Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, Eylül 2015, Sayı 35, 87 -111

Kayıp Veriler Yerine Yaklaşık Değer Atamada Kullanılan Farklı Yöntemlerin Model Veri Uyumu Üzerindeki Etkisi

The Effects of Different Methods Used for Value Imputation Instead of Missing Values on Model Data Fit Statistics

Sait ÇÜM

Selahattin GELBAL[1]

Öz

Bu çalışmanın genel amacı, farklı oranlarda ve farklı örüntü yapılarında kayıp veri içeren veri setleri üzerinde, farklı yöntemler kullanılarak yürütülen değer atamalarının, örtük değişkenlerle oluşturulan bir yapısal eşitlik modelinden elde edilen model veri uyum değerleri üzerindeki etkilerinin karşılaştırmalı olarak incelenmesidir. Bu amaç doğrultusunda çalışma, PISA 2012'ye Türkiye'den katılan 15 yaş grubundaki 4848 öğrenci arasından seçilen 1578 öğrenciden elde edilen puanların oluşturduğu veri seti üzerinde yürütülmüştür. Tam veri seti üzerinden eksiltmeler yapılarak tamamıyla rastlantısal olarak dağılan yaklaşık %20 ve %30 oranlarında ve tamamıyla rastlantısal olarak dağılmayan yaklaşık %20 oranında kayıp verilerin oluşturulduğu çalışmada, 10 farklı yöntemle yapılan yaklaşık değer atamalarının model veri uyum değerlerini nasıl etkilediği incelenmiştir. Kayıp verilerin tamamıyla rastlantısal olarak dağıldığı durumlarda regresyonla atama yöntemi sonrası elde edilen veri yapısının modele uyum değerlerinin tam veri setinin modele uyum değerlerine en yakın değerler olduğu sonucuna ulaşılmıştır. Tamamıyla rastlantısal olarak dağılmayan kayıp verilerin bulunması durumunda ise bayesci veri atama ve stokastik regresyonla atama yöntemleri ile yapılan değer atamalarıyla tam veri setinin modele uyum değerlerine en yakın değerlerin elde edildiği belirlenmiştir. Çalışmada ayrıca, kayıp veriler yerine yaklaşık değer atama yöntemlerinin veri dağılımlarını önemli ölçüde değiştirdiği belirlenmiş ve bilinçsizce yapılan atamaların daha sonra yapılacak olan analizlerin sonuçlarını araştırmacıları yanıltacak şekilde etkileyebileceği sonucuna ulaşılmıştır.

Anahtar Kelimeler: Kayıp Veri Analizi, Değer Atama Yöntemleri, Yapısal Eşitlik Modeli, Uyum İyiliği

Abstract

The general purpose of this study is to comparatively research on impacts on data fit statistics which are obtained from structural equation model that is formed by latent variables, data sets including missing data in different pattern structure and at a different rate, value imputation implementing by using different methods. In accordance with this purpose, this study has been conducted on data set formed by points of 1578 students who were chosen among 4848 students- age group of 15-who participated in PISA 2012 from Turkey. In the study of forming lost data of 20% ratios that is not dispersed totally random and of 20% and 30% ratios that are dispersed totally random by subtracting through full data set, it is researched that how value imputations that are conducted in 10 different methods affect model-data fit statistics. In the cases of missing data that is totally random dispersed, it is concluded that model-data fit statistics that is obtained by regression imputation method is close to full data set's model-data fit statistics. In the case of there are data which are not dispersed totally random, it is determined that the most proximate values to data set fit values of model have been obtained by Bayesian data imputation and stochastic regression imputation methods. It is also determined that approximate value imputation instead of missing data have changed data distribution dramatically and it is concluded that analysis results to be conducted after unaware imputation might mislead researchers.

Keywords: Missing Value Analysis, Imputation Methods, Structural Equation Model, Goodness of Fit

[1] Hacettepe Üniversitesi, Eğitim Fakültesi

Giriş

Bilimsel araştırmalar kapsamında üzerinde çalışılmak istenilen veriler bazı nedenlerle istenildiği gibi eksiksiz bir şekilde toplanılamayabilir. İnsanlar üzerinde yapılan araştırmalarda bu eksiklikler araştırmaya katılan bireylerin bilinçli ya da bilinçsiz bir şekilde bazı soruları yanıtsız bırakması, belirlenen süre içerisinde bazı maddelere erişememesi ve çeşitli nedenlerle veri toplama sürecinin bazı aşamalarında bulunamaması gibi katılımcılardan kaynaklı olarak meydana gelebilmektedir. Ayrıca araştırmada kullanılan veri toplama aracının teknik özelliklerinin yetersizliği, aracın uygulanma koşullarının elverişli olmaması, araştırmacının veri girişi sürecinde dikkatsizliği ya da yorgunluğu gibi katılımcılardan kaynaklı olmayan nedenlerden dolayı da eksik veriler ortaya çıkabilmektedir. Veri setlerindeki bu eksiklikler kayıp veriler olarak adlandırılmaktadır. Kayıp veriler analizler için kullanılacak olan klasik ve modern istatistiksel yöntemlerin hemen hemen hepsi için önemli bir sorun oluşturur çünkü tüm yöntemler veri setinin eksiksiz olduğu varsayımı altında geliştirilmiştir. (Pigott, 2001; Allison, 2003; Osborne, 2013).

Sosyal bilimlerde yapılan araştırmalar için kayıp veriler bütün araştırmacıların karşılaştığı yaygın bir sorundur. Özellikle büyük gruplar üzerinde yürütülen çalışmalarda eksiksiz veri setlerinin elde edilmesi neredeyse olanaksızdır (Cool, 2000). Araştırmacılar, elde ettikleri veriler üzerinde gerçekleştirdikleri analizlerden doğru sonuçlar elde edebilmek için kayıp verilerle başa çıkmak zorundadırlar. Bunun için 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 dahil edilen başka değişkenlerle ilişkili olduğu durumlarda yapılacak olan silme işlemi önemli bir yanlılığa yol açabilir (Tabachnick ve Fidell, 1996; Schafer, 1999; Osborne, 2013). 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. Öte yandan Little ve Rubin'e (1987) göre, kayıp veriler yerine bilinçsizce atanan

değerler var olan sorunları ortadan kaldırmadığı gibi ortaya çözümü daha güç olan yeni sorunlar çıkarmaktadır.

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. Kayıp veri miktarının az sayıda ve tamamen rastlantısal olarak dağıldığı (TROK) durumlarda basit atamaya dayalı yöntemler yeterli olabilir fakat aksi durumlarda çoklu atamaya dayalı yöntemlerin daha güvenilir sonuçlar vereceği belirtilmektedir (Schafer, 1999; Osborne, 2013). Günümüzde, birçok farklı istatistik paket programı kullanıcılara farklı yöntemlerle kayıp verilerle başa çıkma olanağı sağlamaktadır. Fakat henüz hangi değişken yapısında hangi miktardaki kayıplar için hangi yöntemin kullanılmasının daha uygun olacağına dair kesin ifadeler kullanmak mümkün değildir. Son yıllarda yurtiçi ve yurtdışı alanyazında çeşitli analiz yöntemlerinden elde edilen parametreler bağlamında farklı yaklaşık değer atama yöntemlerinin karşılaştırıldığı çalışmaların artmaya başladığı görülmektedir. Fakat kayıp verilerle başa çıkma konusunda daha güçlü öneriler getirilebilmesi bakımından mevcut çalışmaların sayısının henüz çok yetersiz olduğu iddia edilebilir.

Alanyazında kayıp veri konusunda yapılan çalışmalar şöyle özetlenebilir: Witta (2000), 35 maddeden oluşan beş kategorili likert tipi bir ölçekle toplanan veriler içerisinde yer alan kayıplara farklı yöntemlerle atamalar yaparak faktör analizi sonucu elde edilen değerleri incelemiştir. Garrett (2009), kayıp verilerin varlığında farklı yöntemlerle atamalar yaparak bu yöntemleri bir değişen madde fonksiyonu belirleme çalışması kapsamında karşılaştırmıştır. Çokluk ve Kayrı'nın (2011), kayıp değerlere yaklaşık değer atama yöntemlerinin ölçme araçlarının geçerlik ve güvenirliği üzerindeki etkisini araştırdıkları çalışmada, yaygın kullanılan beş farklı yaklaşık değer atama yöntemi karşılaştırılmış ve yapılan atamaların hem açıklanan varyans oranlarında hem de güvenirlik değerlerinde düşüşe yol açtığı belirtilmiştir. Demir (2013), kayıp verilerin varlığında, çoktan seçmeli testlerde, farklı kayıp veri yöntemleri kullanılarak kestirilen madde ve test parametreleri arasındaki ilişkileri incelediği çalışmasında en çok olabilirlik ve çoklu veri atama yöntemlerinin kayıp verilerin tamamıyla rastlantısal dağılmadığı durumlarda kullanılmasının uygun olduğu sonucuna ulaşmıştır. Köse ve Öztemur

(2014), kayıp veri ele alma yöntemlerinin t-testi ve anova parametreleri üzerindeki etkisini inceledikleri çalışmalarında, düşük birimli veri setlerinde regresyon ve beklenti maksimizasyonu yöntemlerinin tam veri setinden elde edilen değerlere en yakın sonuçları ürettiğini, yüksek birimli veri setlerinde ise regresyon ve ortalama atama yöntemlerinin tam veri setiyle daha tutarlı sonuçlar verdiğini belirtmişlerdir. Köse (2014), kayıp değer ele alma yöntemlerinin doğrulayıcı faktör analizinden elde edilen uyum değerlerine etkisini araştırdığı çalışmasında, en çok olabilirlik kestirimiyle atama sonucu elde edilen uyum değerlerinin tam veri setinden elde edilen uyum değerlerine en yakın sonuçları ürettiği sonucuna ulaşmış fakat çalışmanın tamamıyla rastlantısal kayıp veri örüntüsüne sahip simülasyon veriler üzerinde yürütülmesini bir sınırlılık olarak belirtmiştir. Ayrıca, yurtdışı alanyazında, yaklaşık değer atama yöntemlerinin az sayıda veriden oluşan küçük örnekler üzerinde karşılaştırıldığı çalışmaların olduğu görülmüştür. Kayıp veri alanında yapılan çalışmaların çoğunda simülasyon veriler kullanılmış ve genelde tamamıyla rastlantısal olarak oluşturulan kayıp veri örüntüleri üzerinde çalışılmıştır.

