T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
İSTATİSTİK ANA BİLİM DALI

SAYI İLE İFADE EDİLEN ZAMAN SERİLERİNDE KAYIP GÖZLEM ANALİZİ: TRAFİK KAZASI ÖRNEĞİ

Yüksek Lisans Tezi

Furkan KOÇAL

Danışman
Prof. Dr. Mehmet Ali CENGİZ

SAMSUN
2021

T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
İSTATİSTİK ANA BİLİM DALI

SAYI İLE İFADE EDİLEN ZAMAN SERİLERİNDE KAYIP GÖZLEM ANALİZİ: TRAFİK KAZASI ÖRNEĞİ

Yüksek Lisans Tezi

Furkan KOÇAL

Danışman

Prof. Dr. Mehmet Ali CENGİZ

SAMSUN
2021

TEZ KABUL VE ONAYI

Furkan KOÇAL tarafından, **Prof. Dr. Mehmet Ali CENGİZ** danışmanlığında hazırlanan **"Sayı İle İfade Edilen Zaman Serilerinde Kayıp Gözlem Analizi: Trafik Kazası Örneği"** başlıklı bu çalışma, jürimiz tarafından 14.7.2021 tarihinde yapılan sınav sonucunda oy birliği ile başarılı bulunarak Yüksek Lisans Tezi olarak kabul edilmiştir.

	Unvanı Adı Soyadı **Üniversitesi** **Ana Bilim/Ana Sanat Dalı**	**İmza**	**Sonuç**
Başkan	Doç. Dr. Talat ŞENEL Ondokuz Mayıs Üniversitesi İstatistik Anabilim Dalı		☐ Kabul ☐ Ret
Üye **(Danışman)**	Prof. Dr. Mehmet Ali CENGİZ Ondokuz Mayıs Üniversitesi İstatistik Anabilim Dalı		☐ Kabul ☐ Ret
Üye	Doç. Dr. Tolga ZAMAN Çankırı Karatekin Üniversitesi İstatistik Anabilim Dalı		☐ Kabul ☐ Ret

Bu tez, Enstitü Yönetim Kurulunca belirlenen ve yukarıda adları yazılı jüri üyeleri tarafından uygun görülmüştür.

ONAY
… / … / …
Prof. Dr. Ali BOLAT
Enstitü Müdürü

BİLİMSEL ETİĞE UYGUNLUK BEYANI

Hazırladığım Yüksek Lisans tezinin bütün aşamalarında bilimsel etiğe ve akademik kurallara riayet ettiğimi, çalışmada doğrudan veya dolaylı olarak kullandığım her alıntıya kaynak gösterdiğimi ve yararlandığım eserlerin Kaynaklar'da gösterilenlerden oluştuğunu, her unsurun enstitü yazım kılavuzuna uygun yazıldığını ve TÜBİTAK Araştırma ve Yayın Etiği Kurulu Yönetmeliği'nin 3. bölüm 9. maddesinde belirtilen durumlara aykırı davranılmadığını taahhüt ve beyan ederim.

İmza
24 /06 / 2021
Furkan KOÇAL

TEZ ÇALIŞMASI ÖZGÜNLÜK RAPORU BEYANI

Tez Başlığı: Sayı İle İfade Edilen Zaman Serilerinde Kayıp Gözlem Analizi: Trafik Kazası Örneği

Yukarıda başlığı belirtilen tez çalışması için şahsım tarafından 24.06.2021 tarihinde intihal tespit programından alınmış olan özgünlük raporu sonucunda;

Benzerlik oranı : % 14
Tek kaynak oranı : % 4 çıkmıştır.

İmza
24/06/ 2021
Prof. Dr. Mehmet Ali CENGİZ

ÖZET

SAYI İLE İFADE EDİLEN ZAMAN SERİLERİNDE KAYIP GÖZLEM ANALİZİ: TRAFİK KAZASI ÖRNEĞİ
Furkan KOÇAL
Ondokuz Mayıs Üniversitesi
Lisansüstü Eğitim Enstitüsü
İstatistik Anabilim Dalı
Yüksek Lisans, Haziran/2021
Danışman: Prof. Dr. Mehmet Ali CENGİZ

Sayım verisi bir olayın sayısını veya olayın meydana gelme sayısını ifade eder. Bu sayılar negatif olmayan tam sayı değerli değişkenlerlerden oluşur. Bu tarz değişkenleri incelerken sıklıkla kullanılan yöntemlerden birisi Poisson zaman serisidir. Bu çalışma Poisson zaman serisinde kayıp verinin yarattığı etkiyi farklı kayıp veri doldurma yöntemleri ile incelemektedir. Bu amaçla Poisson zaman serisine uygun gerçek bir veri seti üzerinde inceleme yapılmıştır. Çalışmada kayıp veri doldurma yöntemlerinden çoklu doldurma, çoklu doldurma zincir denklemi, interpolasyon, son gözlemi ileriye taşıma, Kalman ve hareketli ortalama yöntemleri ele alınmıştır. Bu yöntemlerin alt yöntemleri ile birlikte toplamda sekiz farklı kayıp veri doldurma yöntemi kullanılmıştır. Veri setinde kayıp veriler oluşturularak bu yöntemler ile doldurulmuş ve yeni veri setleri elde edilmiştir. Elde edilen sekiz farklı veri seti üzerinde Poisson zaman serisi modeli oluşturulmuş ve tam veri seti ile elde edilen katsayılar ile kıyaslama yapılmıştır. Sonuç olarak en başarılı çoklu kayıp veri doldurma yönteminin Kalman yöntemi olduğu, en kötü yöntemin ise MICE midastouch yöntemi olduğu görülmüştür.

Anahtar Sözcükler: Poisson Zaman Serisi, MI, MICE, Enterpolasyon Yöntemi, Kayıp Veri Analizi, MA, Kalman

ABSTRACT

MISSING CASE ANALYSİS İN TIME SERIES WITH COUNT DATA: TRAFFIC ACCIDENT

Furkan KOÇAL
Ondokuz Mayıs University
Institute of Graduate Studies
Department of Statistics
Master, July/2021
Supervisor: Prof. Dr. Mehmet Ali CENGİZ

Census data refers to the number of events or the number of times an event occurred. These numbers consist of non-negative integer-valued variables. Poisson time series is one of the frequently used methods when examining such variables. This study examines the effect of missing data in Poisson time series with different missing data imputation methods. For this purpose, a real data set suitable for Poisson time series was examined. In the study, missing data imputation methods such as multiple imputation, multiple imputation chain equations, interpolation, last observation carried forward, Kalman and moving average methods are discussed. A total of eight different missing data imputation methods were used together with the sub-methods of these methods. Missing datasets are drawed from data set and new full datasets are obtained by imputing using selected imputation methods. The Poisson time series model was created on the eight different data sets obtained and the comparison was made with the coefficients obtained with the full data set. As a result, it was concluded that the most successful multiple loss data filling method was Kalman method, and the worst method was MICE midastouch method.

Keywords: Poisson Time Series, MI, MICE, Interpolation Method, Lost Data Analysis, MA, Kalman

ÖNSÖZ VE TEŞEKKÜR

Akademik eğitim sürecimin bir üst noktası olan yüksek lisans tez çalışmam boyunca yardım ve desteğini benden esirgemeyen danışman hocam Prof. Dr. Mehmet Ali CENGİZ'e ve yardımları dolayısıyla Arş. Gör. Fatih SAĞLAM'a teşekkürü borç bilirim.

Her konuda fedakarlıktan kaçınmayan ve desteğini esirgemeyen değerli eşim Büşra KOÇAL ve oğlum Alperen Tuna KOÇAL'a sonsuz teşekkürler.

Ağustos, 2021
Furkan KOÇAL

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

ARIMA	Otomatik Regresif Entegre Hareketli Ortalama
EM	Beklenti Maksimasyonu
EMA	Üstel Hareketli Ortalama
EWMA	Üstel Ağırlıklı Hareketli Ortalama
GLM	Genelleştirilmiş doğrusal model
İNTERPOLATİON	Enterpolasyon ile Imputation
KALMAN	Kalman Yumuşatma ve Durum Uzayı Modellerinden İfade
LOCF	Son Gözlemi İleri Taşıma
LQG	İneer Kuadratik Gauss Kontrol Problemini
LQR	Lineer Kuadratik Regülatör
MA	Ağırlıklı Hareketli Ortalamaya göre Hesaplama
MCAR	Çoklu Rastgele Atama Yöntemi
MI	Çoklu İfade
MICE	Çoklu İfade Zincirli Denklemler
MNAR	Rastgele Eksik Olmayan
NMAR	Rastgele Rksik Olmayan
NOCB	Sonraki Gözlemi Geriye Doğru Taşıma
PMM	Tahmini Ortalama Eşleştirme Yöntemi, Predictive Mean Matching
SEKK	Sıradan en küçük kareler
SMA	Basit Hareketli Ortalama
WMA	Ağırlıklı Hareketli Ortalama

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

1. GİRİŞ

Günümüzde sayım zaman serileri için regresyon modellerinden artan bir ilgi söz konusudur. Artan bu ilgi, bakıldığında literatürde önemli sayıda yayın olarak yansımaktadır. Belirli zaman aralığında meydana gelen olayların sayısal olarak ifade edilen bu tür zaman serilerinde, günlük sağlık kuruluşlarına kabul edilen hastaların sayıları, belli zaman içine düşen borsa işlem hacmi, üretim tesislerinde oluşan hatalı ürün sayıları, günümüzde oldukça güncel olan Covid-19 virüsünün günlük enfekte sa–yıları veya ölüm sayıları örnek verilebilecek alanlardır. Bu tür zaman serilerinin modellenmesi, gözlemlerin negatif olmayan tamsayılar olduğu ve gözlemler arasındaki bağımlılığın göz önüne alınması gerektiği için oldukça zor ve karmaşıktır. Zaman serisi analizinde gözlemlerin birbirlerine eşit aralıklarla elde edilmesi gerekmektedir. Bu nedenle çoğu zaman serisi analizinde eksik gözlem sorunu yaşanabilmektedir. Böyle bir sorunla karşılaşmamak için verilerin değişmeyen bir sistem içinde ya da büyük bir özenle gözlemlenmesi gerekmektedir. Eksik gözlemlerle zaman serisi analizinin yapılması özellikle yorum bakımından birçok sorunlara neden olduğundan, bu eksik verilerin uygun yöntemlerle tahmin edilmesi gereklidir.

Verilerin sağlıklı analizi ve yorumlanabilmesi için en uygun metotla tahmin ederek eksikliğin giderilmesi çok önemlidir. Çoğu zaman veriler önümüze gizli kayıp verilerle gelmektedir. Genelde veriler analize başlanmadan önce birkaç eksik gözlem varsa bile görmezden gelerek veya o eksik kısımları sistemden çıkartarak analizleri yapılmaktadır. Bu iş, temelinde yanlış bir yöntemdir. Eksik gözlemlerin tamamlanması konusunda, Benmouzi ve Cheknane (2016)'da ARIMA (Otoregresif Entegre Hareketli Ortalama) modelleri kullanılarak bir günlük bir sürecin saatlik tahminlerilerini üretirken eksik gözlemler tamamlanmıştır. Hussain ve Al Alili (2016)'da saatlik bazda küresel yatay güneş ışınımının verilerindeki eksiklik için de ARIMA modelini kullanmıştır.

Son gözlemi ileriye taşıyarak doldurma (LOCF), eksik gözlemin giderilmesi için kullanılan yaygın bir yöntemdir. Bu yöntemin en çok tercih edilme nedeniyse, hesaplamasının kolaylığıdır. LOCF, veri sayısı yeterince büyükse ve eksik veri sayısı da oldukça küçükse, son veriye bakarak, bir sonraki eksik veriyi tahmin etme yöntemidir. ABD Gıda ve İlaç İdaresi'nin, Psikiyatrik İlaç Ürünleri Bölümü, geleneksel olarak LOCF'yi en çok tercih ettikleri yöntem olarak açıklamaktadır. Muhtemelen veri büyüklüğü açısından ve eksik gözlem azlığından dolayı bu yöntemi

önermektedirler. ABD Gıda ve İlaç İdaresi birincil analiz yöntemi olarak LOCF kullandıktan sonra, bu yöntemleri psikiyatri dergilerinde tercih edilen bir yöntem haline gelmiştir. Araştırmacılar ve yayıncı hakemleri bu konuda olumsuz bir düşünce paylaşmamıştır (Hamer ve Simpon, 2009).

Eksik gözlem tahmini en yaygın olarak tıp alanında ihtiyaç duyulmaktadır. Çünkü genelde tıp alanında toplanan veriler, başta deneğin vefatı, yer değiştirmesi, bildirmeden ayrılması vb. gibi nedenlerden dolayı eksik kalabilmektedir. Tıp alanında yapılan bazı yayınlarda eksik gözlem tahmini kullanan bazı yayınları şu şekilde açılabiliriz: Klinik araştırmalarda Wood, White ve Thompson (2004), Powney ve diğerleri (2014), Diaz-Ordaz ve diğerleri (2014), Kanser araştırmalarında Burton ve Altman (2004) , Eğitim Araştırmalarında Peugh ve Enders (2004), Epidemiyoloji Araştırmalarında Klebanoff ve Cole (2008), Karahalios ve vd. (2012), Gelişim Psikolojisi alanında (Jeliĉić, Phelps ve Lerner, 2009) ve Genel Tıp alanında Mackinnon (2010) çeşitli yöntemler kullanmışlardır.

