlı bir sonuç elde etme ihtimalinden dolayı problem teşkil edebilmektedir. Bu atama yönteminin uygulandığı değişkenlerin gerçek dağılımı farklılaşabilir ve varyans olduğundan düşük çıkabilir. Bahsedilen bu sonuçlar ise yöntemi bir hayli dezavantajlı kılmaktadır.

Yakın Noktalar Ortalama Ataması: Burada yapılan işlem ise, veri setindeki ilgili eksik verinin bulunduğu hücrenin bir üstündeki değer ile bir altındaki değerin ortalaması eksik veriye atanır. Eğer bir üstündeki veri de eksikse, onun bir üstündeki veriyi eksik gözleme atar. Burada hücrenin üstündeki 2 değer ve altındaki 2 değerin yani toplamda 4 değerin ortalama ataması da yapılabilir. SPSS'de "span of nearby points" seçeneği ile bu sayıya araştırmacı karar verebilir. Bu çalışmada "span of nearby points" seçeneği 3 seçilerek işlem yapılmıştır.

Yakın Noktalar Medyan Ataması: Burada yapılan işlem, "span of nearby points" seçeneği 2 seçildiği takdirde, veri setindeki ilgili eksik verinin bulunduğu hücrenin iki üstündeki değer ile iki altındaki değerin medyanını eksik veriye atar. SPSS'de span of nearby points seçeneği ile bu sayıya araştırmacı karar verebilir. Bu çalışmada "span of nearby points" seçeneği 3 seçilerek işlem yapılmıştır.

Lineer İnterpolasyon Ataması: Eksik veriden önceki son tam gözlem değeri ve eksik değerden sonraki ilk tam gözlem değerinin eksik verilerin yerine atanması yöntemidir. Serideki ilk gözlem veya son gözlem eksik olduğu takdirde atama işlemi yapılamaz (Çokluk ve Kayrı, 2011:289).

Noktanın Doğrusal Eğimi (Linear Trend at a Point): Değer, mevcut yapının (örneğin değerler ilk denekten, son deneğe doğru yükselme eğilimi gösteriyorsa) sergilediği trend ile uyumlu veya tutarlı olarak belirlenir. Mevcut serilerin 1'den n'ye kadar ölçeklendirildiği bir indeks değişkeninde eksik verilere öngörülen değerler yerleştirilir (Çokluk ve Kayrı, 2011:289).

4. SINAMA ÇEŞİTLERİ

Çalışmamızda eksik veri atama yöntemlerinin performanslarını karşılaştırmak amacıyla üç farklı sınama çeşidi kullanılmıştır. Bu üç yöntem ve açıklamaları aşağıdaki gibi açıklanmıştır:

Ortalama Mutlak Hata (MAE): Ortalama mutlak hata iki sürekli değişkenin değerleri arasındaki mutlak farkların toplamının değişken sayısına bölümüyle elde edilmektedir. Ortalama mutlak hatanın genel formülü eşitlik 1'de belirtilmiştir.

(1)

Burada Yo, orijinal değişkene ait değeri, Ym ise eksik veri ataması yapılmış değişkene ait değeri temsil etmektedir.

Ortalama Kare Hata (MSE): Ortalama kare hata, iki sürekli değişkenin değerleri arasındaki farkların karelerinin, hesaba katılan değişken sayısına bölümüyle elde edilmektedir. Ortalama kare hatanın genel formülü eşitlik 2'de verilmiştir.

$$\frac{1}{n}\sum_{i=1}^{n}(Yoi - Ymi)^2 \qquad (2)$$

Burada Yo orijinal değişkene ait değeri ve Ym ise eksik veri ataması yapılmış değişkene ait değeri temsil etmektedir.

Pearson Korelasyon Katsayısı: Bu sınama uygulanırken, tüm veri setleri bir değişken olarak düşünülmüştür. Amacımız atama yöntemlerini karşılaştırmak olduğundan, veri setlerine ait değişkenler alt alta gelecek şekilde 3848 x 5 = 19240 gözleme sahip iki değişkenin Pearson korelasyon katsayıları hesaplanarak atama yöntemlerinin performansları karşılaştırılmıştır.