Bu çalışmada, diğer araştırmalardan farklı olarak hem tamamıyla rastlantısal hem de tamamıyla rastlantısal olmayan kayıp veri örüntülerine sahip PISA (Programme for International Student Assessment) 2012 veri setlerinin kullanılmış olması çalışmanın önemini artırmaktadır. Yapılan tartışmalar doğrultusunda bu çalışmanın genel amacı, PISA 2012 verilerinden oluşturulan, farklı oranlarda ve farklı örüntü yapılarında kayıp veri içeren veri setleri üzerinde, farklı yöntemler kullanılarak yürütülen değer atamalarının, örtük değişkenlerle oluşturulan bir yapısal eşitlik modelinden elde edilen model veri uyum değerleri üzerindeki etkilerinin karşılaştırmalı olarak incelenmesidir. Bu genel amaç doğrultusunda aşağıdaki sorulara yanıt aranmıştır:

1. Tamamıyla rastlantısal olarak dağılan yaklaşık %20 oranında kayıp veri içeren veri setinde kayıp veriler yerine hangi yöntem veya yöntemlerle yaklaşık değerler atanması sonucu elde edilen model veri uyum değerleri tam veri setinden elde edilen model veri uyum değerlerine en yakın sonuçları vermektedir?
2. Tamamıyla rastlantısal olarak dağılan yaklaşık %30 oranında kayıp veri içeren veri setinde kayıp veriler yerine hangi yöntem veya yöntemlerle yaklaşık değerler atanması sonucu elde edilen model veri uyum değerleri tam veri setinden elde edilen model veri uyum değerlerine en yakın sonuçları vermektedir?
3. Tamamıyla rastlantısal olarak dağılmayan yaklaşık %20 oranında kayıp veri içeren veri setinde kayıp veriler yerine hangi yöntem veya yöntemlerle yaklaşık değerler

atanması sonucu elde edilen model veri uyum değerleri tam veri setinden elde edilen model veri uyum değerlerine en yakın sonuçları vermektedir?

Yöntem

Araştırmanın Modeli

Bu çalışmada, kayıp veriler yerine yaklaşık değer atamada kullanılan farklı yöntemler, bu yöntemler kullanılarak elde edilen veri yapılarının oluşturulan yapısal eşitlik modeline uyumunu gösteren değerler üzerinden karşılaştırılmıştır. Araştırma, bilgi üretmeye yönelik kuramsal bir çalışma özelliği taşıması bakımından temel araştırma niteliğindedir.

Çalışma Grubu

Bu çalışma, PISA 2012'ye Türkiye'den katılan 15 yaş grubundaki 4848 öğrenci arasından seçilen matematik öz yeterliği, matematik kavramlarına aşinalık ve matematik okuryazarlığı alt boyutlarının hepsine katılım göstermiş ve karşılaştığı maddelerin tümüne hiç kayıp veri oluşturmayacak şekilde tepkide bulunmuş olan 1578 öğrenciden elde edilen puanların oluşturduğu veri seti üzerinde yürütülmüştür.

Veri Toplama Aracı

Çalışmada, PISA 2012'de yer alan matematik öz yeterliği, matematik kavramlarına aşinalık, matematik okuryazarlığı ölçek ve anketleri ile toplanmış olan veriler kullanılmıştır. PISA 2012, öğrencilerin, okuma, matematik, fen, problem çözme ve finans alanlarında sahip oldukları yetenek, bilgi ve becerileri belirlemek amacıyla OECD'ye üye 34 ülkede ve OECD ülkeleri ile ekonomik işbirliği içerisinde olan 31 ülkede olmak üzere toplam 65 ülkede uygulanmıştır. Ayrıca PISA kapsamında öğrencilerle ilgili daha derinlemesine bilgi edinebilmek amacıyla öğrenci, okul ve veli anketleri de uygulanmaktadır (OECD, 2014).

Verilerin Analizi

Çalışmada öncelikle, matematik öz yeterliği ve matematik kavramlarına aşinalık yapılarının matematik okuryazarlığını etkilediği varsayımıyla teorik bir model oluşturulmuştur. Söz konusu değişkenler modele dahil edilmeden önce her bir değişkenin yapısı açıklayıcı faktör analizi yöntemi ile incelenmiştir. Bu incelemeler sonucu matematik öz yeterliği yapısının özdeğeri 1'den büyük olan iki faktör tarafından yaklaşık %58 oranında açıklandığı görülmüştür. Yapıyı tek boyuta indirgemek amacıyla her iki faktör altında da yüksek miktarda yük verdiği gözlemlenen 4. ve 6. maddeler analiz dışı bırakılarak yapının tek bir faktör altında incelenmesi istenilmiştir. Yapılan son incelemede tüm maddelerin tek faktör altında toplandığı fakat 8. maddenin düşük bir faktör yüküne sahip olduğu belirlenmiş ve bu maddenin de modele dahil edilmemesi konusunda karar alınmıştır. Tablo 1'de söz konusu

maddeler çıkarıldıktan sonra matematik öz yeterliği yapısında yer alan maddelerin faktör yük değerleri verilmiş ve Şekil 1'de yapıya ilişkin yamaç birikinti grafiği sunulmuştur.

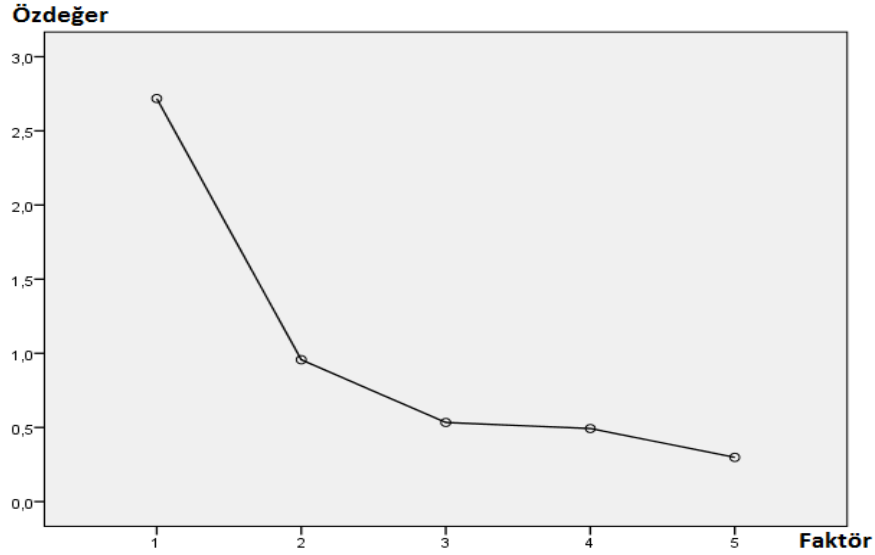
Tablo 1.

Matematik Öz Yeterliği Madde Faktör Yükü Değerleri

	Madde	**Faktör Yükü**
Faktörün özdeğeri 2,718; açıklanan varyans %54,364	1	,712
	2	,727
	3	,734
	5	,765
	7	,747

Şekil 1.

Matematik Öz Yeterliği Yapısına İlişkin Yamaç Birikinti Grafiği



Benzer süreç matematik kavramlarına aşinalık yapısı üzerinde de yürütülmüş ve yapının tek boyuta indirgenmesi amacıyla 1, 4, 6, 8, 11, 13 ve 15. maddeler modele dahil edilmemiştir. Tablo 2' de söz konusu maddeler çıkarıldıktan sonra matematik kavramlarına aşinalık yapısında yer alan maddelerin faktör yük değerleri verilmiş ve Şekil 2'de yapıya ilişkin yamaç birikinti grafiği sunulmuştur.

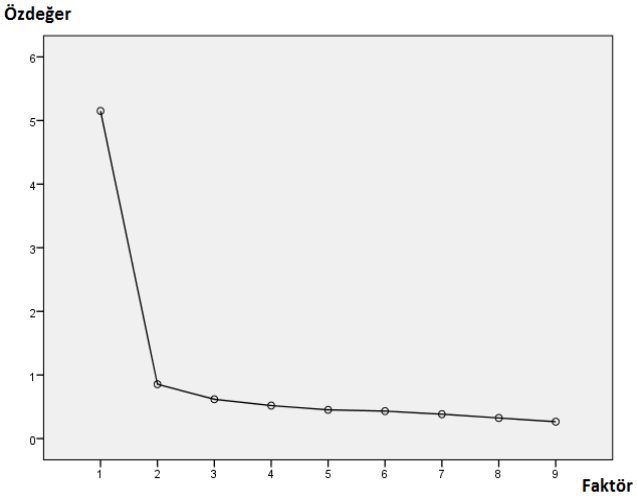
Tablo 2.

Matematik Kavramlarına Aşinalık Madde-Faktör Yükü Değerleri

	Madde	Faktör Yükü
Faktörün özdeğeri 5,150; açıklanan varyans %57,228	2	,745
	3	,664
	7	,752
	9	,829
	10	,809
	12	,773
	16	,667
	17	,762
	19	,790

Şekil 2.