Eksik gözlemler neredeyse her alanda oluşturulan verilerde karşımıza çıkabilmektedir. Bu alanlardan birisi de, nüfus istatistiklerinde olmaktadır. Fritz Scheuren (Scheuren, 2005) ABD'de 2015 yılında gerçekleşen nüfus sayımında Mart gelir ekindeki eksik verilerin giderilmesi için çoklu kayıp veri doldurma yöntemini uygulayarak çözmüştür. ABD Sosyal Güvenlik İdaresi bu yöntemi günümüzde halen kullanmaktır (Scheuren, 2005).

Uygulama alanında sayım zaman serilerinde kayıp gözlem analizi daha karmaşık bir yöntemdir. Bu çalışmada, farklı dağılımlı bu tür zaman serilerinde kayıp gözlem yöntemleri incelenerek gerçek veri setleri üzerinde en uygun yöntemin hangisi olduğuna karar verilecektir.

2. SAYIM VERİSİ

Bir olayın sayısı, bir olayın meydana gelme sayısını ifade eder. Bu sayı, negatif olmayan tam sayı değerli bir rastgele değişkenin gerçekleşmesidir. Olay sayımlarının tek değişkenli bir istatistiksel modeli, olayın meydana gelme sayısının olasılık dağılımını belirtir. Bu dağılımın, genellikle bazı parametreleri bilinmiyor olabilir. Bu tür modellerde tahmin ve çıkarım, bilinmeyen parametrelerle ilgilidir. Böyle bir spesifikasyon başka değişken içermez ve olay sayısının bağımsız olarak aynı dağılıma sahip oldukları varsayılır.

Poisson dağılımı, Poisson (1837) tarafından iki terimli bir sınırlayıcı durum olarak türetilmiştir. İlk uygulamalar arasında, Prusya ordusunda katırlar tarafından tekmelenmeden kaynaklanan yıllık ölüm sayılarının klasik Bortkiewicz (1898) çalışması yer almaktadır. Poisson'un standart bir genellemesi, negatif binom dağılımıdır. Greenwood ve Yule (1920) tarafından, gözlemlenmemiş heterojenlikten kaynaklanan görünür bulaşıcılığın bir sonucu olarak ve gerçek bulaşıcılığın bir sonucu olarak Eggenberger ve Polya (1923) tarafından türetilmiştir.

Sayım modellerinin başlıca kullanılan alanları ekonomi, siyaset bilimi, sosyoloji aktüerya bilimi, demografide ve biyoistatistik yayın olarak kullanılmaktadır. Verilerin, uygulama alanlarındaki başlıca özellikleri, bu modellerin kapsamını genişletmiş ve gelişmesini sağlamıştır.

Sayım veri regresyon modellerinin geliştirilmesindeki önemli bir dönüm noktası, Poisson regresyonunun özel bir durum olduğu, ilk olarak Nelder ve Wedderburn (1972) tarafından tanımlanan ve McCullagh ve Nelder (1989) tarafından detaylandırılan "genelleştirilmiş doğrusal modellerin" ortaya çıkmasıdır. Bu katkılardan yola çıkarak, Gourieroux, Monfort ve Trognon'un (1984) makaleleri ve Hausman, Hall ve Griliches'in (1984) boylamsal veya panel sayımı veri modelleri üzerine çalışmaları da uygulamalı çalışmayı teşvik etmede çok etkili olmuştur.

Sayım verileri farklı bir soyutlama düzeyinde, bir olayın belirli meydana gelme hızı tarafından oluşan, noktasal sürecin gerçekleşmesi olarak düşünülebilir. Olayların sayısı, belirli bir zaman dilimi boyunca bu tür gerçekleşmelerin toplam sayısı olarak nitelendirilebilir. Olay sayısının ikilisi, olaylar arasındaki sürenin uzunluğu olarak tanımlanan, varışlar arası süredir. Sayım veri regresyonu, bazı ortak değişkenler üzerinde koşullu zaman birimi başına oluşum oranının incelenmesinde yararlıdır. Bunun yerine, ortak değişkenlere bağlı olarak varışlar arası zamanların dağılımı

incelenebilir. Bu durum bekleme sürelerinin regresyon modellerine yol açar. Mevcut veri türü, kesitsel, zaman serileri veya boylamsal, istatistiksel çerçevenin seçimini etkileyecektir. Sayım verilerini işlemek için "özel" yöntemlerin gerekli olup olmadığı veya standart Gauss doğrusal regresyonunun yeterli olup olmayacağıdır. Doğrusal regresyon modelindeki sıradan en küçük kareler gibi daha yaygın regresyon tahmin edicileri ve modelleri, bağımlı değişken için sınırlı desteği göz ardı eder. Bu, sayımların ortalaması yüksek olmadığı sürece, önemli eksikliklere yol açabilir. Bu durumda normal yaklaşım ve ilgili regresyon yöntemleri tatmin edici olabilir.

Sayım verileri artan ilgilin başlıca nedeni, bireysel olarak ekonometrinin farklı yönlerini modellemeye yönelik ilgilin artmasından kaynaklanabilir. Örneğin, Pudney (1989) geniş bir mikroekonometri verisini "köşelerin, kıvrımların ve deliklerin ekonometrisi" olarak tanımlamaktadır. Sayım veri modelleri, belirli ayrık veri regresyon türleridir. Ayrık ve sınırlı bağımlı değişken modeller, ekonometride büyük ilgi çekmiş ve mikroekonometride zengin bir uygulama kümesi bulmuştur (Dempster, 1983). Özellikle ekonometri uzmanları birçok alternatif örnek veri türü ve örnekleme çerçevesi için modeller geliştirmeye çalışmaktadır. Poisson regresyonu birçok analiz için bir başlangıç noktası sağlasa da gözlemi ve veri toplamayı yöneten çok sayıda gerçek yaşam koşulunu barındırma girişimleri, ek ayrıntılara ve komplikasyonlara yol açar. Sayım veri modellerinin kapsamı çok geniştir. Bu ekonomik verilerin özelliklerine özel olarak odaklanarak, sayılar için regresyon modellerinde ortaya çıkan sorunları ele almaktadır. Ancak çoğu durumda, kapsanan materyal, ekonomik verilerin özelliklerini her zaman paylaşmayan sosyal ve doğa bilimlerinde kullanılmak üzere kolayca uyarlanabilir (Cameron ve Trivedi, 2013).

3. ZAMAN SERİSİ

Zaman serisi analizi, verilerin anlamlı istatistiklerini ve diğer özelliklerini çıkarmak için zaman serisi verilerini analiz etmeye yönelik yöntemleri içerir. Zaman serisi tahmini, önceden gözlemlenen değerlere dayalı olarak gelecekteki değerleri tahmin etmek için bir modelin kullanılmasıdır. Regresyon analizi genellikle bir veya daha fazla farklı zaman serisi arasındaki ilişkileri test edecek şekilde kullanılırken, bu tür analize genellikle "zaman serisi analizi" denmez; bu, özellikle tek bir zaman içindeki farklı noktalar arasındaki ilişkileri ifade eder. Bir zaman serisi, istenilen zaman içinde sıralanmış (grafik veya listeli şekilde) bir dizi veri noktalarıdır. En yaygın olarak, bir zaman serisi, kronolojik sıralanan, veri satırları belilirli aralıklarla yenilenen (yıl, ay, gün, saat, dakika vb.) eşit aralıklı noktalarda alınan bir dizidir.

Zaman serileri, zamansal çizgi çizelgesi aracılığıyla çizilir. Zaman serileri istatistik, hava tahmini, sağlık, ekonometri, deprem tahmini, sinyal işleme, mühendislik alanları gibi zamansal ölçümleri içeren uygulamalı bilimlerin herhangi bir alanında kullanılır.

Zaman serisi analizi, verilerin farklı özelliklerini bulmak ve istatistiksel olarak anlamlı analiz gerçekleştirmek için, zaman serilerisi çeşitli analiz etme yöntemlerini barındırır. Zaman serisi tahmini, belirli zaman içerisinde gözlemlenen değerlere bağlı olarak, gelecek zaman içerisindeki değerlerini tahmin etmek için bir modelin kullanılarak analiz edilmesidir. Kesikli zaman serisi analizi, verilerdeki belirgin olmayan ama seriyi etkileyebilecek bazı müdahalelerin, bu müdahale öncesi ve sonrasında zaman serisinde oluşabilecek değişikliklerin tespit etmek için kullanılır.

Zaman serisi analiz yöntemi iki başlığa ayırılabilir: Birincisi frekans alanı yöntemleri ikincisi zaman alanı yöntemleri. Frekans alanı yöntemlerini, spektral analiz ve damlacık (wavelet) analizini içerir. Zaman alanı yöntemleri ise otokorelasyon (öz ilinti) ve çapraz korelasyon zaman seri analizi türlerini içermektedir.

Zaman serileri, anlık gerçek verilere, zaman içerisinde sürekli değişken verilerine, kesikli verilere yada kesikli sembolik türlü veri çeşitlerine uyulanabilir.

Ayrıca zaman serisi analizi teknikleri parametrik veya parametrik olmayan yöntemler olmak üzere iki yönteme ayrılabilir. Parametrik yöntemler, verinin çeşidinin altında yatan durağan stokastik sürecin, az sayıda parametre ile tanımlabilen bir yapıdaki veri türü olduğunu varsayar. Bu parametrik yaklaşımlarda amaç, stokastik süreci anlatan modilin parametlerini uygun bir şekilde tahmin etmektir. Parametrik

olmayan yaklaşımlar ise, veri sürecinin belli bir yapıya sahip olmadan, sürecin spektrumunu ve kovaryansını tahmin eder.

Zaman serisi analizi tek değişkenli ve çok değişkenli ile doğrula ya da doğrusal olmayan değişkenli olmak üzerede çeşitlere ayrılabilir.

4. SAYI ZAMAN SERİSİ

İstatistikte, sayım verileri bir istatistiktel tür olduğu gibi gözlemlerin sadece negatif olmayan tam sayıları (0, 1, 2, …) alabileceği bir veri türüdür. Sayım verilerinin istatistiksel olarak ele alınması, gözlemlerin genellikle 0 ve 1 ile temsil edilen yalnızca iki değer alabildiği, ikili verilerden ve aynı zamanda tam sayılardan oluşabilen ancak tek tek değerlerin bir üzerine düştüğü sıralı verilerden farklıdır.

Tek bir sayım veri parçası genellikle bir sayım değişkeni olarak adlandırılır. Böyle bir değişken rastgele bir değişken olarak ele alındığında, dağılımı temsil etmek için Poisson, Binom ve Negatif Binom dağılımları yaygın olarak kullanılır.

Örneklem varyansını stabilize etme özelliğine sahip olacak şekilde seçilen veri dönüşümlerinin kullanımı, sayım verilerinin grafiksel incelemesine yardımcı olabilir. Özellikle, verilere bir Poisson dağılımı ile yaklaşılabildiğinde (diğer dönüşümler makul ölçüde geliştirilmiş özelliklere sahip olmasına rağmen) karekök dönüşümü kullanılabilirken, bir binom dağılımı tercih edildiğinde ters sinüs dönüşümü kullanılabilir.

5. POISSON SAYIM ZAMAN SERİSİ

Regresyon modelini oluşturmadan önce verilerimiz hakkında bilinmesi gereken bazı varsayımları açıklayalım. Tipik olarak, "sayım verileri" terimi, sonlu üst sınır olmaksızın tüm negatif olmayan tam sayıların örnek uzayından ayrı veriler için ayrılmıştır. Bu tanımlanmış örnek uzay, belirli bir zaman aralığında belirli bir olayın meydana gelme sayısını kaydeden Poisson verilerinin özelliğidir. Bir Poisson rastgele değişkeni için, birbirini izleyen olayların bağımsız olarak ve aynı orvea gerçekleştiğini varsayıyoruz.

Poisson rasgele değişkeni olan Y için olasılık kütle fonksiyonu aşağıdaki formülle verilir:

$$P(Y = y|\mu) = \frac{e^{-\mu}\mu^{y}}{y!}, y = 0,1, \dots \quad \textbf{(5.1)}$$

Denklemde görülebileceği gibi Poisson dağılımı, olayların meydana gelme hızı olan pozitif değerli μ ile tanımlanır.

Poisson dağılımının önemli bir özelliği, ortalama ve varyans ilişkisidir. Hem ortalama hem de varyans oran parametresi μ'ye eşittir. Belirli sayılar için regresyon modelimizi oluştururken, ortalama varyans ilişkisi, $E(Y) = var(Y)$, dikkatli bir şekilde dikkate alınmalıdır.

Heteroskedastisite olarak bilinen farklı varyanslar olgusu, doğrusal regresyon analizinde yaygın bir sorundur. Sıradan en küçük kareler (SEKK) regresyonu, hataların bağımsız ve aynı şekilde dağıtılmış normal rastgele değişkenler olduğu varsayımını gerektirir. Verilerimizin Poisson rastgele değişkenlerinin gerçekleşmeleri olduğunu varsayarsak, model hataları her gözlem için aynı olmaz. Çünkü varyans ortalamayla artar. SEKK regresyonunda heteroskedastisite için yaygın bir yöntem, yanıt verilerinin log dönüşümünü yapmaktır. Bu, iki nedenden dolayı sayım verileri için önerilmez. İlk olarak, $y = 0$ ise, $log(y)$ tanımsızdır. İkinci olarak, $exp\big(log(y)\big) = y$ olmasına rağmen $E(y)$'yi tahmin etmek istiyoruz, ancak $exp(E[log(y)]) \neq E(y)$. Bunun yerine, daha geniş bir model sınıfına giren Poisson regresyonuna, yani bir tür genelleştirilmiş doğrusal model (GLM) olan log-doğrusal modele dönüyoruz.