5. VERİ SETİ

Bu çalışmada eksik veri atama yöntemlerinin performanslarını karşılaştırmak için TIMSS* (Trends in International Mathematics and Science Study) veri setinden yararlanılmıştır. TIMSS, öğrencilerin matematik ve fen alanlarında kazandıkları bilgi ve becerilerin değerlendirilmesine yönelik bir tarama araştırmasıdır. Değerlendirme 4. ve 8. sınıf öğrencilerine uygulanmaktadır. Elde edilen çalışmaların sonuçları 4 yılda bir yayınlanmaktadır ve en son 2015 yılında yayınlanmıştır. İlköğretim 4. Sınıf öğrencilerine yönelik herhangi bir ülkeye ait öğrenciler ile ilgili 5 değişkenli, eksik veri bulunmayan ve 3848 gözleme sahip veri seti ele alınmıştır. Bu çalışmada amaçlanan, herhangi tam bir gerçek dünya veri setine sahip olmak olduğundan dolayı bu yazıda veri setinin hangi veri seti olduğu ve neden bu veri setinin seçildiği konusunda daha fazla üzerinde durulmayacaktır.

Burada herhangi bir TIMSS veri setinden, öğrencilerin matematik alanında kazandıkları bilgi ve becerilerin değerlendirilmesi sonucu yayınlanan ve eksik veri bulunmayan 5 değişkenli ve 3848 gözleme sahip bir veri seti üzerinde çalışılmıştır.

*TIMSS 2015 Assessment Frameworks. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Eksiksiz bu veri setinden 5 farklı eksik veri yüzdesi aralığına sahip veri setleri oluşturulmuştur. Bu veri setlerinin her bir değişkeni tamamen rastlantısal olan eksik veri (TROK) yapısındadır. TROK yapıda olan bu veri setlerinin değişkenleri sırasıyla %0-5, %5-10, %10-15, %15-20, %20-25 arasında rastgele yüzdeye sahip eksik verili bir yapıya dönüştürülmüştür. Oluşturulan 5 veri setine ilişkin eksik veri yüzdeleri ve eksik veri sayıları Tablo 1'de belirtilmiştir.

Eksiksiz veri seti 5 farklı eksik veri yüzde aralığına sahip veri setlerine dönüştürülürken değişkenler rastgele eksik verili yapıya getirilmiştir. Buradan bütün değişkenlerin TROK yapısında olduğu sonucu çıkmaktadır. Aynı şekilde her yüzde aralığındaki yüzdelerin kaç olacağına karar verirken de bir rastgelelik söz konusu olmuştur. Bütün bu işlemler yapılmış ve oluşan veri setleri Tablo 1'de gösterilmiştir. Tablo 1'e bakılacak olursa ilk veri seti %0 ile %5 arasında eksik verili değişkenlere sahip bir veri setidir. Eksik veri yüzdeleri de %1.11, %3.62, %1.88, %0.63 ve %4.71 olarak rastgele belirlenmiştir. Diğer yüzde aralığındaki veri setleri için de değişkenlere ait eksik veri yüzdeleri ve o yüzdeye ait kaç verinin o değişkenden silinip eksik verili bir yapıya dönüştürüldüğüne dair bilgiler yer almaktadır.

Bu çalışmada bütün istatistiksel analizler ve eksik veri atama yöntemleri uygulanırken SPSS 22.0 paket programından yararlanılmıştır.

Tablo 1. *Oluşturulan Eksik Veri Setleri, Eksik Veri Yüzdeleri ve Eksik Değer Sayıları*

		Değişkenler									
		X1		X2		X3		X4		X5	
Veri Seti	*Yüzde Aralığı*	*Eksik Veri Yüzdesi ve Sayısı*									
1	%0-5	1.11%	43	3.62%	139	1.88%	72	0.63%	24	4.71%	181
2	%5-10	9.69%	373	6.55%	252	7.32%	282	5.87%	226	6.70%	258
3	%10-15	11.00%	423	13.71%	528	10.17%	391	12.74%	490	14.39%	554
4	%15-20	19.31%	743	17.94%	690	19.31%	743	15.70%	604	16.65%	641
5	%20-25	24.80%	954	23.90%	919	20.40%	785	22.50%	864	20.80%	801

Veri silme işleminin amacı, eksik verili değişkene eksik veri atama yöntemleri uygulandıktan sonra orijinal (eksiksiz değişken) verili değişken ile karşılaştırma yapılması açısındandır.