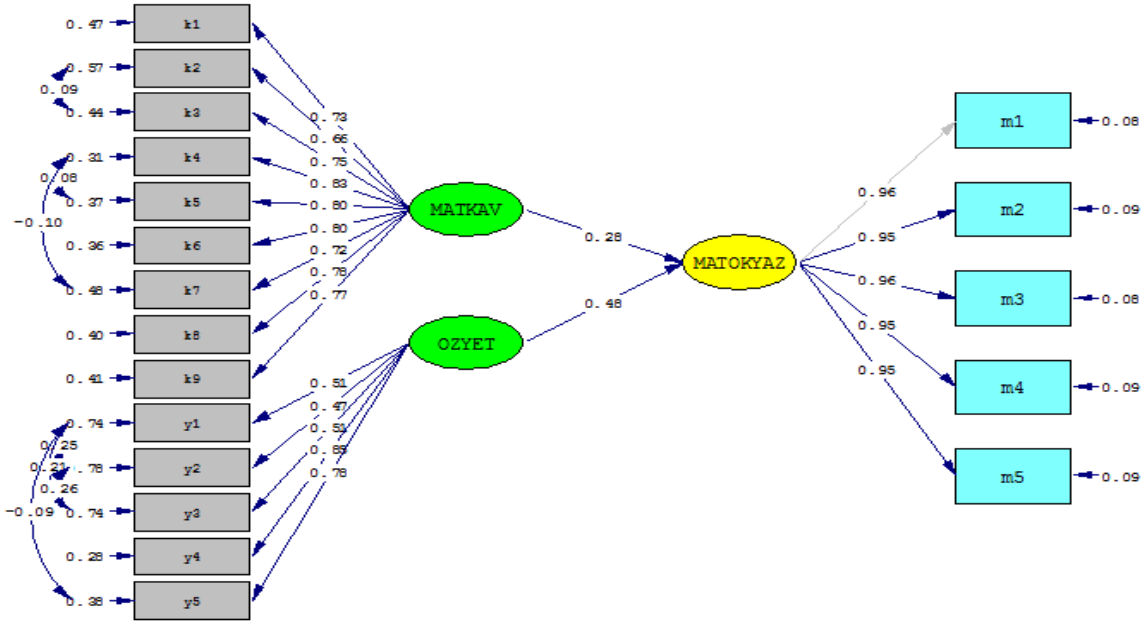
Matematik Kavramlarına Aşinalık Yapısına İlişkin Yamaç Birikinti Grafiği



Bu çalışmanın birincil amacı matematik okuryazarlığını etkileyen değişkenleri belirlemek değil, veriye iyi uyum sağlayan bir model üzerinde kayıp verilere yapılan atamaların model veri uyumuna etkisini incelemektir. Söz konusu yapılar içerisinde yer alan bir çok maddenin modele dahil edilmemesi bu gerekçe ile açıklanabilir. Bu aşamalardan sonra modele eklenen değişkenlerin meydana getirdiği ölçme modelleri ve oluşturulan teorik model LISREL 8.8 programı ile test edilerek modelin veriye uyumu incelenmiştir. Programın model için vermiş olduğu görsel şekil 3'te sunulmuştur.

Şekil 3.

Matematik Okuryazarlığı İçin Yapısal Eşitlik Modeli



Model incelendiğinde, modelde yer alan tüm yolların anlamlı olduğu ve yük değerlerinin yüksek olduğu belirlenmiştir. Matematik öz yeterliği ve matematik kavramlarına aşinalık değişkenleri matematik okuryazarlığı değişkenindeki varyansın %46'sını açıklamaktadır (R^2=0,46). Ayrıca model veri uyum değerleri incelendiğinde, RMSEA=0,049, CFI=0,99, NFI=0,99, NNFI=0,99, SRMR=0,046, AGFI=0,94, GFI=0,96 olduğu belirlenmiştir. Bu değerler modelin veriye iyi uyum sağladığını göstermektedir.

Tam veri seti üzerinde model test edildikten sonra, birinci ve ikinci araştırma sorularına yanıt aramak amacıyla veri seti içerisinde %20 ve %30 oranlarında tamamıyla rastlantısal dağılan kayıp veriler oluşturulmuştur. Dağılımın rastlantısallığının kontrol edilmesi amacıyla oluşturulan veri setleri Little'ın MCAR testine tabi tutulmuştur. Her iki veri seti için de elde edilen p değerlerinin istatistiksel olarak manidar farklılıklara işaret etmediği görülmüştür (%20 kayıp verili grup için p=0,294; %30 kayıp verili grup için p=0,466). Bu bulgular, veri setlerinde yer alan kayıp verilerin tamamıyla rastlantısal dağıldığını göstermektedir.

Üçüncü araştırma sorusuna yanıt aramak amacıyla her bir birey için matematik öz yeterliği, matematik kavramlarına aşinalık ve matematik okuryazarlığı toplam puanları ve bu puanların ortalamaları hesaplanmıştır. Toplam test puanı ortalamanın altında kalan bireyler düşük yetenekli bireyler olarak kabul edilmiş ve bu bireylerin belirli maddelere yanıt veremediği varsayımı altında veri seti içerisinde önceden belirlenmiş maddeler bazında kayıp

veriler oluşturulmuştur. Böylece kayıp verilerin yetenek düzeyi değişkeni ile ilişkisinin kurulması amaçlanmıştır. Eksiltilen veri seti içerisindeki kayıp verilerin genele oranı yaklaşık %20 olarak belirlenmiştir. Kayıp veri dağılımının rastlantısallığı Little'ın MCAR testi ile kontrol edilmiş ve elde edilen p değeri istatistiksel olarak manidar bir farklılığa işaret etmiştir (p=0,000). Bu bulgu, yetenek düzeylerine bağlı olarak oluşturulan kayıp verilerin tamamıyla rastlantısal dağılmadığını göstermektedir.

Belirtilen aşamalardan sonra, farklı oranlarda ve farklı örüntü yapılarında kayıp veriler içeren veri setlerindeki kayıp verilerin yerine seriler ortalaması (SO), beklenti maksimizasyonu (BM), doğrusal değerleme (DD), regresyon ataması (RA), stokastik regresyon ataması (SRA), Markov Zincirleri Monte Carlo (MZM), NIPALS algoritması (NIP), bayesci veri atama (BVA), eğilim skorları eşleştirmesi (ESE), mahalanobis uzaklığı ataması (MUA) yöntemleri ile değer atamaları yapılmıştır. Değer atamaları için, SPSS V22, AMOS V22, LISREL V8.8, XLSTAT 2014 ve SOLAS for Missing Data Analysis V4.02 programları kullanılmıştır. Farklı yöntemlerle yapılan yaklaşık değer atamaları sonucu elde edilen her bir veri yapısının yapısal eşitlik modeline uyumu test edilmiş ve elde edilen değerler tam veri setinden elde edilen değerler ile karşılaştırılarak yorumlanmıştır. Karşılaştırmalar için kullanılan çizgi ve histogram grafikleri için STATISTICA V7 programından yararlanılmıştır.

Bulgular

Bu bölümde, yapılan analizler sonucu elde edilen bulgular ve bulgulara dayalı olarak yapılan yorumlar araştırma sorularının sırasına uygun olarak sunulmuştur.

Öncelikle, ''Tamamıyla rastlantısal olarak dağılan (TROK) yaklaşık %20 oranında kayıp veri içeren veri setinde kayıp veriler yerine hangi yöntem veya yöntemlerle yaklaşık değerler atanması sonucu elde edilen model veri uyum değerleri tam veri setinden elde edilen model veri uyum değerlerine en yakın sonuçları vermektedir?'' sorusuna yanıt aranmıştır. Bu amaçla oluşturulan veri setindeki kayıp veriler yerine farklı yöntemlerle yapılan yaklaşık değer atamaları sonucu elde edilen veri yapılarının yapısal eşitlik modeline uyum değerlerinin karşılaştırılması amacıyla oluşturulan Tablo 3 aşağıda sunulmuştur.

Tablo 3.

Farklı Yöntemlerle %20 Oranında Trok Veriye Atama Yapılması Sonucu Elde Edilen Model Veri Uyum Değerleri

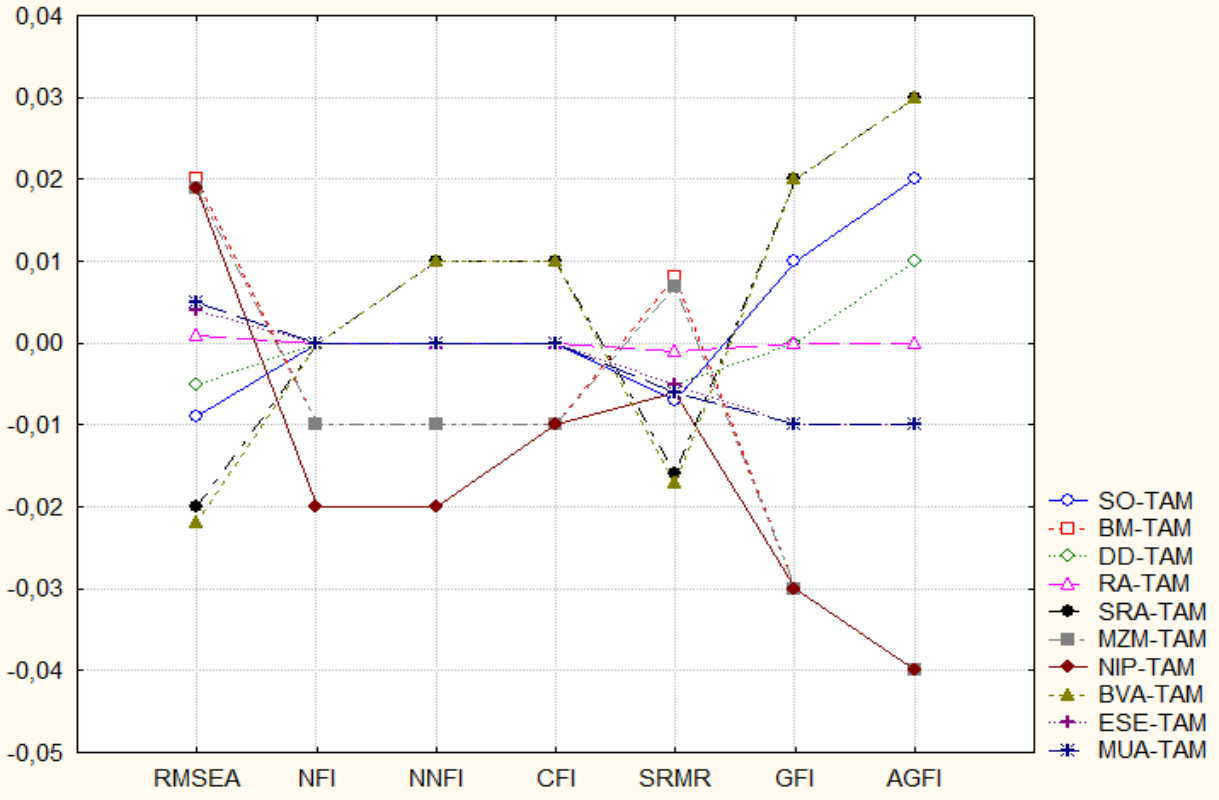
	x^2	RMSEA	NFI	NNFI	CFI	SRMR	GFI	AGFI
Tam veri	673,44	0,049	0,99	0,99	0,99	0,046	0,96	0,94
SO	503,45	0,040	0,99	0,99	0,99	0,039	0,97	0,96