McCullagh ve Nelder (1989), Bölüm 6'da loglineer modeller de dahil olmak üzere GLM'lerin kapsamlı bir tanımını verir. Genelleştirilmiş doğrusal modelin amacı, doğrusal modellemeyi normal dağılımın uygun olmadığı durumlara genişletmektir.

GLM, veriler normal olmadığında kullanılabilir ve ortalama ile tahmin ediciler arasındaki ilişki doğrusal olmadığında kullanılabilir. GLM'nin üç bölümü vardır: dağıtım varsayımı, bağlantı işlevi ve sistematik bileşen. Veriler için bir dağılım ailesini belirtmek için kullanılır ve bağlantı işlevi, ortalamayı doğrusal öngörü olan sistematik bileşene bağlar. Y, $\{Y_t: t = 1, 2, \ldots, n\}$, bağımsız rastgele değişkenlerin bir koleksiyonunu temsil etsin. Öyle ise

$$Y \sim f(y|\theta) \tag{5.2}$$

$$g(E[Y]) = X\beta \tag{5.3}$$

GLM modeli olarak verilebilir. Burada θ, dağılım parametresi, g, bilinen bir link fonksiyonu, X, tasarım matrisi ve β regresyon parametreler vektörüdür.

$\{Y_t: t = 1, 2, \ldots, n\}$, daha sonra tanımlayacağımız geçmiş ve şimdiki bilgilerden bağımsız koşullu rastgele değişkenlerdir. Ayrıca, X_k, t anında k'inci ortak değişken olsun. Öyle ise model

$$Y_t \sim Poisson(\mu_t), t = 1,2, \ldots, n \tag{5.4}$$

$$\log(\mu_t) = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} \tag{5.5}$$

şeklindedir. GLM'nin rastgele bileşeni, Poisson dağılım varsayımını verilere koyar ve Poisson ortalamasını doğrusal tahmin ediciye bağlayan link fonksiyonu olan Eşitlik 2 tipik olarak doğal logaritmadır. Regresyon parametresinin yorumlanması, β_k, açıklayıcı değişkenin beklenen yanıt üzerindeki çarpımsal etkisini içerir. Diğer tüm bağımsız değişkenleri sabit tutmak, x_k bir birim arttıkça, y'nin ortalaması bir e^{β_k} faktörü kadar değişir. Pozitif bir β_k değeri, Y ve $X_k$arasında pozitif bir ilişki olduğunu gösterir.

Bir GLM'nin regresyon parametrelerini tahmin etmek için kullanılan yöntemlerden biri, Wedderburn (1974) tarafından ortaya konan maksimum yarı-olasılık tahmini olarak adlandırılır. Yarı olasılık fonksiyonları, özellikle olabilirlik fonksiyonlarının inşa edilmesinin zor veya imkânsız olduğu durumlarda faydalıdır. Yarı olasılık tahmininin avantajı, gözlemler için bir dağılım belirlememize gerek olmamasıdır. Gözlemlerin varyansını, ortalamanın bazı fonksiyonlarıyla orantılı olarak tanımlamak yeterlidir.

6. KAYIP VERİ ANALİZİ

Kayıp verilere standart yaklaşım, onları silmektir. Veri setlerinde kayıp verileri incelediğimizde, kayıp veri yok gibi görünüyor olsa dahi son veride kayıp veriler sinilerek oluşturulmaktadır. Zhang ve vd. (1994), literatüre katkı sağlayan birçok veri kümesi yayınlamıştır. Bu veriler kullanılarak çok farklı konularda çalışmalar yapılabilmektedir. Koleksiyondaki 510 veri kümesinden yalnızca 13'ü aslında kayıp veriye sahip olarak görülmektedir. Çoğu durumda, eksik veri sorunu veri seti son haline gelene kadar muhtemelen bir şekilde çözülmüştür. Genellikle bize başlangıçta kaç tane eksik değer olduğunu söylemeden önümüze gelmektedir. Zhang (1994)'in kitabındaki çoğu veri seti için orijinal verileri izlemek imkansızdır. Bununla birlikte, bu durumu örneklendirirsek, Seul'deki 1988 Olimpiyat dekatlonunda 10 spor karşılaşmasında, 34 sporcunun puanlarının bir listesi olan 357 veri kümesi için kolayca bulabiliriz. Bu sporcular oyunlarını tamamladı, ancak internette yapılan hızlı bir araştırma, başlangıçta 34 sporcunun yerine 39 sporcunun katıldığını ortaya çıkmaktadır. Bunlardan beşi, 100 metrelik sürat koşusunda üç yanlış başlangıçtan dolayı Alman favori Jürgen Hingsen'in dramatik diskalifiye edilmesi de dahil olmak üzere çeşitli nedenlerle bitiremedi. Bu veriden de anlaşılacağı gibi silme işleminin diğer veri kümelerinin çoğunda anlaşılmadan gerçekleştiğini varsaymak muhtemelen doğrudur.

Kayıp verileri silme eğilimi anlaşılabilir bir durumdur. Kayıp verilerin dayattığı teknik zorlukların yanı sıra, kayıp verilerin ortaya çıkması uzun zaman alır ve özensiz bir araştırmanın işareti olarak kabul edilir.

Bu açıklamada pek çok gerçek olmasına rağmen, bu ideale pratikte ulaşmak imkânsız gibidir. Bilimsel çalışmalarda sıklıkla karşılaşılan bir durum, kayıp verilerin önemsenmemesidir.

Çalışmalar incelendiğinde kayıp verilerin varlığı genellikle metinde açıkça belirtilmez. Liste şeklinde silme gibi varsayılan yöntemler, bunlardan bahsetmeden kullanılır. Genelde bu durumlar çoğu verilerde olmaktadır. Sosyal bilimlerde, bazı katılımcıların katılmayı veya belirli sorulara cevap vermeyi reddetmesi neredeyse kaçınılmazdır. Tıbbi çalışmalarda hastaların verilerinin kaybedilmesi çok yaygındır. Eksik veri problemlerini ele almaya yönelik teori, metodoloji ve yazılım, son on yılda büyük ölçüde genişletildi. Kayıp veri analizinde istatistiksel analiz paketleri artık uygun analizleri gerçekleştirmek için büyük kolaylıklara sahiptir.

Kayıp veriler için bir dizi basit düzeltmeyi incelemeden önce, MCAR, MAR ve MNAR terimlerine kısa bir göz atalım. Rubin (1976) eksik veri sorunlarını üç kategoriye ayırmıştır. Onun teorisine göre her veri noktasının bir miktar eksik olma olasılığı vardır. Bu olasılıkları yöneten süreç, eksik veri mekanizması veya yanıt mekanizması olarak adlandırılır. Süreç için model, eksik veri modeli veya yanıt modeli olarak adlandırılır.

Kayıp olma olasılığı tüm durumlar için aynıysa, verilerin tamamen rastgele eksik olduğu söylenir (MCAR). Bu, kayıp verilerin nedenlerinin verilerle ilgisi olmadığı anlamına gelir. Sonuç olarak, bariz bilgi kaybı dışında, veriler kayıp olduğu için ortaya çıkan karmaşıklıkların çoğunu görmezden gelebiliriz. Bir MCAR örneği, pilleri bitmiş bir tartıdır. Bazı veriler sadece şanssızlık yüzünden eksik olacaktır. Başka bir örnek, her bir üyenin örnekleme dahil edilme şansının aynı olduğu rastgele bir popülasyon örneği aldığımız zamandır. Örnekleme dahil edilmeyen popülasyondaki üyelerin (gözlemlenmemiş) verileri MCAR'dır. Uygun olsa da MCAR eldeki veriler için genellikle gerçekçi değildir.

Kayıp olma olasılığı yalnızca gözlemlenen verilerle tanımlanan gruplar içinde aynıysa, veriler rastgele (MAR) kayıptır. MAR, MCAR'dan çok daha geniş bir sınıftır. Örneğin, yumuşak bir yüzeye yerleştirildiğinde bir tartı, sert bir yüzeye yerleştirilenden daha fazla kayıp değer üretebilir. Bu tür veriler bu nedenle MCAR değildir. Bununla birlikte, yüzey türünü biliyorsak ve yüzey türü içinde MCAR'ı kabul edebilirsek, o zaman veriler MAR'dır. Başka bir MAR örneği, dahil edilme olasılığının bilinen bazı özelliklere bağlı olduğu bir popülasyondan bir örnek aldığımız zamandır. MAR, MCAR'dan daha genel ve daha gerçekçidir. Modern kayıp veri yöntemleri genellikle MAR varsayımından başlar.

Ne MCAR ne de MAR durumu varsa, rastgele olmayan (MNAR) kayıptan söz ederiz. Literatürde, aynı kavram için NMAR (rastgele eksik olmayan) terimi de bulunabilir. MNAR, eksik olma olasılığının bizim bilmediğimiz nedenlerle değiştiği anlamına gelir. Örneğin, tartı mekanizması zamanla yıpranabilir ve zaman ilerledikçe daha fazla eksik veri üretebilir, ancak bunu not edemeyebiliriz. Daha ağır nesneler daha sonra ölçülürse, bozulacak ölçümlerin bir dağılımını elde ederiz. MNAR, ölçeğin daha ağır nesneler için daha fazla eksik değerler üretme olasılığını içerir, bu durum fark edilmesi ve idare edilmesi zor olabilir. Kamuoyu araştırmalarında MNAR'ın bir örneği, zayıf görüşlere sahip olanlar daha az yanıt verirse ortaya çıkar. MNAR'ı ele alma stratejileri, eksikliğin nedenleri hakkında daha fazla veri bulmak veya sonuçların

çeşitli senaryolar altında ne kadar hassas olduğunu görmek için olasılık analizleri yapmaktır.

R kayıp veri setindeki kayıplıp yapısı olsun (değer kayıp ise 1 değil ise 0 olan matris). $Y^{kayıp}$ kayıp veriyi, $Y^{gözlem}$ gözlemlenmiş veriyi ifade etsin. Öyle ise matematiksel ifadeler sırasıyla

- MCAR için $P(R|Y^{gözlem}, Y^{kayıp}) = P(R)$,
- MAR için $P(R|Y^{gözlem}, Y^{kayıp}) = P(R|Y^{gözlem})$, ve
- MNAR için $P(R|Y^{gözlem}, Y^{kayıp}) = P(R|Y^{gözlem}, Y^{kayıp})$

olarak verilebilir (Little ve vd., 2016).

Rubin'in ayrımı teorisi, kayıp bir veri yönteminin geçerli istatistiksel çıkarımlar sağlayabileceği koşulları ortaya koymaktadır. Çoğu basit düzeltme yalnızca kısıtlayıcı ve çoğu zaman gerçekçi olmayan MCAR varsayımı altında çalışır. MCAR mantıksız ise, bu tür yöntemler yanlı tahminler sağlayabilir.

6.1. Kayıp Veri Doldurma

İstatistikte, atama, eksik verileri ikame edilmiş değerlerle değiştirme işlemidir. Bir veri noktasının yerini doldurma, "birim atama" olarak bilinir. Eksik verilerin neden olduğu üç ana sorun vardır: eksik veriler önemli miktarda yanlılığa neden olabilir, verilerin işlenmesini ve analizini daha zor hale getirebilir ve verimlilikte düşüşlere neden olabilir (Barnard ve Meng, 1999). Eksik veriler, verilerin analizinde sorunlar yaratabileceğinden, atama, eksik değerlere sahip vakaların liste halinde silinmesiyle ilgili tuzaklardan kaçınmanın bir yolu olarak görülür. Başka bir deyişle, bir vaka için bir veya daha fazla değer eksik olduğunda, çoğu istatistiksel paket varsayılan olarak eksik bir değeri olan herhangi bir vakayı atar. Bu da yanlılığa neden olabilir veya sonuçların temsil edilebilirliğini etkileyebilir. Kayıp veri doldurma, eksik verileri mevcut diğer bilgilere dayalı tahmini bir değerle değiştirerek tüm durumları korur. Tüm eksik değerler yüklendikten sonra, veri seti tam veri için standart teknikler kullanılarak analiz edilebilir (Gelman ve Hill, 2006). Eksik verileri açıklamak için bilim adamları tarafından benimsenen birçok teori var, ancak bunların çoğu önyargıya yol açıyor. Eksik verilerle başa çıkmak için iyi bilinen birkaç girişim şunları içerir: sıcak güverte ve soğuk güverte ataması, liste ve ikili silme; ortalama atama, negatif olmayan matris çarpanlarına ayırma, regresyon ataması, ileriye taşınan son gözlem, stokastik atama ve çoklu atama.

İngilizce “impute” fiili, hesaplamak, nitelendirmek, hesaba katmak, suçlamak,

atfetmek anlamına gelen Latince imputo'dan gelir. 19. yüzyılda “emsal gelir” kavramı, arazi ve konut gibi mülkten elde edilen geliri belirtmek için kullanılmıştır. İstatistik literatüründe kayıp veri, "verilerin doldurulması" anlamına gelir. Bu anlamda imtiyazdan ilk olarak 1957'de ABD Nüfus Sayım Bürosu'nun (ABD Nüfus Sayımı Bürosu 1957) çalışmasında bahsedilmiştir.