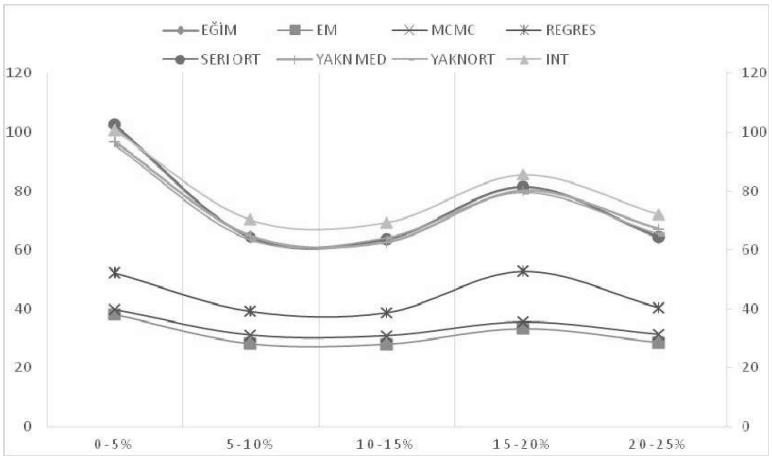
Burada eksik gözlemlerin Tamıyla Rastlantısal Olan Eksik veriler olduğu yorumunu yapılabilir.

6. BULGULAR

Bu bölümde tamamıyla rastlantısal olan eksik veriye sahip değişkenlere çalışmada bahsedilen atama yöntemlerinin tek tek uygulanması sonucu elde edilen bulgulara yer verilmiş ve yöntem sonuçları karşılaştırılmıştır.

Eksik veri atama yöntemlerinin orijinal veri setine ne derece yakın değerler türettiğinin sınanması amacıyla orijinal veri seti ile eksik veri atama yöntemleri sonucunda hesaplanan değerlerin arasındaki ortalama mutlak hatalar hesaplanmıştır ve sonuçlar Grafik 1'de sunulmuştur.

Grafik 1. *Ortalama Mutlak Hata Değerleri*



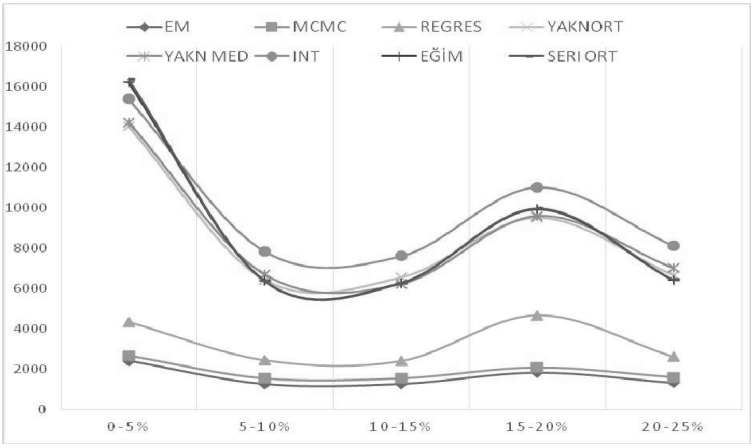
Grafik 1'deki değerler incelendiğinde EM yöntemi ilse Markov Zinciri Monte Carlo yöntemi en iyi sonuçları vermişlerdir. %5-10 ve %20-25 arasında değişen eksik veri setlerinde diğer yüzde aralıklarına göre daha iyi ortalama mutlak hata değerleri gözlenmiştir. Genele bakıldığında %0'dan %25'e kadar olan eksik veri setlerinde ortalama mutlak hata değeri sınama çeşidine göre en iyi sonucu EM yön-

temi vermiştir. Doğrusal İnterpolasyon yöntemi ise bu sınama çeşidine göre en başarısız sonucu vermiştir. Regresyon Atama Yöntemi hariç diğer atama yöntemleri de EM ve Markov Zincirleri Yöntemi 'nin uzağında kalmaktan öteye gidememiştir. Regresyon atama yöntemi EM ve MCMC' nin ardından ortalama mutlak hata değeri sınama çeşidine göre en iyi üçüncü sonucu vermiştir.