BM	1203,80	0,069	0,98	0,98	0,98	0,054	0,93	0,90
DD	583,99	0,044	0,99	0,99	0,99	0,041	0,96	0,95
RA	695,35	0,050	0,99	0,99	0,99	0,045	0,96	0,94
SRA	329,06	0,029	0,99	1,00	1,00	0,030	0,98	0,97
MZM	1165,58	0,068	0,98	0,98	0,98	0,053	0,93	0,90
NIP	1178,32	0,068	0,97	0,97	0,98	0,040	0,93	0,90
BVA	304,59	0,027	0,99	1,00	1,00	0,029	0,98	0,97
ESE	781,87	0,053	0,99	0,99	0,99	0,041	0,95	0,93
MUA	788,16	0,054	0,99	0,99	0,99	0,040	0,95	0,93

Tablo 3 incelendiğinde, farklı yöntemlerle kayıp değerler yerine yapılan yaklaşık değer atamalarının, ki-kare değerlerinde büyük farklılaşmalara neden olduğu görülmektedir. Özellikle bayesci veri atama (BVA) ve stokastik regresyonla veri atama (SRA) yöntemleri ile yapılan atamalardan sonra oluşan veri yapılarının yapısal eşitlik modeli ile test edilmesinden sonra elde edilen ki-kare değerlerinde tam veri setinden elde edilen ki-kare değerine göre önemli bir düşüş meydana geldiği görülmektedir. Ayrıca söz konusu atamalardan sonra elde edilen model veri uyum değerlerinin de tam veri setinden elde edilen uyum değerlerine göre daha iyi bir uyuma işaret eder düzeye geldikleri görülmektedir. Tablo 3'ten, tam veri setinin modele uyum değerlerine en yakın uyum değerlerinin regresyon yöntemi (RA) ile atamaların yapıldığı veri setinden elde edildiği anlaşılmaktadır. Yaklaşık %20 oranında TROK veri yerine farklı yöntemler ile yapılan atamalardan sonra oluşan veri yapılarının model veri uyum değerlerinin tam veri setinden elde edilen model veri uyum değerlerine göre ne kadar farklılaştığının daha açık bir şekilde gözlemlenebilmesi amacıyla çoklu çizgi grafiği oluşturulmuş ve elde edilen bulgular Şekil 4'te sunulmuştur.

Şekil 4.

Farklı Yöntemlerle %20 Oranında Trok Veriye Atama Yapılması Sonucu Elde Edilen Model Veri Uyum Değerleri İle Tam Veri Seti Model Veri Uyum Değerleri Arasındaki Farklar



Şekil 4 incelendiğinde, tam veri setinin yapısal eşitlik modeline uyum değerlerine en yakın uyum değerlerinin RA sonrası oluşan veri yapısından elde edildiği görülmektedir (y=0,00 doğrusu üzerine düşen noktalarda uyum değerleri arasında fark yoktur). Tam veri seti uyum değerlerine göre en çok farklılaşma gösteren uyum değerlerinin ise BVA, NIP ve SRA yöntemleri ile yapılan atamalar sonrası oluşan veri yapılarından elde edildiği görülmektedir. En yüksek farklılaşmanın AGFI uyum indeksi bazında 0,04 düzeyinde olduğu belirlenmiştir.

Kayıp veriler yerine BVA ve SRA yöntemleriyle atama yapıldığında elde edilen birçok uyum değerinin y=0,00 doğrusunun üst kısmında yer aldığı görülmektedir. Başka bir ifadeyle söz konusu yöntemlerle yapılan atamalardan sonra oluşan veri yapıları modele tam veri setinden daha iyi uyum sağlamaktadır. Bu durum, kayıp veriler yerine yaklaşık değer atama yöntemlerinin kullanımında oldukça dikkatli olunması gerektiğinin bir göstergesi olarak öne sürülebilir. Yüksek oranlarda tamamıyla rastlantısal dağılım gösteren kayıp veri içeren veri yapılarının kullanıldığı çalışmalarda kayıp veriler yerine yaklaşık değerlerin atanması sonucu modele gerçekte iyi uyum sağlayamayacak olan bir veri yapısının araştırmacıları yanıltıcı yönde değişime uğraması durumu ortaya çıkabilir. Schafer'a (1999) göre, araştırmacının elinde bulunan veri yapısının çarpık ya da basık bir dağılım göstermesi durumunda kayıp veriler yerine normallik varsayımı altında yürütülen yaklaşık değer atama süreci sonrasında oluşan veri yapısı kullanılarak yapılacak olan analizler tam veri seti üzerinde yapılan analizlere göre daha iyi sonuçlar verebilir. Yaklaşık değer atama

yöntemlerinin birçoğu verilerin normal dağıldığını varsaymaktadır. Bu bağlamda, ortaya çıkan durumun derinlemesine incelenebilmesi amacıyla tamamıyla rastlantısal olarak dağılan yaklaşık %20 oranında kayıp veri yerine yapılan atamalardan sonra elde edilen matematik öz yeterliği toplam puan dağılımları histogram grafikleri incelenmiş ve söz konusu grafikler *EK 2*'de sunulmuştur. Yapılan incelemeler sonrasında tam veri setinden elde edilen matematik öz yeterliği toplam puanları dağılımının (EK 1) kayıp verilere yapılan atamalar sonrası değişime uğradığı görülmüştür. Örneklenen yaklaşık değer atama yöntemlerinin tümü söz konusu puanların dağılımını etkilemiş özellikle bayesci veri atama (BVA) ile yapılan yaklaşık değer atamaları sonrası dağılımın diğer yöntemlere oranla normale daha fazla yaklaştığı belirlenmiştir. Veri yapılarında meydana gelen bu tür değişimlerin yapılacak olan analizlerden elde edilecek olan değerleri etkilemesi olasıdır.

Bu aşamadan sonra, ''Tamamıyla rastlantısal olarak dağılan (TROK) yaklaşık %30 oranında kayıp veri içeren veri setinde kayıp veriler yerine hangi yöntem veya yöntemlerle yaklaşık değerler atanması sonucu elde edilen model veri uyum değerleri tam veri setinden elde edilen model veri uyum değerlerine en yakın sonuçları vermektedir?'' sorusuna yanıt aranmıştır. Bu amaçla oluşturulan veri setindeki kayıp veriler yerine farklı yöntemlerle yapılan yaklaşık değer atamaları sonucu elde edilen veri yapılarının yapısal eşitlik modeline uyum değerlerinin karşılaştırılması amacıyla oluşturulan Tablo 4 aşağıda sunulmuştur.

Tablo 4.

Farklı Yöntemlerle %30 Oranında Trok Veriye Atama Yapılması Sonucu Elde Edilen Model Veri Uyum Değerleri

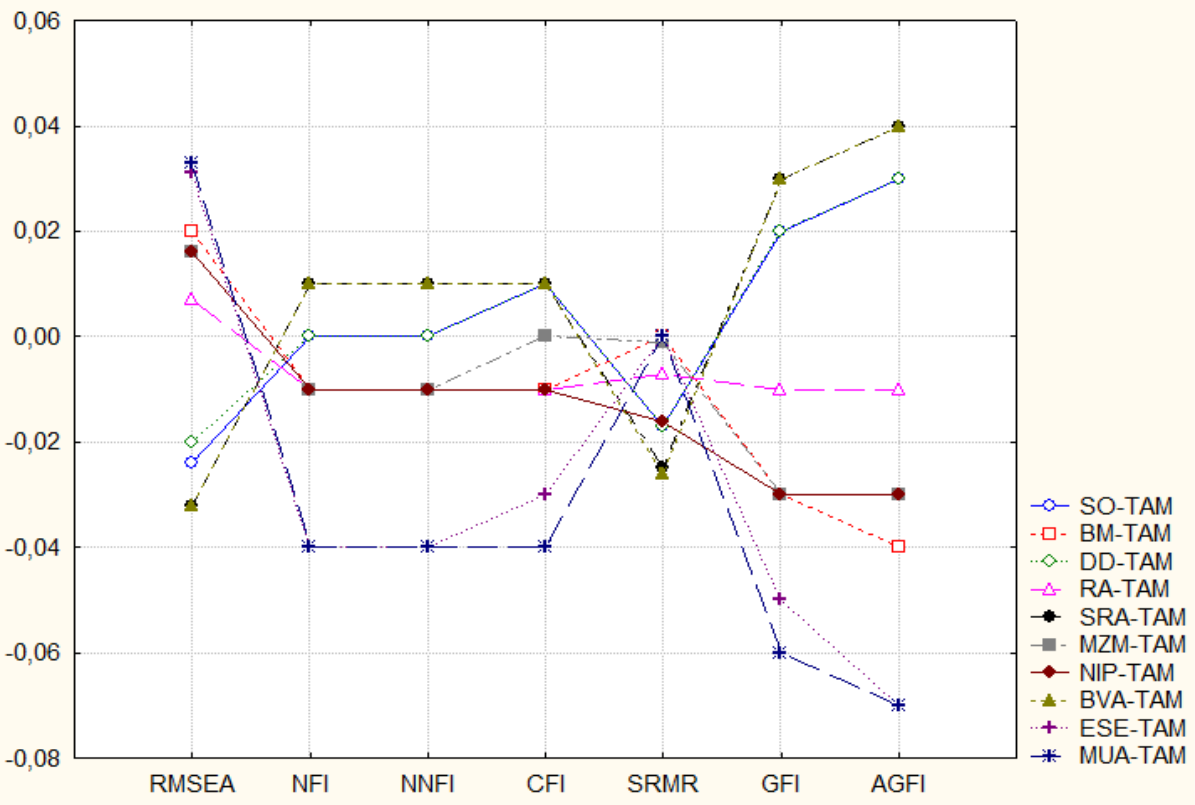
	x^2	RMSEA	NFI	NNFI	CFI	SRMR	GFI	AGFI
Tam veri	673,44	0,049	0,99	0,99	0,99	0,046	0,96	0,94
SO	285,51	0,025	0,99	0,99	1,00	0,029	0,98	0,97
BM	1211,29	0,069	0,98	0,98	0,98	0,046	0,93	0,90
DD	330,00	0,029	0,99	0,99	1,00	0,029	0,98	0,97
RA	852,00	0,056	0,98	0,98	0,98	0,039	0,95	0,93
SRA	203,03	0,017	1,00	1,00	1,00	0,021	0,99	0,98
MZM	1097,81	0,065	0,98	0,98	0,99	0,045	0,93	0,91
NIP	1097,22	0,065	0,98	0,98	0,98	0,030	0,93	0,91
BVA	210,54	0,017	1,00	1,00	1,00	0,020	0,99	0,98
ESE	1560,91	0,080	0,95	0,95	0,96	0,046	0,91	0,87
MUA	1629,80	0,082	0,95	0,95	0,95	0,046	0,90	0,87