Allan ve Wishart (1930), kayıp bir değeri değiştirmek için istatistiksel bir yöntem geliştiren ilk kişilerdir. Tek bir kayıp gözlemin değerini tahmin etmek için iki formül oluşturdular ve verilerdeki tahminin doldurulmasını tavsiye ettiler (Yates 1933). Bu çalışmayı birden fazla kayıp gözlem için genelleştirdi ve böylece ilk adımları, artık klasik EM (Expectation Maximation) algoritmasına (Dempster, Laird ve Rubin 1977) götüren uzun ve verimli bir ara maddeler zinciri yoluyla geliştirdi. İlginç bir şekilde, "kayıp veri" terimi Dempster ve diğerleri tarafından kullanılmadı.

6.1.1. MI (Çoklu İfade) Yöntemi

Artık birçok alanda eksik verileri tamamlamak için en iyi yöntem olarak çoklu kayıp veri doldurma yöntemi kabul edilmektedir. Ancak bu durum her zaman kabul edilmemektedir. Rubin (1970) tarafından geliştirilmiş bir yöntemdir.

Rubin (1987), geliştirdiği yöntemi istatistiksel temelini oluşturmuştur. 1987'den beri çeşitli düzeltmeler yapmış olsa da zamanın kayıp veri teknolojisinin temellerini oluşturur. Tekrarlanan tam veri tahminini (Rubin kuralları olarak bilinir) birleştirmek için gereken formülleri oluşturmuş ve çoklu kayıp veri altında istatistiksel çıkarımların koşullarını başlıca ana hatlarıyla belirtmiştir.

Bu yöntemde veri setindeki oluşan kayıp gözlemlerin yerine iki veya daha fazla değerin verilmesi ile oluşan yöntemdir. Kayıp olan verileri birden fazla kopyasını oluşturarak her bir kopyanın farklı veriler atama koluyla eksik gözlemlerin tamamlandığı yöntemdir (Scheuren, 2005). Bu yöntemde, kayıp olan verilerin olmasından kaynaklanan belirsizliğe bağlı olarak kesin veya tek bir değer ataması yapılmamaktadır. MI, $m > 1$ sayısında oluşacak veri setinde eksiksiz bir şekilde doldurulacak şekilde yapılması, elde edilen m farklı verilerin belli bir analiz tekniğiyle analiz edilmesi ve elde edilen sonuçların birleştirilmesinden oluşan üç bölümlü süreçtir (Schafer ve Graham, 2002). Bu süreçleri doldurma (imputation), analiz (analysis) ve havuzlama (pool) olarak da isimlendirebiliriz.

MI aslında birçok tekniği tanımlayan genel bir isimlendirmedir. Diğerlerinden farklı olmasının temel nedeni eksik verileri tamamlama evresindeki işleyişi ve faklı algoritmalar kullanmasıyla oluşur. Bu teknikler Bileşik Modelleme ile Tam Koşullu

Tayin olmak üzere iki ayrı başlığa ayrılabilir. (Van Buuren ve vd, 2006).

MI uygulamasında, her veri setinin hata miktarları bulunur. Ardından veri setlerinin ortalama değerleriyle son veri kümesi oluşturulur. Burada dikkat edilmesi gereken ayrıntı, standart hatanın bulunması aşamasında veri setlerinin birleştirilmesi için, veri seti sayısı kadar standart hata miktarlarını toplayıp karekökleri alınmasıdır. Özetle MI, çok daha iyi sonuç alabilmek için bu yöntemi birden fazla çalıştırmak önerilmektedir. MI, farklı araştırmacılar tarafından da tek bir yöntemle bulunan tahmin yöntemlerine göre güvenilir olduğu kanaatine varılmıştır. Bu yöntemin olumsuz yanıysa, birden fazla çalıştırmak gerektiği için işlem sonuçları uzun sürmesidir.

MI şu şekilde örneklendirebiliriz; Q, bilimsel bir tahmin edici ve tüm popülasyonu gözlemlemek için hesaplayabileceğimiz bir ilgi miktarıdır. Örneğin, belli bir nüfusun ortalama gelirini ele alabiliriz. Genel olarak Q, bu kitlenin bilinen bir fonksiyonu olarak tanımlanabilir. Eğer birden fazla nicelikle ilgileniyor olsaydık, Q bir vektör olacaktı.

Q, bu örnekteki popülasyonun bir özelliğidir. Bilimsel tahmin, bu popülasyonun ortalamasını, varyansını, korelasyonunu, regresyon katsayılarını ve bilinen katmanlarını da hesaplamaktır. Bilimsel tahmin işleminde eksik gözlemlerin yerleri doldurularak örnek ortalamalar, hatalar ve test istatistikleri verilebilir. Ancak Q'yu yanlızca bu popülasyonun verileri tam olarak biliniyorsa hesaplayabiliriz. Ancak veriler çoğu zaman eksik gözlem içerdiği için, Q'yu çoğu zaman hesaplayamayabiliriz. Çoklu kayıp verilerin amacı, tarafsız ve güvenilir olan bir Q tahmini bulabilmektir. (Rubin 1996). Tarafsız ve güvenilir bir Q için bu yöntemi aşağıda açıklıyorum. Sapmazsızlık, popülasyondaki tüm Y örnekleri üzerindeki ortalama $\hat{Q}$'nın Q'ya eşit olduğunu anlamındadır. Formülü şu şekildedir:

$$E(\hat{Q}|Y) = Q \qquad \textbf{(6.1)}$$

$\hat{Q}$'nun tahmini varyans – kovaryans matrisi U olsun. Bu durumda güvenilirlik, tüm olası örnekler üzerindeki U ortalamasının, $\hat{Q}$ varyansına eşit veya büyük olması durumudur. Bu durumu şu şekilde gösterebiliriz:

$$E(U|Y) \geq V(\hat{Q}|Y) \qquad \textbf{(6.2)}$$

Burada $V(\hat{Q}|Y)$ işlevi, örneklemenin varyansını gösterir. Bu durumda hipotezin reddetme oranı %5 olan bir istatistiksel testte, hipotezin gerçekte olduğu durumlarda en fazla %5'inde hipotezi reddetmelidir. Bu hipotez tutulursa, yöntemin güvenilir

olduğu söylenir.

Özet olarak, çoklu kayıp verilerin amacı, popülasyondaki bilimsel tahminlerin tahminini elde etmektir. Bu tahmin ortalama olarak popülasyonun parametrelerine eşit olmalıdır. Ayrıca, güven aralıkları ve hipotez testlerinin belirtilen normal değerlerine de ulaşmalıdır.

6.1.2. MICE (Çoklu İfade Zincirli Denklemler) Yöntemi

MICE yöntemi kayıp verileri tamamlama sırasında sıralı denklemler oluşturur ve eksik verileri içeren değişkeni, bu oluşturduğu denklemleri atamalar yaparak kayıp verileri tamamlar. MICE yönetiminde tamamlanan kayıp veriler koşulludur, nedeni kayıp olarak tamamlanan veriler, verinin değişkenlerinin bilgisine dayanarak oluşturulmuştur (Rijnhart ve vd., 2019). MICE yöntemini 9 basamakta özetleyerek gösterebiliriz (Van Buuren , 2018).

I en fazla yenileme sayısını ve t yenilemenin kaç kere olduğunu tanımlasın,

1. $j = 1, \dots, p$'ye kadar olan veriyi Y_j için modeli $P\left(Y_j^k \middle| Y_j^g, Y_j, R\right)$ şeklinde oluştur.
2. Tüm j için Y_j^g' den rastgele seçim ile Y_j^0'i bitir.
3. $t = 1, \dots, I$ şeklinde tekrarla
4. $j = 1, \dots, p$ şeklinde tekrarla
5. $Y_{-j}^t = (Y_1^t, \dots, Y_{j-1}^t, Y_{j+1}^{t-1}, \dots, Y_p^{t-1})$, t zamanında Y_j dışında bitmiş verilerdir.
6. $\emptyset_j^t \sim P\left(\emptyset_j^t \middle| Y_j^g, Y_{-j}^t, R\right)$
7. Eksik verileri tamamla $Y_j^t \sim P\left(Y_j^k \middle| Y_j^g, Y_{-j}^t, R, \emptyset_j^t\right)$
8. j'yi bitir.
9. t'yi bitir.

MICE yönteminde eksik veri tamamlamada ilk olarak rastgele seçimler yaparak başlatır. Her yenileme verinin tamamı için bir döngü oluşturur. MICE yönteminde genellikle 5-10 yenilemede ekik gözlemlerin tamamlanması için yeterli olması beklenir (Van Buuren, 2018).

Tüm tamamlanmış eksik gözlemler, tamamlanmış gözlemlerin derlemesi olduğundan, MICE yöntemi aslında bir markov zinciri yöntemidir. Aslında uygun şartlar ile MICE yöntemi, Gibbs (bir bayesyen simülasyonu) örnekleyicisidir (Gelfand ve Smith, 1990).

Sabit bir dağılama yakınsamak için, Markov zincirinin üç önemli özelliği bulunmaktaıdır (Roberts, 1998). Van Buurn (2018) nadir durumlar dışında MICE

yönteminin bu özelliklere sahip olduğunu belirtmiştir.

1. İndirgenemez (irreducible): Zincir, verinin tüm kısımlarına ulaşabilmedir.
2. Periyodik olmamalı (aperiodic): Zincir, farklı durumlar arasında olmalıdır
3. Tekrarlamalı (recurrence): Verideki ilgili alanlara, sınırsız sıklıkla uşabilmelidir.

MI yönteminde, veri setleri bağımsız paralel bir şekilde elde edilir. MICE metodu MI'a bağlı olarak oluşan adımların sırası ile uygular. Böylelikle Y, $P(Y|\theta)$ şeklinde çok değişkenli bütün bir veri kümesi olur. MICE algoritması, koşullu dağılımlarından yinelemeli örnekleme yaparak θ'nin önceki dağılımını elde eder.

$$\begin{matrix} P(Y_1 \backslash Y_{-1}, \theta_1) \\ \vdots \\ P(Y_p \backslash Y_{-p}, \theta_p). \end{matrix} \quad \textbf{(6.3)}$$

Burada, $\theta_1, \dots, \theta_p$ parametreleri yapının karşılıklı yoğunluğuna göre hesaplanır. Gözlemlenen değişken, dağılımlardan rastgele başlayarak, peş peşe denklemlerin t-inci yinelemesi ile oluşan bir Gibbs örnekleyicisidir.

$$\begin{matrix} Y_1^{*(t)} \sim P\left(Y_1 | Y_1^{obs}, Y_2^{(t-1)}, \dots, Y_p^{(t-1)}, \theta_1^{*(t)}\right) \\ \vdots \\ \theta_p^{*(t)} \sim P\left(\theta_p | Y_p^{obs}, Y_1^{(t)}, \dots, Y_{p-1}^{(t)}\right) \\ Y_p^{*(t)} \sim P\left(Y_p | Y_p^{obs}, Y_1^{(t-1)}, \dots, Y_p^{(t-1)}, \theta_p^{*(t)}\right). \end{matrix} \quad \textbf{(6.4)}$$

Burada, $Y_j^{(t)} = \left(Y_j^{obs}, Y_j^{*(t)}\right)$, t yinelemesinde j-inci atanmış değişkendir. Bu işlem, belirli bir sınıra veya maksimum yineleme sayısına ulaşıldığında sona erer. Her adımda çeşitli farklı yöntemler de kullanılabilir (Buuren ve vd., 2010).

MICE yönteminin diğer markov zinciri yöntemlerine göre en büyük avantajı, çok yüksek sayıda yenilemeye ihtiyaç duymamasıdır. (Van Buuren, 2018)'de belirtildiği gibi genellikle 5-10 yenilemede eksik gözlemler tamamlanabilmektir.

6.1.2.1. MICE Random Forest (MICE rastgele orman)

Tarafsız tahminler elde etmede analizler için atama modellerinin doğru bir şekilde belirtilmesi önemlidir ve rastgele orman, ikincisinin atama modelleri yanlış belirtilirse parametrik MICE ile oluşabilecek yanlılıktan kaçınmaya yardımcı olabilir. Rastgele orman, doğrusal olmayan durumları ve etkileşimleri otomatik olarak barındırması gerektiğinden, öngörücü değişkenler arasındaki ilişkileri araştırma ihtiyacını azaltır. Tahmin modelleri aynı zamanda asli model ile uyumlu olmalıdır ve rastgele orman, ortak değişkenler için atama modellerinde sonucun nasıl

koşullandırılacağını belirtme ihtiyacını ortadan kaldırır. Rastgele ormanın bir dezavantajı, "modellerin" karmaşık olması ve kolayca yorumlanamamasıdır, ancak bu, tartışmaya açık bir şekilde, kayıp veri doldurmada bir eksiklik değildir. Diğer bir dezavantaj, aralıklarının uç noktalarındaki sürekli değişkenlerin rasgele orman tahminlerinin daha az uç değerlere eğilimli olması nedeniyle bazı durumlarda rastgele ormanın yanlı olabilmesidir. Bunun nedeni, rastgele bir orman tahmininin, tahmin edilen değişkenin gözlenen değerlerinin ağırlıklı ortalamasından oluşmasıdır. Modele dayalı tahminden farklı olarak, gözlemlenen değerlerin ötesine geçemez (Shah ve vd., 2014).