Eksik veri atama yöntemlerinin orijinal veri setine ne derece yakın değerler türettiğinin sınanması amacıyla orijinal veri seti ile eksik veri atama yöntemleri sonucunda hesaplanan değerlerin arasındaki hata kareler ortalaması değerleri hesaplanmıştır ve sonuçlar Grafik 2'de sunulmuştur.

Hata kareler ortalamaları değerlerine göre en iyi ve en iyi sonuçları EM, Markov Zinciri Monte Carlo Yöntemi ve Regresyon Atama Yöntemi vermiştir. Diğer yöntemler ise hata kareler ortalamaları sınamasına göre bu üç yönteme göre çok başarısız sonuçlar vermiştir, aynı zamanda diğer yöntemler birbirine oldukça yakın sonuçlar vermişlerdir. Hata kareler ortalamaları sınama çeşidine göre en iyi sonucu EM yöntemi, en iyi ikinci sonucu Markov Zinciri Monte Carlo Yöntemi ve en iyi üçüncü sonucu Regresyon Atama yöntemi vermiştir. Beklenti Maksimizasyonu ve Markov Zinciri Monte Carlo Yöntemi' nin değişik yüzde aralıklarında en istikrarlı sonuçları verdikleri söylenebilir. Bu iki yöntemin haricinde diğer yöntemlerin MSE değerleri değişik yüzdelerde oldukça istikrarsız artıp azalırken bu iki yöntemde artışların ya da azalışların oldukça istikrarlı ve düzgün oldukları söylenebilir.

Grafik 2. *Hata Kareler Ortalamaları Değerleri*

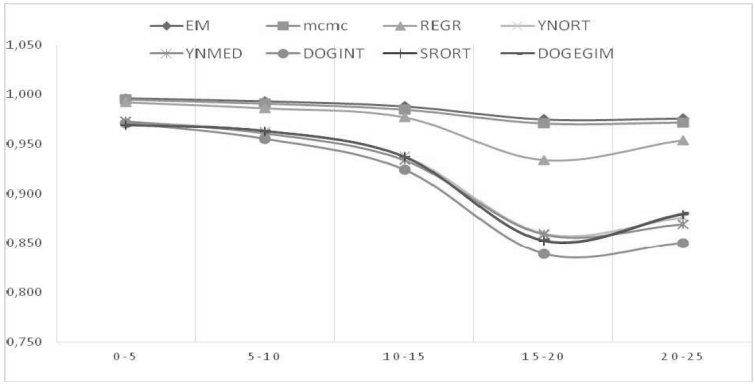


Eksik veri atama yöntemlerinin orijinal veri setine ne derece yakın değerler türettiğinin sınanması amacıyla orijinal veri seti değişkenleri ile eksik veri bulunan veri setlerinin değişkenleri arasındaki Pearson korelasyon katsayılarının yüzde aralıklarına göre ve atama yöntemlerine göre değişen değerleri grafik 3'de belirtilmiştir.

Grafik 3 incelendiğinde, tıpkı MSE' de olduğu gibi en istikrarlı ve en iyi sonuçları sırasıyla Beklenti Maksimizasyonu ve Markov Zinciri Monte Carlo yöntemi vermiştir. Regresyon Atama yöntemi diğer iki sınama yöntemine göre en iyi üçüncü sonucu vermişti ve bu sınama çeşidine göre de en başarılı üçüncü sonucu vermiştir. Diğer iki sınama çeşidine göre diğer atama yöntemleri birbirlerine yakın sonuçlar vermişlerdi ve aynı şekilde korelasyon sınama çeşidine göre de birbirlerine yakın sonuçlar vermişlerdir. Eksik veri yüzdesi arttıkça veri setleri arasındaki korelasyonlar azalışa geçerken %20-25 eksik veri setinde yapılan atama sonuçlarında Beklenti Maksimizasyonu ve Markov Zinciri Monte Carlo yöntemi harici diğer yöntemlerde korelasyon sayılarında bir artış söz konusudur. Bu ise beklenmedik bir sonuç olup, en iyi atama yöntemleri olan Beklenti Maksimizasyonu ve Markov Zinciri Monte Carlo yöntemlerinin önemini azaltmamıştır. Yani korelasyon değerlerindeki bu artış bu çalışmanın sonuçlarını hiçbir şekilde etkilememiştir.