Tablo 4 incelendiğinde, tamamıyla rastlantısal olarak dağılan yaklaşık %30 oranında kayıp veri yerine yapılan yaklaşık değer atamaları sonucu elde edilen veri yapılarının modele uyum değerlerinin tam veri setinin modele uyum değerlerine göre önemli düzeyde

farklılaştığı görülmektedir. Tablo 1'de verilen bulgulara benzer olarak bayesci veri atama (BVA) ve stokastik regresyonla veri atama (SRA) yöntemleri ile yapılan atamalardan sonra elde edilen ki-kare değerlerinde tam veri setinden elde edilen ki-kare değerine göre büyük bir düşüş; eğilim skorları eşitlemesi (ESE) ve mahalanobis uzaklığı ile veri atama (MUA) yöntemleri ile yapılan atamalardan sonra elde edilen ki-kare değerlerinde ise büyük bir artış olduğu belirlenmiştir. Ayrıca yine daha önceki bulgulara paralel olarak tam veri setinden elde edilen model veri uyum değerlerine en yakın uyum değerlerinin regresyon ataması (RA) sonrası elde edildiği görülmektedir. Yaklaşık %30 oranında TROK veri yerine farklı yöntemler ile yapılan atamalardan sonra oluşan veri yapılarının model veri uyum değerlerinin tam veri setinden elde edilen model veri uyum değerlerine göre ne kadar farklılaştığının daha açık bir şekilde gözlemlenebilmesi amacıyla çoklu çizgi grafiği oluşturulmuş ve elde edilen bulgular Şekil 5'te sunulmuştur.

Şekil 5.

Farklı yöntemlerle %30 oranında TROK veriye atama yapılması sonucu elde edilen model veri uyum değerleri ile tam veri seti model veri uyum değerleri arasındaki farklar



Şekil 5 incelendiğinde, kayıp veriler yerine RA yöntemi ile atama yapılarak oluşturulan veri yapısının tam veri setinin modele uyum değerlerine en yakın ve tutarlı uyum değerlerini verdiği belirlenmiştir. Ayrıca, tamamıyla rastlantısal olarak dağılan kayıp veri oranının artmasıyla birlikte atama yapılan veri yapılarının modele uyum değerlerinin tam veri

setinden elde edilen uyum değerlerinden daha fazla farklılaştığı belirlenmiştir. En yüksek farklılaşma AGFI uyum indeksi bazında 0,07 düzeyindedir. Yaklaşık %20 oranında TROK veri üzerinden elde edilen bulgulara paralel olarak yaklaşık %30 oranındaki TROK veriye BVA ve SRA yöntemleri ile atama yapılması sonucu oluşan veri yapılarının modele uyum değerleri tam veri setinin modele uyum değerlerinden daha iyi uyumu göstermektedir. Bu bağlamda, ortaya çıkan durumun derinlemesine incelenebilmesi amacıyla tamamıyla rastlantısal olarak dağılan yaklaşık %30 oranında kayıp veri yerine yapılan atamalardan sonra elde edilen matematik öz yeterliği toplam puan dağılımları histogram grafikleri incelenmiş ve söz konusu grafikler *EK 3*'te sunulmuştur. Yapılan incelemeler sonucunda tamamıyla rastlantısal olarak dağılan kayıp veri oranının artması sonucu yapılan değer atamalarının matematik öz yeterliği toplam puanlarının dağılımlarını normal dağılıma bir miktar daha yaklaştırdığı belirlenmiştir.

Son olarak, ''Tamamıyla rastlantısal olarak dağılmayan yaklaşık %20 oranında kayıp veri içeren veri setinde kayıp veriler yerine hangi yöntem veya yöntemlerle yaklaşık değerler atanması sonucu elde edilen model veri uyum değerleri tam veri setinden elde edilen model veri uyum değerlerine en yakın sonuçları vermektedir?'' sorusuna yanıt aranmıştır. Bu amaçla oluşturulan veri setindeki kayıp veriler yerine farklı yöntemlerle yapılan yaklaşık değer atamaları sonucu elde edilen veri yapılarının yapısal eşitlik modeline uyum değerlerinin karşılaştırılması amacıyla oluşturulan Tablo 5 aşağıda sunulmuştur.

Tablo 5.

Farklı Yöntemlerle %20 Oranında Trok Olmayan Veriye Atama Yapılması Sonucu Elde Edilen Model Veri Uyum Değerleri

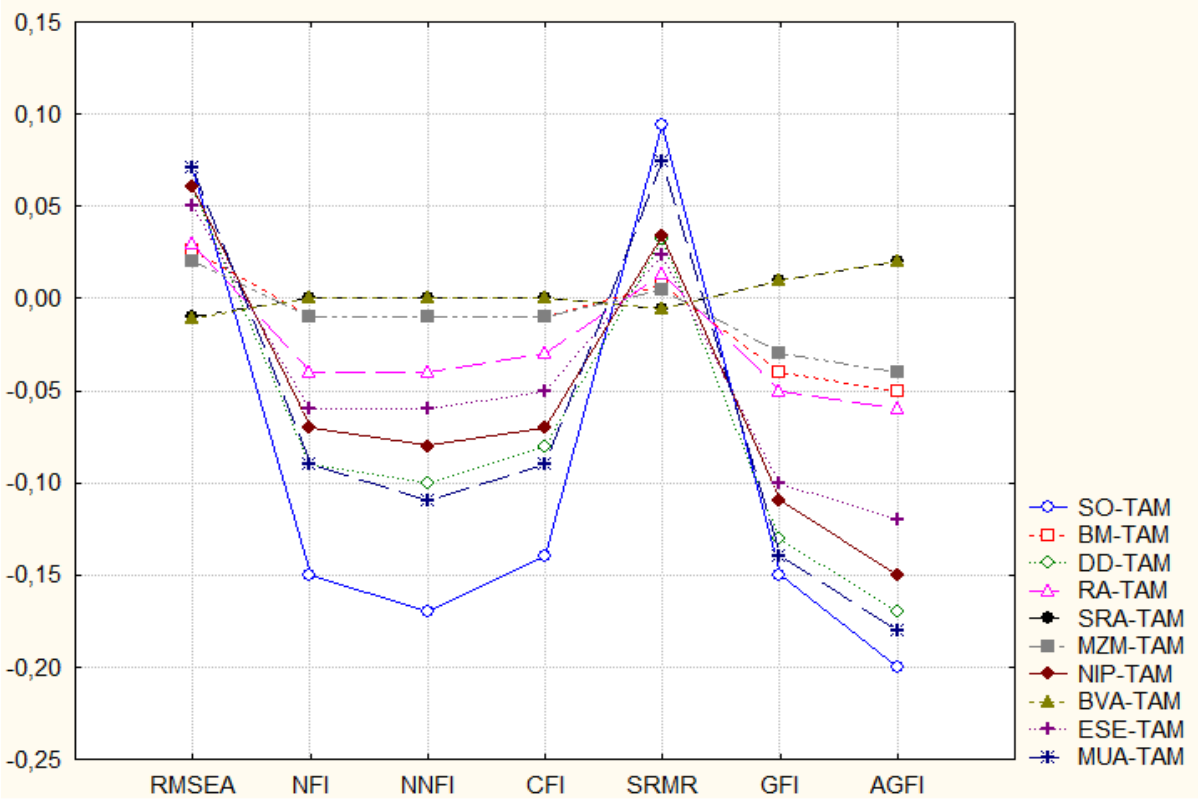
	x^2	RMSEA	NFI	NNFI	CFI	SRMR	GFI	AGFI
Tam veri	673,44	0,049	0,99	0,99	0,99	0,046	0,96	0,94
SO	3580,29	0,12	0,84	0,82	0,85	0,14	0,81	0,74
BM	1388,09	0,075	0,98	0,98	0,98	0,053	0,92	0,89
DD	3039,60	0,11	0,90	0,89	0,91	0,078	0,83	0,77
RA	1541,28	0,079	0,95	0,95	0,96	0,059	0,91	0,88
SRA	481,08	0,039	0,99	0,99	0,99	0,040	0,97	0,96
MZM	1209,81	0,069	0,98	0,98	0,98	0,051	0,93	0,90
NIP	2583,05	0,11	0,92	0,91	0,92	0,080	0,85	0,79
BVA	468,72	0,038	0,99	0,99	0,99	0,040	0,97	0,96
ESE	2353,91	0,099	0,93	0,93	0,94	0,070	0,86	0,82
MUA	3230,74	0,12	0,90	0,88	0,90	0,12	0,82	0,76

Tablo 5 incelendiğinde, kayıp verilerin tamamıyla rastlantısal olarak dağılmadığı durumda yapılan yaklaşık değer atamaları sonrası oluşan bazı veri yapılarının modele uyum

değerlerinin kabul edilebilir uyum sınırının dışarısına çıktığı görülmektedir. Tabloda görüldüğü üzere seriler ortalaması (SO), doğrusal değerleme (DD), NIPALS algoritması (NIP), mahalanobis uzaklığı ataması (MUA) yöntemleri ile yapılan yaklaşık değer atamaları sonucu oluşan veri yapılarının modele uyum değerleri kabul edilebilir değerlerin dışındadır. Tam veri setinden elde edilen model veri uyum değerlerine en yakın değerlerin ise stokastik regresyon ataması (SRA) sonrası elde edilen veri setinden elde edildiği belirlenmiştir. Yaklaşık %20 oranında TROK olmayan veri yerine farklı yöntemler ile yapılan atamalardan sonra oluşan veri yapılarının model veri uyum değerlerinin tam veri setinden elde edilen model veri uyum değerlerine göre ne kadar farklılaştığının daha açık bir şekilde gözlemlenebilmesi amacıyla çoklu çizgi grafiği oluşturulmuş ve elde edilen bulgular Şekil 6'da sunulmuştur.