6.1.2.2. Tahmini Ortalama Eşleştirme Yöntemi (PMM)

Tahmine dayalı ortalama eşleştirme yöntemi, bir v değişkenindeki eksik bir değerin, bir donörden gelen v değeri ile değiştirildiği, regresyon tahmin puanına sahip bir katılımcının, kendisi için regresyon tahmini puanına en yakın olduğu stokastik bir regresyon tekniğidir. Diğer stokastik regresyon yöntemleri gibi, standart hata tahmini açısından hem ortalama hesaplama hem de deterministik regresyon yöntemlerinden üstündür (Kalton, 1986; David ve vd. 1986). Diğer stokastik tekniklerle karşılaştırıldığında, kolayca operasyonel hale getirilebilir ve empoze edilecek değerin gerçek değerleri atandığından, ayrık ve sürekli ölçümler yapmak için uygundur. Tek değer atama teknikleri ayrıca çoklu değer atama teknikleri ile karşılaştırılır. İlkinde, eksik bir değeri değiştirmek için tek bir tahmin kullanılır. İkincisi ile, ilgili bir istatistiğin ve bunun standart hatasının hesaplanmasında çeşitli tahminler birleştirilir.

MICE midastouch yöntemi ise, MICE PMM yöntemine benzemektedir. Mesafe destekli verici seçimi ile tahmine dayalı ortalama eşleştirmeyi kullanarak tek değişkenli eksik verileri uygular.

6.1.3. Enterpolasyon

Enterpolasyon yöntemi ile kayıp veri tahmininde, basit tabirle veri setinde mevcut olan değerler ile, kayıp gözlemin tahmin edilmesi metodudur. Bu yöntemde amaç $x_0, x_1, x_2, \ldots, x_n$ noktaları için $f_0, f_1, f_2, \ldots, f_n$ hesaplamalarından yararlanarak $x_i - x_{i+1}$arasındaki bir x için, F_x ara değeri hesaplamaktır.

Doğrusal enterpolasyon yönteminde verilen $(x_0, f(x_0)), (x_1, f(x_1))$ noktalarından yararlanarak $x_0 - x_1$ arsındaki herhangi bir x değeri için $F(x)$ değeri, bu iki nokta arasındaki değişimin doğrusal olduğu düşünülerek bulunabilir.

Doğrusal enterpolasyon fonksiyonu denklemi bu şekildedir (Chapra ve Canale, 1998):

$$f(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0) \quad \textbf{(6.5)}$$

Burada x bir bağımsız değişken, x_1 ve x_0 bu bağımsız değişkenin bilinen değerleridir. Ayrıca $f(x)$, bağımsız değişkenin bir x değeri için, bağımlı değişkenin değeridir.

İki veri noktasını düz bir çizgiyle birleştirme yöntemi, en basit enterpolasyon yöntemi olarak tanımlanabilir. Bu tekniğe doğrusal enterpolasyon yöntemi denir. Doğrusal enterpolasyon yöntemi fonksiyonu şu şekildedir (Noor ve vd., 2014).

$$f_1(x) = b_0 + b_1(x - x_0) \quad \textbf{(6.6)}$$

Burada x bağımsız değişken, x_0 bağımsız değişkenin bilinen bir değeridir. Ayrıca $f_1(x)$ bağımsız değişkenin bir x değeri için bağımlı değişkenidir.

Son denklem (6.6)'den fonksiyon aşağıdaki şekilde olur,

$$b_0 = f(x_0) \quad \textbf{(6.7)}$$

ve

$$b_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \quad \textbf{(6.8)}$$

Eğer üç veri noktası mevcutsa, tahminin, ikinci dereceden bir polinom kullanarak yapılır. Bu tahmin fonksiyonu aşağıdaki gibi verilmiştir:

$$f_2(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1) \quad \textbf{(6.9)}$$

Burada x bağımsız değişken, x_0 ve x_1 bağımsız değişkenin bilinen değerleridir ve b_0 ile b_1 bilinmeyen katsayılardır. Ayrıca $f_2(x)$ ikinci derecen bir enterpolasyon polinomdur. b_0 ve b_1'in katsayılarını belirleme fonksiyonu, denklem (6.7) ve (6.8) ile aynı şekilde hesaplanır. b_2 için katsayılar aşağıdaki gibi elde edilir (Noor ve vd., 2014):

$$b_2 = \frac{\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0} \quad \textbf{(6.10)}$$

şeklinde elde edilir.

Eğer dört veri noktası mevcutsa, üçüncü dereceden bir polinom kullanılarak yapılır. Bu yöntemine kübik enterpolasyon olarak da isimlendirilebilir. Formülü şu şekildedir. Denklem (6.10)'in devamı

$$f_3(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1) + b_3(x - x_0)(x - x_1)(x - x_2)$$

Burada b_0, b_0 ve b_2'nin katsayılarının belirlenme yöntemi denklem (6.8) - (6.10) ile aynı şekildedir. B3'ün hesaplanması şu şekildedir (Noor ve vd., 2014):

$$b_3 = \frac{\frac{f(x_3) - f(x_2)}{x_3 - x_2} - \frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_3 - x_0} \quad \textbf{(6.11)}$$

şeklinde elde edilir.

6.1.4. Son Gözlemi İleri Taşıyarak Doldurma (LOCF)

Kayıp verileri ele almanın bir yöntemi, basitçe mevcut verilere dayalı değerleri atamak veya doldurmaktır. Bunu yapmak için standart bir yöntem, Son Gözlemi İleri Taşıma (LOCF) yöntemidir.

LOCF yönteminin tarihsel olarak başlangıcı belirsizdir. İlk ortaya çıkış zamanlarında tarafsız bir sonuç verdiği konusunda şüpheli yaklaşımlardan dolayı tek bir hakemli yayını olmamıştır. Ancak bu durum LOCF yönteminin yaygınlığı konusnuda olumsuz bir etki yapmamıştır. Günümüzde en çok atıf alan yöntemlerden biri olmayı başarmıştır (Lachin., 2016).

LOCF yöntemi verilerde çok sık karşılaşılan, veri akışının kesilmesinden kaynaklanarak oluşan kayıp veriyi, son gözlemiş ileri alarak çözümde bulunan kolay bir yöntemdir. Bu yöntem genellikle sağlık alanında yaygın olarak kullanılabilmektedir. Veri oluşumunda (kişi, değerler vb.) değerinde ilk gözlemlenmiş değer, kendinden bir önceki veri ile yerine eklenir ve bu süreç veri tamamen bitene kadar devam eder (Little ve Rubin, 2019). Farklı olarak veriye sonradan ekleme olduğunda, eksik gözlemi geriye taşıma yöntemine gidilebilir. Bu yönteme NOCB (next observation carried backward) denilir (Engels, 2003).

LOCF yaklaşımının avantajları analizden elenen denek sayısını en aza indirmesi ve analizin yalnızca son noktaya odaklanmak yerine zaman içindeki eğilimleri incelemesine olanak tanımasıdır.

LOCF yöntemi ne kadar kullanışlı ve basit bir yöntem olsa dahi, klinik bir veri setinde, kişinin (deneğin) ara vermesi veya tamamen ayrılması sonucunda, eksik kalan verinin son ölçülmüş veriye eşit olacağından sağlıklı bir eksik gözlem tamamlama yöntemi olmayacaktır (Molenberghs ve Kenward, 2007). Bununla birlikte, Ulusal Bilimler Akademisi, Gıda ve İlaç İdaresi'ne klinik araştırmalardaki eksik verilerle ilgili bir danışma raporunda, LOCF gibi yöntemlerin kritik olmayan kullanımına karşı tavsiyede bulunarak, LOCF ve benzeri yöntemlerinin altında yatan varsayımlar bilimsel olarak doğrulanmadıkça, eksik verilerin tedavisine yönelik birincil yaklaşım olarak kullanılmaması gerektiğini söylemiştir (Little vd., 2012).

LOCF'nin altında yatan temel varsayım, yani eksik verilerin sanki yokmuş gibi

analizin yapılması, çoğu zaman doğru değildir. Örneğin, birçok ilaç, hastaların gözlem altındayken kötüleşmesi veya ölmesi beklenen kanser, kalp yetmezliği veya AIDS gibi durumları tedavi eder. Ayrıca tedavi edici ilaçların bile zararlı ve bazen ölümcül yan etkileri ve güvenlik sorunları olabilir. Bu tür araştırma için, sanki geçmiş değişmeden devam ediyormuş gibi eksik verilerin ele alınması, etkinliğin fazla rapor edilmesine veya zararlı güvenlik sorunlarının eksik bildirilmesine neden olabilir ve sonuçları araştırma tedavisinin gerçekte olduğundan daha güvenli veya daha etkili görünmesini sağlayacak şekilde önyargılı hale getirebilir.

Ek olarak, uygun olmayan yanlılık eklemeseler bile, basit değerlendirme yöntemleri, tahminlerin kesinliğini ve güvenilirliğini ve denemenin tedaviyi değerlendirme gücünü abartır. Veriler eksik olduğunda, tahminlerin dayandığı örnek boyutu düşürülür. Basit atama yöntemleri, örneklem büyüklüğündeki bu azalmayı hesaba katmaz ve bu nedenle sonuçların değişkenliğini hafife alma eğilimindedir.

6.1.5. Hareketli Ortalama Kayıp Veri Doldurma (MA)

Hareketli ortalama, kısa vadeli dalgalanmaları yumuşatmak ve uzun vadeli eğilimleri veya döngüleri vurgulamak için zaman serisi verileriyle yaygın olarak kullanılır. Kısa vadeli ve uzun vadeli arasındaki eşik değeri, yapılacak uygulamaya bağlıdır ve hareketli ortalamanın parametreleri buna göre ayarlanacaktır. Ekonomide gayri safi yurtiçi hasıla, istihdam veya diğer makroekonomik zaman serilerini incelemek için de kullanılır. Matematiksel olarak, hareketli ortalama bir tür güncellemedir ve bu nedenle sinyal işlemede kullanılan bir alçak geçiren filtre örneği olarak görülebilir. Zaman serisi olmayan verilerle kullanıldığında, hareketli bir ortalama, tipik olarak bir tür sıralama ima edilmesine rağmen, zamana herhangi bir özel bağlantı olmaksızın daha yüksek frekanslı bileşenleri filtreler. Basitçe bakıldığında, verilerin düzgünleştirilmesi olarak kabul edilebilir.

Hareketli ortalama yöntemiyle, hatalı, aykırı veya beklenmedik verileri zamana bağlı akışı üzerindeki etkileri azaltarak daha düz bir çizgide olmasını sağlamaktır. Hareketli ortalamanın belirli bir algoritması bulunmamaktadır (Seker, 2015).

6.1.5.1. Basit Hareketli Ortalama (SMA)

Finansal uygulamalarda basit hareketli ortalama (SMA), önceki n veri noktasının ağırlıksız ortalamasıdır. Bununla birlikte, bilim ve mühendislikte ortalama, normalde merkezi bir değerin her iki tarafındaki eşit sayıda veriden alınır. Bu, ortalamadaki varyasyonların, zaman içinde kaydırılmak yerine verilerdeki varyasyonlarla hizalanmasını sağlar. Zaman serisinde, verilerin kendinden önceki n

veri noktasının ortak bir ortalamasındır. Verideki her nokta eşit aralıktadır. Bundan dolayı bu noktaların hiçbirine uygulanan ağırlıklandırma faktörü yoktur (Hansun, 2013).

Basit hareketli ortalamanın formülü bu şekildedir:

$$SMA = \frac{P_M + P_{M-1} + \cdots + P_{M-(n-1)}}{n} \quad \textbf{(6.12)}$$

Yukarıdaki formülde P_M, M zamanındaki veri noktasının kısaltması, n hesaplamada kullanılan veri sayısıdır. Ardışık olarak kullanıldığında, formülün toplamına yeni değer gelir ve en eski veri çıkartılır. Bu durumun formülü aşağıdaki gibidir:

$$SMA_{Bugün} = \frac{P_M}{n} + SMA_{Dün} - \frac{P_{M-n}}{n} \quad \textbf{(6.13)}$$

şeklinde oluşmaktadır.

Kullanılan veriler ortalama etrafında ortalanmazsa, basit bir hareketli ortalama, örnek genişliğinin yarısı kadar en son verinin gerisinde kalır. Bir SMA, eski verilerin düşmesinden veya gelen yeni verilerden de orantısız bir şekilde etkilenebilir. SMA'nın bir özelliği, verilerin periyodik bir dalgalanması varsa, o zaman o döneme ait bir SMA'nın uygulanması bu varyasyonu ortadan kaldıracaktır. Ancak mükemmel düzenli bir döngüye nadir rastlanır (Chou, 1975).

SMA'nın önemli bir dezavantajı, pencere uzunluğundan daha kısa olan önemli miktarda sinyalin geçmesine izin vermesidir. Daha da kötüsü, aslında onu tersine çevirir. Bu, verilerde çukurların olduğu yerde görünen düzleştirilmiş sonuçtaki tepeler gibi beklenmedik yapaylıklara yol açabilir. Ayrıca, bazı yüksek frekanslar düzgün bir şekilde kaldırılmadığından, sonucun beklenenden daha az düzgün olmasına yol açar.

6.1.5.2. Ağırlıklı Hareketli Ortalama (WMA)

Ağırlıklı ortalama, örnek penceresindeki farklı konumlardaki verilere farklı ağırlıklar vermek için çarpan faktörlerine sahip bir ortalamadır. Matematiksel olarak, ağırlıklı hareketli ortalama, sabit bir ağırlık fonksiyonu ile verilerin güncellemedir.