Sonuç olarak, sınama çeşitlerine göre en başarılı ve istikrarlı sonucu Beklenti Maksimizasyonu yöntemi vermiştir. Markov Zinciri Monte Carlo yöntemi de diğer yöntemlerin arasından sıyrılmış ve Beklenti Maksimizasyonu'nun sonuçlarına benzer sonuçlar vermiştir fakat hiçbir şekilde bu yöntemi geçememiştir.

Grafik 3. *Eksik Değerli Veri Setleri ile Eksiksiz Veri Seti Arasındaki Pearson Korelasyon Katsayısı Değerleri*



7. SONUÇ

Eksik verilerde yapılacak atamalarla ilgili; eksik verinin rastgele olup olmamasıyla ilgili, eksik verinin yüzdesini de göz önünde bulundurarak hangi yüzde aralığında hangi yöntemlerin kullanılmasının sağlıklı olup hangilerinin kullanılmaması gerektiğini bir istatistikçi, ilgili veri setini tanıdığı ve değişkenler ile ilgili bilgisi olduğu varsayımı altında en mantıklı yöntemi uygulamalıdır.

Bu çalışmanın sonucu olarak, eksik veri setlerinde eksik veri atama yöntemi olarak yaygın kullanılan bir yöntem olan Seri Ortalamaları yönteminin çoğu zaman ne kadar yanlı sonuçlar verebileceği ispatlanmıştır. Bunun yerine daha iyi sonuçlar veren yöntemlerin mevcut olduğu vurgulanmıştır. Genel olarak bakıldığında Beklenti Maksimizasyonu atama yönteminin bütün farklı eksik verili yüzdeye sahip veri setlerinde en başarılı ve en istikrarlı sonuçları verdiği görülmüştür. Aynı şekilde Markov Zinciri Monte Carlo atama yönteminin de Beklenti Maksimizasyonu kadar iyi olmasa da başarılı ve istikrarlı sonuçlar verdiği görülmüştür. Sürekli değişkenlere sahip eksik veri setlerinde Beklenti Maksimizasyonu (EM) atama yöntemini uygulamak daha mantıklıdır.

KAYNAKÇA

AKBAŞ, U., & TAVŞANCIL, E. (2015). Farklı örneklem büyüklüklerinde ve eksik veri örüntülerinde ölçeklerin psikometrik özelliklerinin eksik veri baş etme teknikleri ile incelenmesi. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6(1).

Allison, P. D. (2001). Missing data (Vol. 136). Sage publications.

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

Byrne, B. M. (2016). Structural equation modeling with AMOS: Basic concepts, applications, and programming. Routledge.

Cool, A. L. (2000). A Review of Methods for Dealing with Missing Data.

Çokluk, Ö., & Kayrı, M. (2011). Eksik değerlere yaklaşık değer atama yöntemlerinin ölçme araçlarının geçerlik ve güvenirliği üzerindeki etkisi. Kuram ve Uygulamada Eğitim Bilimleri, 11(1), 289-309.

Çüm, S., Demir, E. K., Gelbal, S., & Kışla, T. Eksik Veriler Yerine Yaklaşık Değer Atamak için Kullanılan Gelişmiş Yöntemlerin Farklı Koşullar Altında Karşılaştırılması. Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, (45), 230-249.

Demir, E., & Parlak, B. (2012). Türkiye'de eğitim araştırmalarında eksik veri sorunu. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 3(1).

Enders, C. K. (2010). Applied missing data analysis. Guilford press.

Mertler, C. A., & Reinhart, R. V. (2016). Advanced and multivariate statistical methods: Practical application and interpretation. Routledge.

Osborne, J. W. (2013). Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data. Sage.

Pigott, T. D. (2001). A review of methods for missing data. Educational research and evaluation, 7(4), 353-383.

Schafer, J. L. (1999). Multiple imputation: a primer. Statistical methods in medical research, 8(1), 3-15.

Suhaimi, N., Ghazali, N. A., Nasir, M. Y., Mokhtar, M. I. Z., & Ramli, N. A. (2017). MARKOV CHAIN MONTE CARLO METHOD FOR HANDLING MISSING DATA IN AIR QUALITY DATASETS. Malaysian Journal of Analytical Sciences, 21(3), 552-559.

Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics. Allyn & Bacon/Pearson Education.