Şekil 6.

Farklı yöntemlerle %20 oranında TROK olmayan veriye atama yapılması sonucu elde edilen model veri uyum değerleri ile tam veri seti model veri uyum değerleri arasındaki farklar



Şekil 6 incelendiğinde, kayıp veriler yerine SRA, BVA ve MZM yöntemleri ile yapılan değer atamaları sonrası oluşan veri yapılarının modele uyum değerlerinin tam veri setinin modele uyum değerlerinden en az farklılaşan değerler olduğu görülmektedir. Ayrıca farklı yöntemlerle TROK verilere yapılan atamalar sonrası elde edilen model veri uyum değerlerinin aksine TROK olmayan verilere yapılan atamalar sonrası elde edilen uyum

değerlerinin tam veri setinin modele uyum değerlerine göre daha fazla farklılaştığı görülmektedir. En fazla farklılaşmanın AGFI uyum indeksinde meydana geldiği ve 0,20 düzeyinde olduğu belirlenmiştir. Tamamıyla rastlantısal olarak dağılmayan kayıp veriler yerine yapılan atamalar sonrası matematik öz yeterliği toplam puan dağılımları incelenmiş ve *EK 4*'te sunulmuştur. Yapılan incelemeler sonrasında, tamamıyla rastlantısal olarak dağılmayan kayıp verilere yapılan atamalar sonrası elde edilen matematik öz yeterliği toplam puan dağılımlarının tamamıyla rastlantısal olarak dağılan kayıp verilere yapılan atamalar sonrası oluşan dağılımlardan büyük ölçüde farklılaştığı görülmektedir. Özellikle bayesci veri atama (BVA), stokastik regresyonla atama (SRA) ve regresyonla atama (RA) yöntemleri ile yapılan yaklaşık değer atamaları sonucu elde edilen matematik öz yeterliği toplam puan dağılımlarının tam veri setindeki dağılıma benzer şekilde sola çarpık olduğu görülmektedir. Elde edilen bulgular, verilerin tamamıyla rastlantısal dağılmadığı durumlarda BVA ve SRA yöntemleri ile yapılan yaklaşık değer atamalarının diğer yöntemlere kıyasla daha iyi sonuçlar verdiğine bir kanıt olarak sunulabilir.

Elde edilen bulgular toparlanacak olursa verilerin tamamıyla rastlantısal dağıldığı durumlarda kayıp veriler yerine farklı yaklaşık değer atama yöntemleri ile atama yapılması sonucu oluşan veri yapılarının modele uyum değerlerinin tam veri setinin modele uyum değerlerine göre çok büyük farklılıklar göstermediği elde edilen tüm veri setlerinin modele uyum değerlerinin kabul edilebilir sınırlar içerisinde olduğu belirlenmiştir. Hem yaklaşık %20 hem de yaklaşık %30 oranlarında tamamıyla rastlantısal dağılan kayıp veri içeren veri setlerine regresyonla değer atanması (RA) sonucu oluşan veri yapılarından tam veri setinin modele uyum değerlerine en yakın uyum değerlerinin elde edildiği belirlenmiştir. Diğer taraftan verilerin tamamıyla rastlantısal dağılmadığı durumda ise RA ile yapılan değer ataması sonucu elde edilen veri yapısının modele uyum değerlerinin tam veri setinin modele uyum değerlerinden uzaklaştığı görülmüştür. Ayrıca TROK olmayan veriler yerine seriler ortalaması (SO), doğrusal değerleme (DD) gibi basit atamaya dayalı yöntemlerle değer atanması sonucu oluşan veri yapılarının modele uyum değerlerinin tam veri setinin modele uyum değerlerinden önemli derecede farklılıklar gösterdiği görülmüş; stokastik regresyon ataması (SRA), bayesci veri atama (BVA) ve Markov Zincirleri Monte Carlo (MZM) yöntemleriyle yapılan değer atamalarından tam veri setine daha yakın uyum değerlerinin elde edildiği belirlenmiştir.

Tartışma ve Sonuç

Bu çalışma, PISA 2012'ye Türkiye'den katılan 15 yaş grubundaki 4848 öğrenci arasından seçilen 1578 öğrenciden elde edilen puanların oluşturduğu veri seti üzerinde

yürütülmüştür. Tam veri seti üzerinden eksiltmeler yapılarak tamamıyla rastlantısal olarak dağılan yaklaşık %20 ve %30 oranlarında ve tamamıyla rastlantısal olarak dağılmayan yaklaşık %20 oranında kayıp verilerin oluşturulduğu çalışmada, 10 farklı yöntemle yapılan yaklaşık değer atamalarının model veri uyum değerlerini nasıl etkilediği incelenmiştir. Kayıp verilerin tamamıyla rastlantısal olarak dağıldığı durumlarda regresyonla atama (RA) yöntemi sonrası elde edilen veri yapısının modele uyum değerlerinin tam veri setinin modele uyum değerlerine en yakın değerler olduğu sonucuna ulaşılmıştır. Kayıp verilerin tamamıyla rastlantısal olarak dağıldığı durumlarda kayıp veriler yerine farklı yöntemlerle yapılan yaklaşık değer atamaları sonucu oluşan modele uyum değerlerinin tam veri setinin modele uyum değerlerinden önemli farklılıklar göstermediği belirlenmiştir. Yaklaşık %20 oranında TROK veriler yerine yapılan atamalar sonucu oluşan veri yapılarının modele uyum değerleri içerisinde en yüksek farklılaşma NIPALS algoritması (NIP) ile atamaların yapıldığı veri setinin AGFI uyum indeksinde 0,04 düzeyinde olduğu belirlenmiştir. Yaklaşık %30 oranında TROK veriler yerine yapılan atamalar sonucu oluşan veri yapılarının modele uyum değerleri içerisinde en yüksek farklılaşma eğilim skorları eşitlemesi (ESE) ve mahalanobis uzaklığı ataması (MUA) yöntemleri ile atamaların yapıldığı veri setlerinin AGFI uyum indekslerinde 0.07 düzeyindedir. Kayıp verilerin (yaklaşık %20) tamamıyla rastlantısal olarak dağılmadığı durumda ise atamalar sonrası elde edilen modele uyum değerlerindeki farklılaşmalar önemli derece artmıştır. En fazla farklılaşmanın seriler ortalaması (SO) ile değer ataması yapılan veri setinin AGFI uyum indeksinde 0.20 düzeyinde olduğu belirlenmiştir.

Elde edilen bulgulardan hareketle kayıp veri dağılımının yapılacak olan analizlerin doğruluğu üzerinde önemli bir etken olduğu ve kayıp verilerin tamamıyla rastlantısal dağılmadığı durumlarla başa çıkmanın daha fazla alan bilgisi gerektirdiği sonucuna ulaşılabilir. Ayrıca veri setinde TROK olmayan verilerin bulunması durumunda seriler ortalaması (SO) ve doğrusal değerleme (DD) gibi basit atamaya dayalı yöntemlerin iyi sonuçlar vermediği, bayesci veri atama (BVA) ve stokastik regresyonla atama (SRA) yöntemleri ile yapılan değer atamalarıyla tam veri setinin modele uyum değerlerine en yakın değerlerin elde edildiği belirlenmiştir. Yapılan ek incelemelerde kayıp veriler yerine yaklaşık değer atama yöntemlerinin veri dağılımlarını önemli ölçüde değiştirdiği belirlenmiş ve bilinçsizce yapılan atamaların daha sonra yapılacak olan analizlerin sonuçlarını araştırmacıları yanıltacak şekilde etkileyebileceği sonucuna ulaşılmıştır. BVA ve SRA gibi yöntemlerle yapılan değer atamalarının model veri uyum değerlerini artırıcı yönde etkilerinin olduğu göz ardı edilmemelidir. Farklı araştırmalar kapsamında söz konusu yöntemler modele iyi uyum sağlamayan veri yapıları üzerinde de denenerek etkileri daha derinlemesine incelenebilir. Bu

çalışmada, farklı yöntemlerle değer atamaları sonucu oluşturulan veri yapılarının yapısal eşitlik modeli ile test edilmesi sonucu elde edilen model veri uyum değerlerinin analiz sonuçlarını etkileyebilecek ölçüde birbirlerinden farklılaştığı belirlenmiştir. Bu bulgudan hareketle kullanılan kayıp veriler ile başa çıkma yönteminin bilimsel araştırmaların sonuçları için önemli bir etken olduğu sonucuna varılabilir. Bu nedenle araştırmacıların özellikle fazla miktarlarda kayıp veri içeren veri setleri üzerinde yaptıkları araştırmalarda kayıp verilerin miktarını, rastlantısal olarak dağılıp dağılmadığını ve kayıp verilerle başa çıkmak için kullandıkları yöntemi raporlamaları araştırmalarının tekrarlanabilirliği bakımından büyük önem taşımaktadır. Bu araştırma, PISA 2012 uygulaması içerisinde yer alan matematik öz yeterliği, matematik kavramlarına aşinalık ve matematik okuryazarlığı yapıları ile oluşturulmuş bir yapısal eşitlik modeli üzerinde yürütülmüştür. Kayıp değerler yerine yaklaşık değer atama yöntemleri, farklı veriler kullanılarak farklı istatistiksel analiz türleri üzerinden karşılaştırılabilir.

Kaynaklar

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.