Basit hareketli ortalamanın bir iyileştirme şeklidir. Ağırlıklı hareketli ortalama eski veriye nazaran yeni verilere daha fazla ağırlık verir. Ağırlık faktörleri, bilinen zaman serisi verilerinde kullanılan günlerin toplamından hesaplanır. WMA aşağıdaki gibi hesaplanır:

$$WMA = \frac{nP_M + (n-1)P_{M-1} + \cdots + 2P_{M-n+2} + P_{M-n+1}}{n + (n-1) + \cdots + 2 + 1} \quad \textbf{(6.14)}$$

6.1.5.3. Üstel Hareketli Ortalama (EMA)

Üstel ağırlıklı hareketli ortalama (EWMA) olarak da bilinen üstel hareketli ortalama (EMA), üstel olarak azalan ağırlıklandırma faktörlerini uygulayan birinci dereceden bir sonsuz dürtü yanıt filtresidir. Her eski veri için ağırlık katlanarak azalır ve asla sıfıra ulaşmaz.

Üstel hareketli ortalama ilk olarak, son günün veriye olan etkisini alan α değeri ile hesaplanır. Bu α değeri, 0 ile 1 arasında olacaktır.

EMA aşağıdaki gibi hesaplanır:

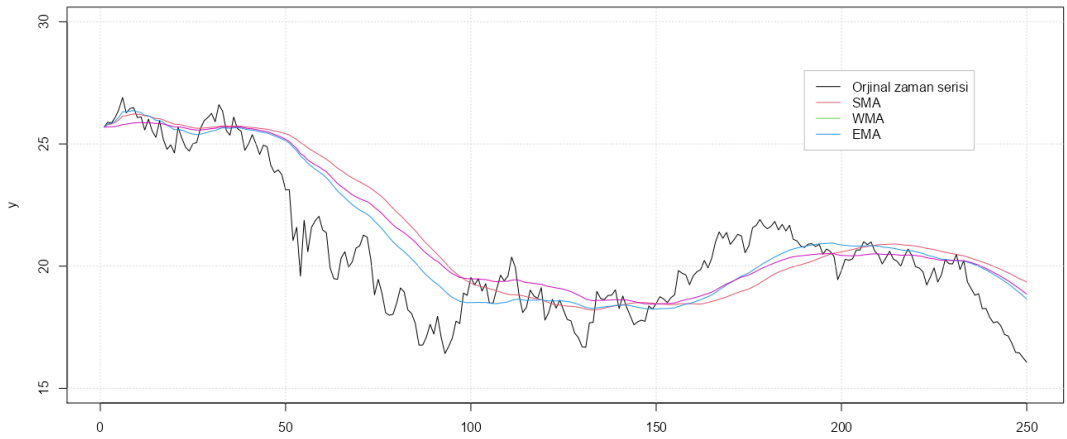
$$S_1 = Y_1$$
$$t\ için > 1, S_t = \alpha \times Y_t + (1-\alpha) \times S_{t-1} \quad \textbf{(6.15)}$$

Şeklinde oluşmaktadır. Buda Y_t, t zamanındaki değeridir. S_t, t zamanındaki üstel hareketli ortalamanın değeridir. α ağırlık düşüşünün derecesini temsil eder (Hansun, 2013).

Bura da α genellikle şu şekilde hesaplanır: $\alpha = \frac{2}{n+1}$. Hunter (1986)'ya göre

$$S_t = \alpha[Y_t + (1-\alpha)Y_{t-1} + (1-\alpha)^2 Y_{t-2} + \cdots + (1-\alpha)^k Y_{t-k}] + (1-\alpha)^{k+1} S_{t-(k+1)} \quad \textbf{(6.16)}$$

formülünün tekrar tekrar farklı t'ler için tekrar edilmesi ile Y_t veri noktlaarının ağırlıklı toplamını S_t olarak yazabiliriz. Burada $k \in \{0,1,2,\ldots\}$ ve Y_{t-i} genel veri noktasının ağırlığı $\alpha(1-\alpha)^i$ şeklindedir. Şekil 1'de farklı hareketli ortalama yöntemlerine ait eğri örnekleri gösterilmiştir.



Şekil 6.1. Farklı hareketli ortalama yöntemlerine ait eğri örneği

6.1.6. Kalman Yöntemi

Kalman yöntemi, optimal en düşük ortalama varyans tahmin edicisidir ve Kalman filtresi olarak adlandırılır. İlgili sistemlerin hata özellikleri hakkında güçlü istatistiksel bilgiler verebilmektedir. Bunu, verinin belli alanına bakmadan, tüm bilgileri kullanarak yapar. Hata analizi için çok kullanışlıdır.

Kalman filtresinde, değişkenlerin belirsizliklerini kullanılarak tahminler yapılır. Bir sonraki ölçümün sonucu gözlemlendiğinde, ağırlıklı ortalamalar kullanılarak bu tahminler güncellenir. Belirsizliği az olan ölçüme daha fazla ağırlık verilir. Öz yinelemeli olan bu algoritma için tüm geçmiş bilgiye ihtiyaç yoktur. Mevcut girdi ölçümlerinin yanı sıra geçmişe ait durum ve belirsizlik matrisinin olması yeterlidir.

Kalman filtresi, ilk olarak Swerling (1958), Kalman (1960) ve Kalman ve Bucy (1961) tarafından teknik makalelerde tanımlanmış ve kısmen geliştirilmiştir. Kalman filtresi, hataların normal dağılımlı olduğunu varsayar. Bu yöntemi bulan Kalman (1960), bu hataların normal dağılımlı ve sıfır ortalamalı olduğu varsayımını yapmıştır. Kalman filtresi, süreç ve ölçüm kovaryansları biliniyor iken hata kareler ortalaması bakımından mümkün olabilecek en iyi doğrusal tahmin edicidir (Humpherys, 2012).

Gürültülü veriler, sistem gelişimini tanımlayan denklemlerdeki yaklaşımlar ve tümü hesaba katılmayan dış faktörler sistemin durumunu belirlerken kısıtlar oluşturur. Kalman filtresi, gürültülü sensörlü verilerinden ve bir dereceye kadar rastgele dış faktörlerden kaynaklanan belirsizliklerle etkin bir şekilde ilgilenir. Ağırlıklı ortalama kullanarak yeni ölçümün ortalamasını sistem durumunun bir tahmini olacak şekilde üretir. Ağırlıkların amacı, daha iyi (yani daha küçük) tahmini belirsizliğe sahip değerlerin daha fazla "güvenilir" olmasını sağlamaktır. Ağırlıklar, sistemin durumuna ilişkin tahminin tahmini belirsizliğinin bir ölçüsü olan kovaryanstan hesaplanır. Ağırlıklı ortalamanın sonucu, tahmin edilen ve ölçülen durum arasında yer alan ve tek başına olduğundan daha iyi tahmin edilen belirsizliğe sahip yeni bir durum tahminidir. Bu süreç, yeni tahmin ve sonraki yinelemede kullanılan tahmini bildiren kovaryansı ile her zaman adımında tekrarlanır. Bu, Kalman filtresinin özyinelemeli çalıştığı ve yeni bir durumu hesaplamak için bir sistemin tüm geçmişinden ziyade yalnızca son "en iyi tahminini" gerektirdiği anlamına gelir.

Ölçümlerin ve mevcut durum tahmininin göreceli kesinliği önemli bir husustur. Filtrenin tepkisi, Kalman filtresinin kazancı açısından tartışılan bir konudur. Kalman kazancı, ölçümlere ve mevcut durum tahminine verilen göreli ağırlıktır ve belirli bir performansı elde etmek için ayarlanabilirdir. Yüksek kazançlı filtre, en son ölçümlere

daha fazla ağırlık verir ve böylece onları daha duyarlı bir şekilde takip eder. Düşük kazanç ile filtre, model tahminlerini daha yakından takip eder. Uç noktalarda, bire yakın yüksek bir kazanç daha ürkek bir tahmini yörüngeye neden olurken, sıfıra yakın düşük bir kazanç gürültüyü yumuşatacak ancak tepkiyi azaltacaktır.

Kalman Filtresi için gerçek hesaplamalar yapılırken, durum tahmini ve kovaryanslar, tek bir hesaplama setinde yer alan çoklu boyutları işlemek için matrislere kodlanır. Bu, herhangi bir geçiş modeli veya kovaryans içinde farklı durum değişkenleri (konum, hız ve ivme gibi) arasındaki doğrusal ilişkilerin temsiline olanak verir.

Kalman filtresi, bir dizi gürültülü ölçümden doğrusal bir dinamik sistemin iç durumunu tahmin eden verimli bir özyinelemeli filtredir. Radar ve bilgisayarlı görüden yapısal makroekonomik modellerin tahminine kadar çok çeşitli mühendislik ve ekonometrik uygulamalarda kullanılır (Andreasen, 2008; Strid ve Walentin, 2009) ve kontrol teorisi ve kontrol sistemleri mühendisliğinde önemli bir konudur. Lineer-kuadratik regülatör (LQR) ile birlikte Kalman filtresi, lineer-kuadratik-Gauss kontrol problemini (LQG) çözer.

Dempster-Shafer teorisinde, her durum denklemi veya gözlemi, doğrusal bir fonksiyonunun özel bir durumu olarak kabul edilir. Kalman filtresi, bir birleştirme ağacı veya Markov ağacında doğrusal işlevlerini birleştirmenin özel bir durumudur. Ek yaklaşımlar, durum denklemlerinde Bayes veya kanıtsal güncellemeleri kullanan filtrelerini içerir.

Kalman filtreleri, zaman alanında ayrıklaştırılmış doğrusal dinamik sistemlere dayanmaktadır. Gauss gürültüsünü içerebilecek hatalarla bozulan lineer operatörler üzerine inşa edilmiş bir Markov zinciri üzerinde modellenmiştir. Hedef sistemin durumu, gerçek sayıların bir vektörü olarak temsil edilen, ilgilenilen temel gerçek sistem konfigürasyonunu ifade eder. Her ayrık zaman artışında, yeni durumu oluşturmak için duruma bir lineer operatör uygulanır. Ardından, daha fazla gürültü ile karıştırılan başka bir doğrusal operatör, gerçek ("gizli") durumdan ölçülebilir çıktılar (yani gözlem) üretir. Kalman filtresi, gizli Markov modelinde olduğu gibi ayrık bir durum uzayının aksine, gizli durum değişkenlerinin sürekli bir uzayda değerler almasıyla, gizli Markov modeline benzer olarak kabul edilebilir. Kalman Filtresi denklemleri ile gizli Markov modelinin denklemleri arasında güçlü bir analoji vardır.

Gürültülü gözlemler dizisi verilen bir sürecin içsel durumunu tahmin ederken Kalman filtresi kullanabilmek için bu süreç belirli bir çerçevede modellenmelidir.

Yalnızca bir dizi gürültülü gözlem verilen bir sürecin iç durumunu tahmin etmek için Kalman filtresini kullanmak için, süreci aşağıdaki çerçeveye göre modellemek gerekir. Bu çerçevede F_k, durum geçiş modeli; H_k, gözlem modeli; Q_k, süreç gürültüsü kovaryansı; R_k, gözlem gürültüsü kovaryansı; ve bazen B_k, her bir k zaman adımı için kontrol-girdi modeli belirlenir.

Kalman filtre modeli, k anındaki gerçek durumun $k-1$ anındaki durumdan Denklem 6.17'e göre evrimleştiğini varsayar.

$$x_k = F_k x_{k-1} + B_k u_k + w_k \quad \textbf{(6.17)}$$

Burada, F_k, önceki $x_k - 1$ durumuna uygulanan durum geçiş modelidir; B_k, u_k kontrol vektörüne uygulanan kontrol-girdi modelidir; w_k, kovaryanslı bir sıfır ortalama çok değişkenli normal dağılımdan geldiği varsayılan süreç gürültüsüdür, $Q_k: w_k \sim N(0, Q_k)$. k anında x_k gerçek durumunun z_k gözlemi Denklem 6.18'ye göre yapılır.

$$z_k = H_k x_k + v_k \quad \textbf{(6.18)}$$

Burada H_k, gerçek durum uzayını gözlenen uzaya eşleyen gözlem modelidir; v_k, kovaryanslı sıfır ortalama Gauss beyaz gürültüsü olduğu varsayılan gözlem gürültüsüdür: $R_k: v_k \sim N(0, R_k)$. Başlangıç durumu ve her adımda $\{x_0, w_1, \ldots, w_k, v_1, \ldots, v_k\}$ gürültü vektörlerinin hepsinin karşılıklı olarak bağımsız olduğu varsayılır.

Birçok gerçek zamanlı dinamik sistem bu modele tam olarak uymaz. Aslında, modellenmemiş dinamikler, girdi olarak bilinmeyen stokastik sinyallerle çalışması gerektiğinde bile filtre performansını ciddi şekilde düşürebilir. Bunun nedeni, modellenmemiş dinamiklerin etkisinin girdiye bağlı olması ve dolayısıyla tahmin algoritmasını kararsızlığa getirebilmesidir. Öte yandan, bağımsız beyaz gürültü sinyalleri algoritmanın sapmasına neden olmaz. Ölçüm gürültüsü ile modellenmemiş dinamikler arasında ayrım yapma sorunu zor bir sorundur ve kontrol teorisinde sağlam kontrol çerçevesinde ele alınmaktadır (Ishihara vd., 2006; Terra vd., 2014).

Kalman filtresi özyinelemeli bir tahmin edicidir. Bu, mevcut durum için tahminin hesaplanması için yalnızca önceki zaman adımından tahmin edilen durum ve mevcut ölçümün gerekli olduğu anlamına gelir. Toplu tahmin tekniklerinin aksine, hiçbir gözlem veya tahmin geçmişine gerek yoktur. Buradan itibaren $\hat{x}_{n|m}$, $m \leq n$ zamanında ve $m \leq n$ zamanına kadar olan n anındaki x'in tahminini ifade edecektir.

Filtrenin durumu iki değişkenle temsil edilir: $\hat{x}_{k|k}$, k dahil olacak şekidle k

zamanına kadar olan gözlemlerin sonsal durum tahmini; $P_{k|k}$, kovaryans matrisinin sonsal tahmini.

Kalman filtresi tek bir denklem olarak yazılabilir, ancak çoğunlukla iki farklı aşama olarak kavramsallaştırılır: "Tahmin" ve "Güncelleme". Tahmin aşaması, mevcut zaman adımında bir durum tahmini üretmek için önceki zaman adımındaki durum tahminini kullanır. Bu tahmin edilen durum tahmini aynı zamanda önsel durum tahmini olarak da bilinir, çünkü bu, mevcut zaman adımındaki durumun bir tahmini olmasına rağmen, mevcut zaman adımından gözlem bilgilerini içermez. Güncelleme aşamasında, yenilik (uygulama öncesi kalıntı), yani mevcut önsel tahmin ile mevcut gözlem bilgisi arasındaki fark, optimal Kalman kazancı ile çarpılır ve durum tahminini iyileştirmek için önceki durum tahmini ile birleştirilir. Mevcut gözleme dayalı bu geliştirilmiş tahmin, sonsal durum tahmini olarak adlandırılır.

Kalman filtresi aşamaları, tahminin durumunun bir sonraki programlanmış gözleme kadar ilerlemesi ve gözlemi içeren güncelleme ile değişir. Ancak bu gerekli değildir. Herhangi bir nedenle bir gözlem kullanılamıyorsa, güncelleme atlanabilir ve birden fazla tahmin adımı gerçekleştirilebilir. Benzer şekilde, aynı anda birden fazla bağımsız gözlem mevcutsa, birden fazla güncelleme adımı gerçekleştirilebilir (Kelly, 1994).

Tahmin aşamasında Tahmin edilmiş durum tahmini,

$$\hat{x}_{k|k-1} = F_k \hat{x}_{k-1|k-1} + B_k u_k \quad \textbf{(6.19)}$$

tahmin edilmiş kovaryans tahmini

$$P_{k|k-1} = F_k P_{k-1|k-1} F_k^\top + Q_k \quad \textbf{(6.20)}$$

şeklinde hesaplanır. Güncelleme aşamasında ise yenilik veya ölçüm öncesi uyum artığı,

$$\tilde{y}_k = z_k - H_k \tilde{x}_{k|k-1} \quad \textbf{(6.21)}$$

yenilik kovaryansı,

$$S_k = H_k P_{k|k-1} H_k^\top + R_k \quad \textbf{(6.22)}$$

optimal Kalman,

$$K_k = P_{k|k-1} H_k^\top S_k^{-1} \quad \textbf{(6.23)}$$

güncellenmiş durum tahmini,

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k \tilde{y}_k \quad \textbf{(6.24)}$$

güncellenmiş kovaryans tahmini,

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} \quad \textbf{(6.25)}$$

uydurma sonrası artık,

$$\tilde{y}_{k|k} = z_k - H_k\hat{x}_{k|k} \quad \textbf{(6.26)}$$

şeklinde hesaplanır.

Güncellenmiş sonsal tahmin kovaryansı için formül, artık hatayı en aza indiren optimal K_k kazancı için geçerlidir. $\hat{x}_{k|k}$ güncellenmiş durum tahmini, daha sezgisel bir şekilde

$$\hat{x}_{k|k} = (I - K_k H_k)\left(\hat{x}_{k|k-1}\right) + (K_k)(H_k x_k + v_k) \quad \textbf{(6.27)}$$

şeklinde ifade edilebilir. Bu ifade, $[0,1]$ aralığındaki t için $x = (1-t)(a) + t(b)$ doğrusal interpolasyonuna benzemektedir. Bu durumda t, 0 ile I arasında değer alan bir matris şeklindeki K_k Kalman kazancıdır. a, modelden tahmin edilen değerdir. b ise ölçümden elde edilen değerdir.

7. TRAFİK KAZASI VERİ SETİNDE EKSİK GÖZLEM TAMAMLAMA

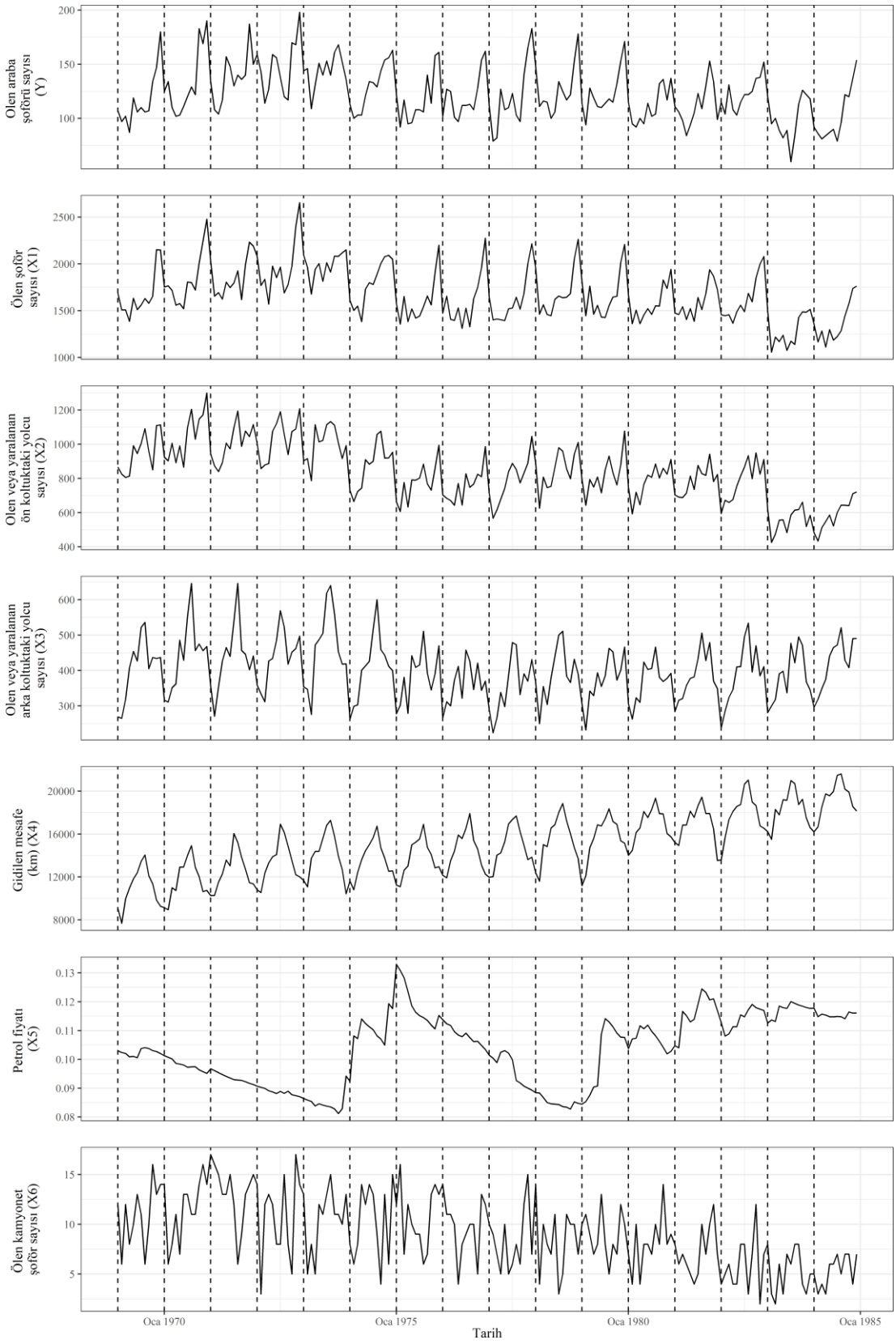
Bu çalışmada literatürde ulaşılması mümkün olan Seatbelts isimli veri seti kullanıldı (Harvey, 1989; Durbin ve Koopman, 2001). Bu veri seti 1969-1984 arasında Birleşik Krallık trafik kazaları hakkında aylık bilgileri içermektedir. Veri setinde bulunan değişkenler içerisinde bazı değişkenler kullanılmış, bazıları ise kullanılmamıştır. Kullanılan değişkenler Tablo 7.1'de açıklamalarıyla birlikte verilmiştir. Bağımlı değişken ölen araba şoförü sayısı, yani bir sayım verisidir. Veri setindeki Petrol Fiyatı değişkeni hariç diğer değişkenlerde mevsimsellik bulunmaktadır. Şekil 7.1'de değişkenlere ait çizgi grafikleri görülmektedir. Şekillerdeki dikey kesikli çizgiler, mevsimler arası kesim noktalarını göstermektedir. Ölen araba şoforü, şoför ve kamyonet şoförlerinin kış mevsiminde artış gösterip, yaz mevsiminde azaldığı görülmektedir.

Tablo 7.1. Veri setine ait değişkenler ve açıklamaları

Değişken	Açıklama
Y	Ölen araba şoforü sayısı
X_1	Ölen şoför sayısı
X_2	Ölen veya yaralanan ön oltuktaki yolcu sayısı
X_3	Ölen veya yaralanan arka oltuktaki yolcu sayısı
X_4	Gidilen mesafe (km)
X_5	Petrol fiyatı
X_6	Ölen kamyonet şoförü sayısı

Y değişkeninin Poisson dağılımına uygun olduğu varsayımı altında bahsedilen değişkenler ile Poisson zaman serisi modeli oluşturuldu. 0.1, 0.25, 0.5, 0.75 oranlarında her bir oran için 100 farklı olacak şekilde kayıp veri içeren veri setleri oluşturuldu. Oluşturulan tüm veri setlerinde kayıp veri türü MAR olarak ayarlandı. Poisson zaman serisi için R programında glarma paketi (Dunsmuir ve Scott, 2015) içinde glarma fonksiyonu kullanıldı. Veri seti üzerinden kayıp veri seti oluştururken mice paketindeki (Buuren ve Groothuis-Oudshoorn, 2010) ampute fonksiyonu kullanıldı. Ardından bu yöntemler farklı kayıp veri doldurma yöntemleri kullanılarak dolduruldu. Uygulamada kullanılan kayıp veri doldurma yöntemleri kullanım şekilleri

ile birlikte Tablo 7.2'de görülmektedir. Veri seti mevsimsellik içerdiği için kayıp veri doldurma yöntemleri, her bir mevsim yani her yıl için ayrı ayrı uygulanmıştır.



Şekil 7.1. Veri setindeki değişkenlerin zamana göre çizgi grafikleri

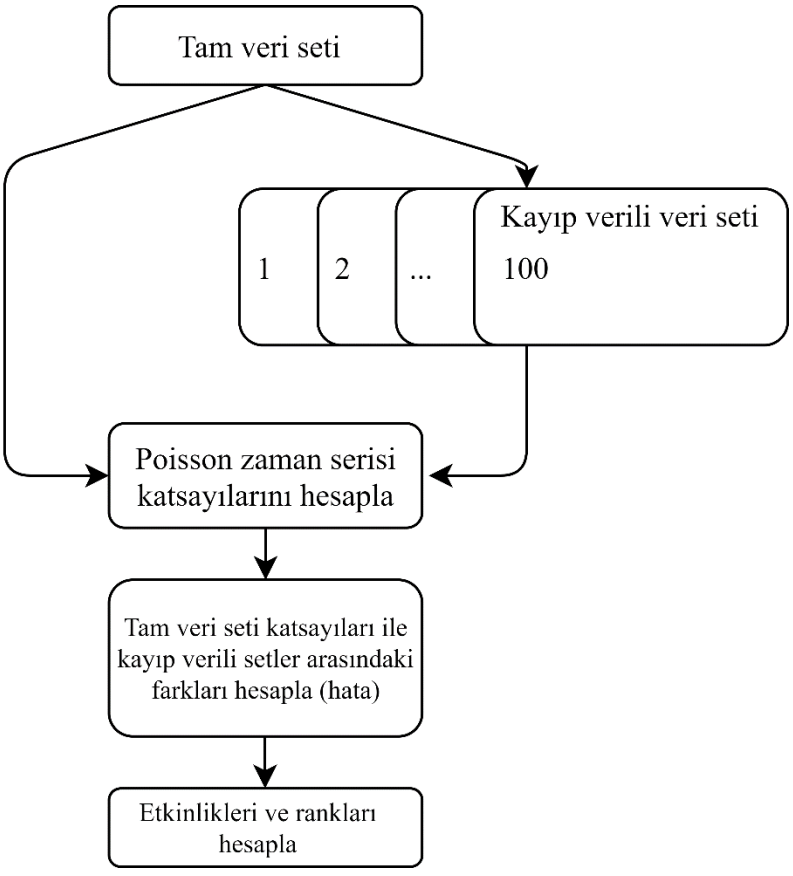
Tablo 7.2. Çalışmada kullanılan kayıp veri doldurma yöntemler ve kullanım şekilleri

Yöntem	Fonksiyon	Paket
MICE Random Forest	mice(…, method = “rf)	mice
MICE Midastouch	mice(…, method = “midastouch”)	mice
MICE PMM	mice(…, method = “pmm”)	mice
MI AMELIA	amelia(…)	amelia
Enterpolasyon	na_seadec(…, algorithm = “interpolation”)	imputeTS
LOCF	na_seadec(…, algorithm = “locf”)	imputeTS
Kalman	na_seadec(…, algorithm = “kalman”)	imputeTS
MA	na_seadec(…, algorithm = “ma”)	imputeTS

Şekil 7.2’de çalışmanın uygulama kısmını anlatan akış şeması görülmektedir. Poisson zaman serisi modellerinden katsayılar elde edildikten sonra, tam veri setine ait katsayılar ile kayıp veri içeren veri setlerinden elde edilen katsayılar arasındaki farklar hesaplanmış ve “hata” olarak kabul edilmiştir. Ardından gerçek katsayılar ile hatalardan faydalanılarak

$$Etkinlik = 100 - \left|\frac{hata}{gercek\ katsayı}\right| \times 100$$

şeklinde etkinlik değerleri elde edilmiştir. Burada elde edilen etkinlik, gerçek değere kıyasla yüzdelik doğru tahmin edilen katsayıya karşılık gelmektedir.