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.*

Demir, E. (2013). Kayıp verilerin varlığında çoktan seçmeli testlerde madde ve test parametrelerinin kestirilmesi: SBS örneği. *Eğitim Bilimleri Araştırmaları Dergisi, 3(2), 48-68.*

Garrett, P. L., (2009). A monte carlo study investigating missing data, differential item functioning and effect size. Retrieved on 06. 01. 2015, at URL: http://scholarworks.gsu.edu/eps_diss/35

Köse, İ. A. (2014). The effect of missing data handling methods on goodness of fit indices in confirmatory factor analysis. *Educational Resarch 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 Eğitim Fakültesi Dergisi, 14(1), 400-412.*

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

OECD (2014). PISA 2012 technical report. Retrieved on 06. 01. 2015, at URL: http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

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.*

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

Witta, E. L. (2000). *Comparison of missing data treatments in producing factor scores (rapor).* Annual Meeting of the American Educational Resarch Association. Honolulu.

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

Extended Abstract

The general purpose of this study is to comparatively research on impacts on data fit statistics which are obtained from structural equation model that is formed by latent variables, data sets including missing data in different pattern structure and at a different rate, value imputation implementing by using different methods. In accordance with this purpose, this study has been conducted on data set formed by points of 1578 students who were chosen among 4848 students- age group of 15-who participated in PISA 2012 from Turkey. In the study firstly, it is set a theoretical model with the assumption on mathematics self-efficacy and familiarity with mathematics concepts effect literacy mathematics. After the model having been tested on full data set, lost data which are totally random dispersed about in the ratios of 20% and 30% and are not totally random dispersed about in the ratios of %20 were created by subtracting in the data set. After determined steps, approximate values imputation were conducted in 10 various methods (series mean, expectation maximization algorithm, linear interpolation, regression, stochastic regression, MCMC method, NIPALS algorithm, Bayesian imputation, propensity score matching, mahalanobis distance imputation) rather than missing data in data sets including missing data in different pattern structure and different ratios and each obtained data structures fitting to structural equation model was interpreted by comparing obtained values from full data set to obtained values which were tested.

Results

In the case of missing data dispersing in totally random, it is determined that model-dat fit statistics that are created by approximate value imputation in different methods instead of missing data do not significantly differ from full data set model fit statistics. In the case of missing data dispersing in totally random, it is also concluded that data structure of model-data fit statistics which is obtained after regression imputation method are the most proximate values to full data set of model fit statistics. In the cases of lost data not dispersing totally random, model-data fit statistics which are obtained after imputation have dramatically changed. In this cases, it is determined that methods based on simple imputation methods such as series mean and linear interpolation do not work, the approximate values of full data set to model fit statistics are obtained by value imputations with stochastic regression imputation methods and bayesian value imputation. In the cases of missing data both dispersing in totally random and not dispersing totally in random, it is seen that differentiation mostly occurs in AGFI fit index after values imputation are conducted. In the additional reviews, it is concluded that approximate value imputation have changed the data distribution considerably compared with complete data.

Discussion

From the received facts, it is claimed that missing data dispersion has an important impact on accuracy of analysis and it requires more information to overcome in the cases of missing data not dispersing in totally random. It should be taken into consideration that unaware imputations mislead researchers with analysis results to be conducted later as approximate value imputation methods rather than missing data have significantly changed data distribution. It is determined that value imputations conducted by methods such as Bayesian data imputation and stochastic regression effect model-data fit statistics in an increasing way. Methods in question are carefully dealt with particularly in the case of missing data dispersing in totally random.

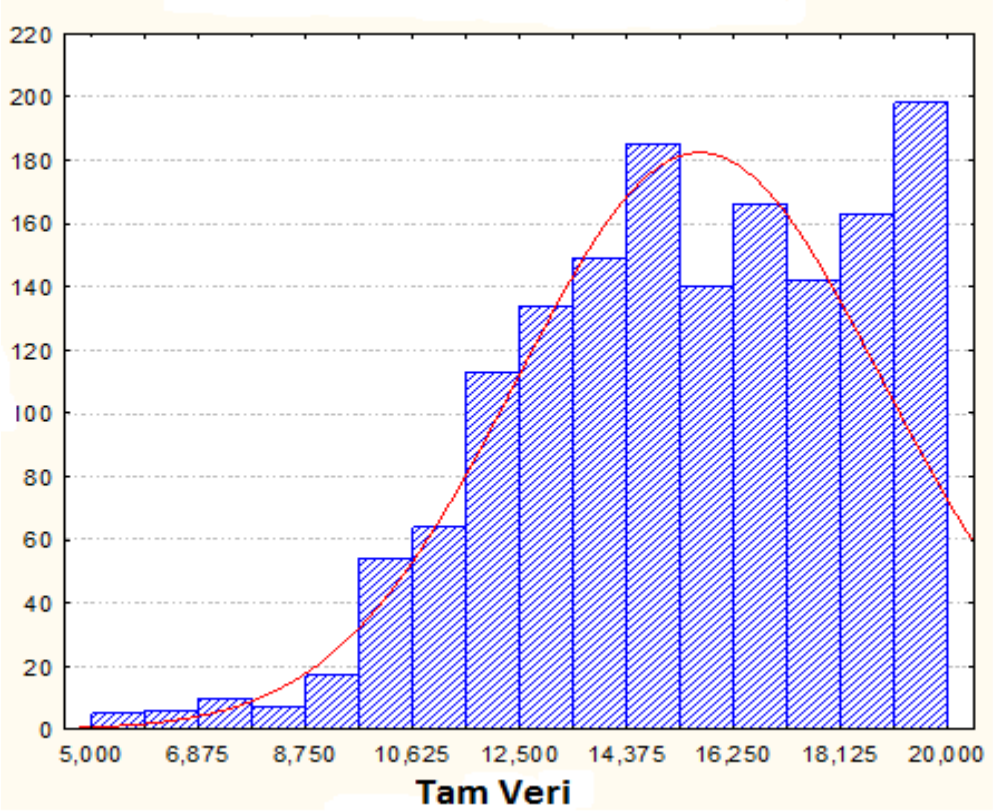
Conclusion

In this study, it is concluded that model-data fit statistics which are obtained by testing data structure with structural equation model formed by value imputations in different methods have differentiated from each other resulting in possibly effects analysis results. From this fact, it might be deduced that overcoming methods of missing data are the most significant factor for results of scientific researches. Therefore, in the studies conducted by researchers on particularly including high quantity of missing data set, quantity of lost data, whether missing data disperse at random or not, and reporting the method used for handling

missing data are highly crucial in terms of repeatability of researches. This research was conducted on a structural equation model which had been formed mathematics literacy, familiarity with mathematics concepts and mathematics self-efficacy within PISA 2012 Application. The methods of imputation of approximate values instead of missing values might be compared on different kind of statistical analysis by using different data. It might be claimed that the numbers of current studies are insufficient to overcome missing data in terms of bringing strong recommendations. It might be foreseen that the increasing number of studies to be conducted in this area will be a significant for overcoming problems which threaten the accuracy of analysis results.

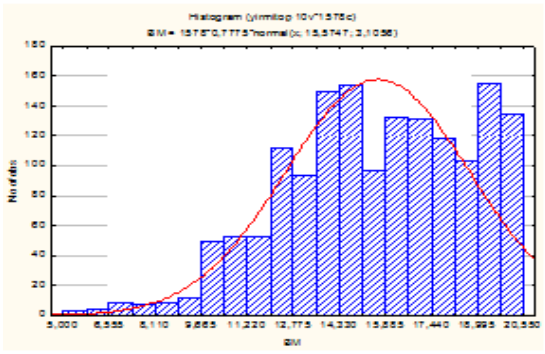
EK 1.

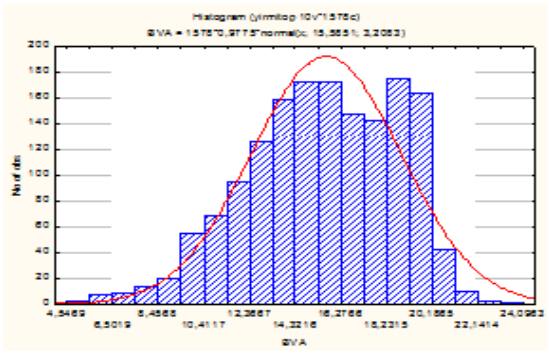
Tam Veri Seti İçin Matematik Öz Yeterliği Toplam Puan Dağılımı

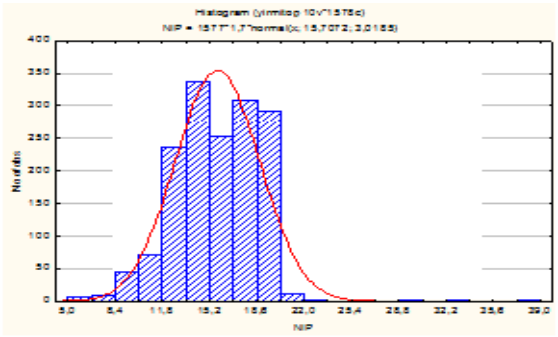


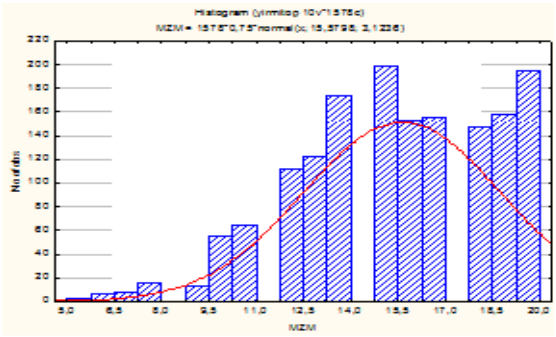
EK 2.

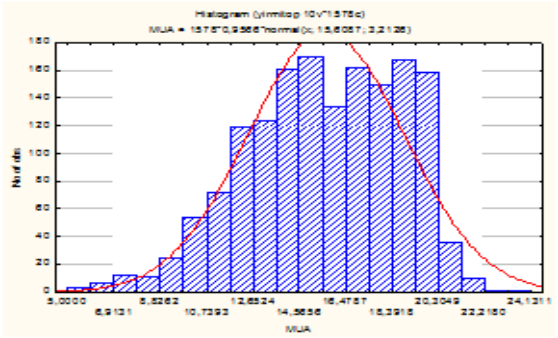
Farklı Yöntemlerle %20 Oranında Trok Veriye Atama Yapılması Sonucu Matematik Öz Yeterliği Toplam Puanlarından Elde Edilen Dağılımlar

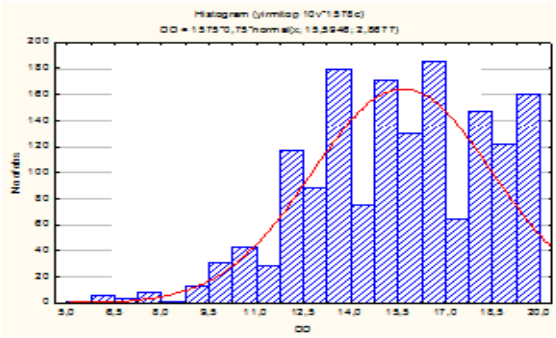

Histogram (yirmiüç 10v*1575c)
SİM = 1575*0,7775*normal(x; 15,5747; 3,1056)
Nordidis
SİM

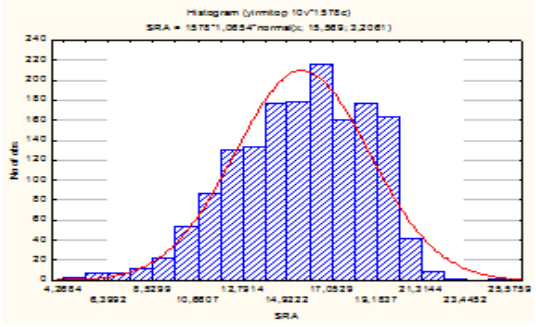

Histogram (yirmiüç 10v*1575c)
SİVA = 1575*0,9775*normal(x; 15,3551; 3,2053)
Nordidis
SİVA

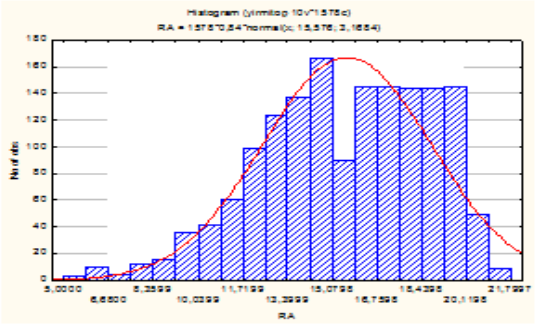

Histogram (yirmiüç 10v*1575c)
NİP = 1577*1,7*normal(x; 15,7072; 3,0155)
Nordidis
NİP

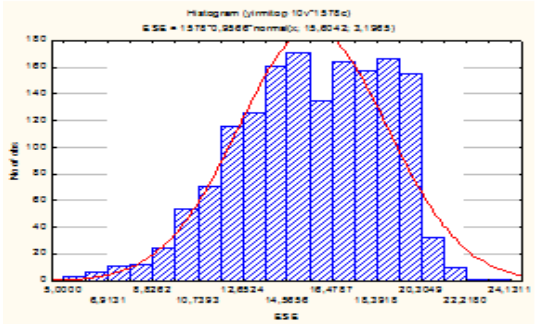

Histogram (yirmiüç 10v*1575c)
MZM = 1575*0,75*normal(x; 15,5795; 3,1236)
Nordidis
MZM

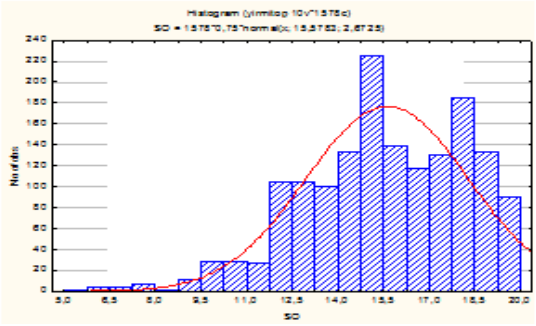

Histogram (yirmiüç 10v*1575c)
MUA = 1575*0,9566*normal(x; 15,6057; 3,2126)
Nordidis
MUA


Histogram (yirmiüç 10v*1575c)
OO = 1575*0,75*normal(x; 15,5946; 2,8677)
Nordidis
OO


Histogram (yirmiüç 10v*1575c)
SRA = 1575*1,0654*normal(x; 15,569; 3,2061)
Nordidis
SRA

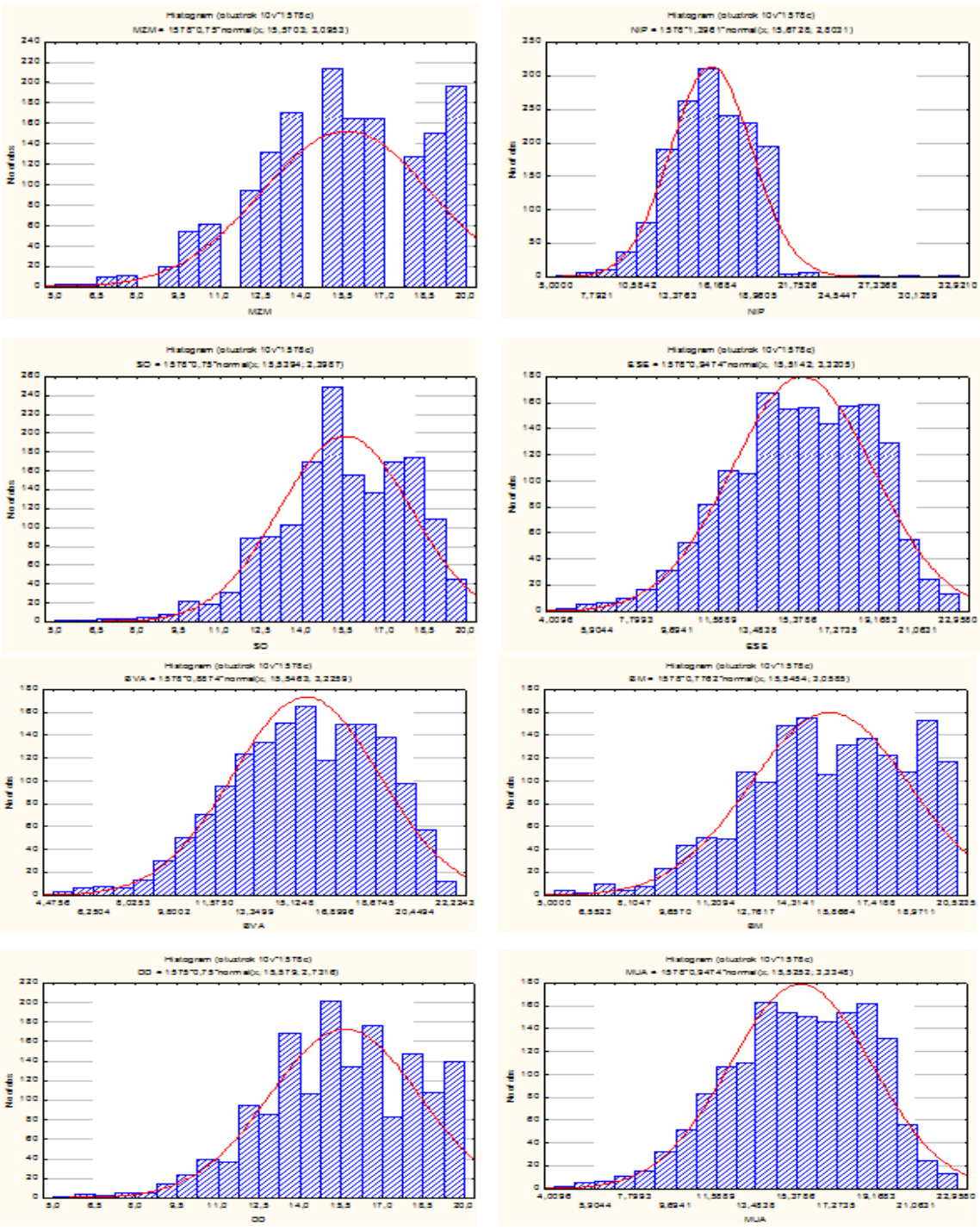

Histogram (yirmiüç 10v*1575c)
RA = 1575*0,64*normal(x; 15,576; 3,1654)
Nordidis
RA

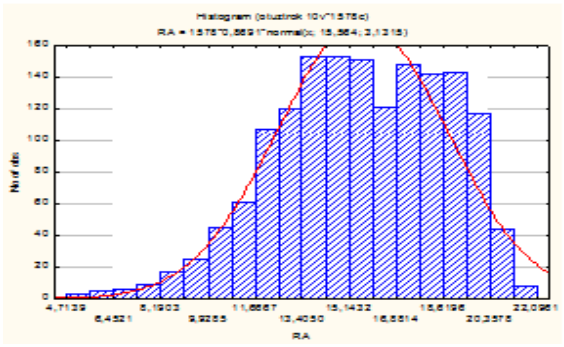

Histogram (yirmiüç 10v*1575c)
SSS = 1575*0,9566*normal(x; 15,6042; 3,1965)
Nordidis
SSS

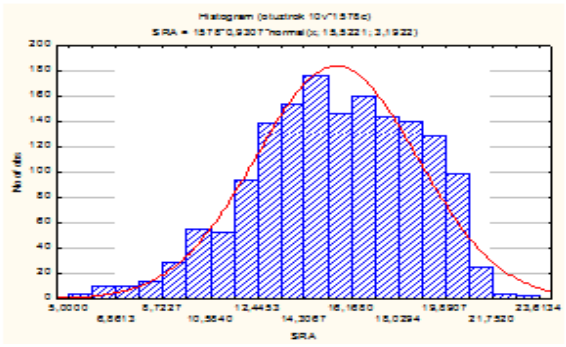

Histogram (yirmiüç 10v*1575c)
SO = 1575*0,75*normal(x; 15,5783; 2,8725)
Nordidis
SO

EK 3.

Farklı yöntemlerle %30 oranında TROK veriye atama yapılması sonucu matematik öz yeterliği toplam puanlarından elde edilen dağılımlar







EK 4.

Farklı Yöntemlerle %20 Oranında Trok Olmayan Veriye Atama Yapılması Sonucu Matematik Öz Yeterliği Toplam Puanlarından Elde Edilen Dağılımlar