Şekil 7.2 Çalışmanın akış şeması

Tablo 7.3'te elde farklı yöntemler kullanılarak elde edilen Poisson zaman serisi modeli katsayıları görülmektedir. Tablo 7.4'te ise katsayılara ait hata miktarları görülmektedir. Burada kayıp veri doldurma yöntemi hücrelerindeki değerler, hesaplanan 100 farklı değerlerin ortalamalarıdır. Tablo 7.3 ve 7.4'te farklılıklar her ne kadar görülüyor olsa da buradan yorum çıkarmak zordur. Bu nedenle Tablo 7.5'te verilen etkinlikler hesaplanmıştır.

Tablo 7.3. Katsayılar Tablosu

Kayıp Oranı	Yöntem	X_1	X_2	X_3	X_4	X_5	X_6
Tam veri	Yok	8.38E-04	1.50E-03	-1.90E-03	7.37E-05	1.55E+01	1.58E-02
0.1	Enterpolasyon	8.52E-04	1.43E-03	-1.73E-03	7.14E-05	1.56E+01	1.56E-02
	Kalman	8.52E-04	1.43E-03	-1.74E-03	7.16E-05	1.55E+01	1.60E-02
	LOCF	8.52E-04	1.42E-03	-1.73E-03	7.14E-05	1.56E+01	1.55E-02
	MA	8.50E-04	1.43E-03	-1.73E-03	7.12E-05	1.56E+01	1.58E-02
	MI AMELIA	8.23E-04	1.55E-03	-2.03E-03	7.50E-05	1.56E+01	1.57E-02
	MICE Midastouch	8.45E-04	1.43E-03	-1.73E-03	6.97E-05	1.59E+01	1.51E-02
	MICE PMM	8.25E-04	1.53E-03	-1.96E-03	7.43E-05	1.56E+01	1.58E-02
	MICE Random Forest	8.55E-04	1.40E-03	-1.67E-03	6.98E-05	1.58E+01	1.52E-02
0.25	Enterpolasyon	8.73E-04	1.35E-03	-1.56E-03	6.89E-05	1.56E+01	1.51E-02
	Kalman	8.75E-04	1.35E-03	-1.56E-03	6.91E-05	1.55E+01	1.59E-02
	LOCF	8.62E-04	1.35E-03	-1.52E-03	6.79E-05	1.58E+01	1.50E-02
	MA	8.70E-04	1.35E-03	-1.53E-03	6.82E-05	1.56E+01	1.56E-02
	MI AMELIA	8.05E-04	1.61E-03	-2.12E-03	7.56E-05	1.58E+01	1.54E-02
	MICE Midastouch	8.51E-04	1.34E-03	-1.46E-03	6.27E-05	1.65E+01	1.40E-02
	MICE PMM	8.10E-04	1.56E-03	-1.99E-03	7.35E-05	1.58E+01	1.53E-02
	MICE Random Forest	8.65E-04	1.31E-03	-1.45E-03	6.57E-05	1.60E+01	1.50E-02
0.5	Enterpolasyon	9.00E-04	1.25E-03	-1.31E-03	6.48E-05	1.56E+01	1.41E-02
	Kalman	9.01E-04	1.25E-03	-1.31E-03	6.52E-05	1.55E+01	1.58E-02
	LOCF	8.77E-04	1.24E-03	-1.17E-03	6.25E-05	1.59E+01	1.42E-02
	MA	8.96E-04	1.25E-03	-1.28E-03	6.41E-05	1.56E+01	1.52E-02
	MI AMELIA	7.78E-04	1.74E-03	-2.38E-03	7.87E-05	1.58E+01	1.48E-02
	MICE Midastouch	8.85E-04	1.19E-03	-1.12E-03	5.64E-05	1.69E+01	1.33E-02
	MICE PMM	7.91E-04	1.61E-03	-2.05E-03	7.40E-05	1.59E+01	1.50E-02
	MICE Random Forest	8.98E-04	1.16E-03	-1.06E-03	5.84E-05	1.64E+01	1.43E-02
0.75	Enterpolasyon	9.18E-04	1.21E-03	-1.18E-03	6.29E-05	1.56E+01	1.34E-02
	Kalman	9.17E-04	1.21E-03	-1.20E-03	6.34E-05	1.54E+01	1.51E-02
	LOCF	8.69E-04	1.20E-03	-9.50E-04	5.93E-05	1.61E+01	1.34E-02
	MA	9.09E-04	1.22E-03	-1.14E-03	6.17E-05	1.56E+01	1.41E-02
	MI AMELIA	7.61E-04	1.83E-03	-2.57E-03	8.08E-05	1.58E+01	1.39E-02
	MICE Midastouch	8.84E-04	1.16E-03	-9.04E-04	4.94E-05	1.75E+01	1.16E-02
	MICE PMM	8.01E-04	1.61E-03	-2.02E-03	7.25E-05	1.61E+01	1.34E-02
	MICE Random Forest	8.95E-04	1.11E-03	-8.71E-04	5.50E-05	1.66E+01	1.30E-02

Tablo 7.4. Hatalar tablosu

Kayıp Oranı	Yöntem	X_1	X_2	X_3	X_4	X_5	X_6
0.1	Enterpolasyon	1.40E-05	-6.99E-05	1.62E-04	-2.24E-06	6.37E-02	-2.19E-04
	Kalman	1.34E-05	-6.93E-05	1.58E-04	-2.11E-06	3.62E-02	1.52E-04
	LOCF	1.36E-05	-7.31E-05	1.65E-04	-2.30E-06	1.07E-01	-3.18E-04
	MA	1.15E-05	-6.82E-05	1.69E-04	-2.46E-06	8.34E-02	-2.29E-05
	MI AMELIA	-1.51E-05	5.71E-05	-1.29E-04	1.36E-06	1.04E-01	-1.09E-04
	MICE Midastouch	7.03E-06	-6.60E-05	1.65E-04	-3.95E-06	4.12E-01	-7.71E-04
	MICE PMM	-1.32E-05	3.63E-05	-6.22E-05	5.71E-07	8.87E-02	-6.77E-05
	MICE Random Forest	1.73E-05	-9.97E-05	2.22E-04	-3.91E-06	2.87E-01	-6.66E-04
0.25	Enterpolasyon	3.49E-05	-1.44E-04	3.35E-04	-4.83E-06	9.03E-02	-7.57E-04
	Kalman	3.66E-05	-1.50E-04	3.35E-04	-4.63E-06	1.87E-02	6.64E-05
	LOCF	2.41E-05	-1.47E-04	3.79E-04	-5.82E-06	2.67E-01	-8.18E-04
	MA	3.23E-05	-1.50E-04	3.65E-04	-5.47E-06	1.15E-01	-2.05E-04
	MI AMELIA	-3.28E-05	1.14E-04	-2.27E-04	1.92E-06	2.63E-01	-4.37E-04
	MICE Midastouch	1.33E-05	-1.57E-04	4.40E-04	-1.10E-05	1.06E+00	-1.86E-03
	MICE PMM	-2.85E-05	6.88E-05	-9.63E-05	-2.00E-07	3.49E-01	-5.01E-04
	MICE Random Forest	2.73E-05	-1.82E-04	4.45E-04	-7.96E-06	5.17E-01	-8.39E-04

0.5	Enterpolasyon	6.22E-05	-2.41E-04	5.88E-04	-8.92E-06	1.44E-01	-1.77E-03
	Kalman	6.31E-05	-2.49E-04	5.82E-04	-8.47E-06	1.69E-02	-8.37E-06
	LOCF	3.87E-05	-2.59E-04	7.27E-04	-1.11E-05	4.49E-01	-1.67E-03
	MA	5.79E-05	-2.46E-04	6.14E-04	-9.55E-06	1.59E-01	-6.33E-04
	MI AMELIA	-6.02E-05	2.41E-04	-4.86E-04	5.05E-06	3.17E-01	-1.07E-03
	MICE Midastouch	4.65E-05	-3.03E-04	7.77E-04	-1.72E-05	1.39E+00	-2.56E-03
	MICE PMM	-4.71E-05	1.15E-04	-1.55E-04	3.04E-07	4.59E-01	-7.92E-04
	MICE Random Forest	6.01E-05	-3.39E-04	8.32E-04	-1.52E-05	8.67E-01	-1.49E-03
0.75	Enterpolasyon	7.94E-05	-2.86E-04	7.12E-04	-1.08E-05	8.30E-02	-2.44E-03
	Kalman	7.86E-05	-2.86E-04	6.98E-04	-1.03E-05	-6.08E-02	-6.94E-04
	LOCF	3.06E-05	-2.95E-04	9.47E-04	-1.44E-05	5.69E-01	-2.44E-03
	MA	7.03E-05	-2.80E-04	7.60E-04	-1.20E-05	1.35E-01	-1.71E-03
	MI AMELIA	-7.68E-05	3.35E-04	-6.69E-04	7.07E-06	3.50E-01	-1.88E-03
	MICE Midastouch	4.63E-05	-3.39E-04	9.92E-04	-2.43E-05	2.03E+00	-4.26E-03
	MICE PMM	-3.71E-05	1.13E-04	-1.23E-04	-1.17E-06	5.72E-01	-2.40E-03
	MICE Random Forest	5.69E-05	-3.81E-04	1.03E-03	-1.87E-05	1.13E+00	-2.87E-03

Etkinlik değerinin 100 olması, kusursuz tahmin yapıldığı anlamına gelmektedir. Bu değer, 100 veya daha küçük değer almaktadır. Yani, ne kadar küçük değer alırsa o kadar başarısız olunduğu anlamına gelmektedir. Tablo 7.5'teki etkinlikler incelendiğinde beklendiği üzere, 0.1 kayıp oranından 0.75 kayıp oranına doğru etkinliklerin azaldığı görülebilmektedir.

Tablo 7.5. Etkinlikler Tablosu

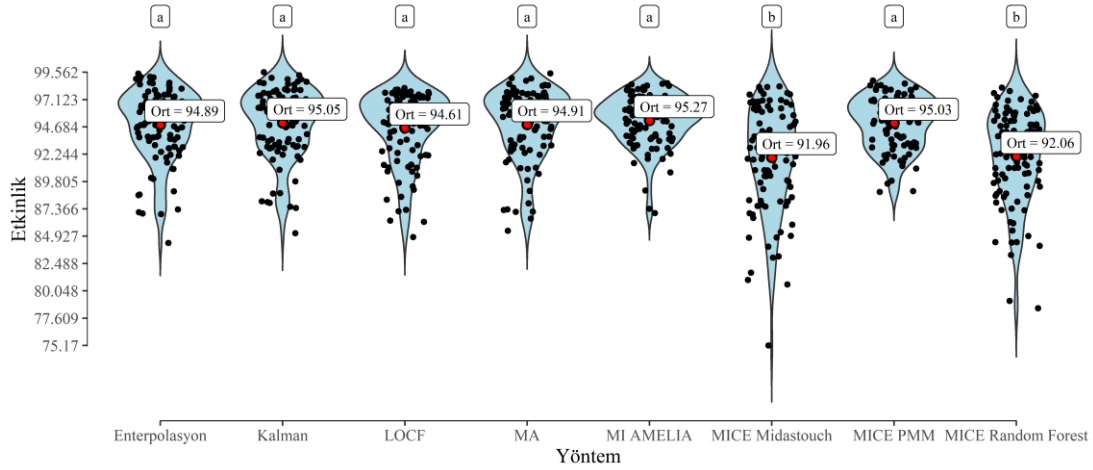
Kayıp Oranı	Yöntem	X_1	X_2	X_3	X_4	X_5	X_6
0.1	Enterpolasyon	97.283	93.896	90.058	96.047	98.815	93.236
	Kalman	97.335	94.058	90.238	96.168	98.889	93.621
	LOCF	96.717	93.386	90.260	96.195	98.657	92.452
	MA	97.304	93.949	89.917	95.922	98.783	93.577
	MI AMELIA	97.151	95.301	92.526	97.131	98.570	90.951
	MICE Midastouch	94.538	91.182	87.939	93.867	96.796	87.434
	MICE PMM	96.533	94.905	93.738	96.765	98.321	89.927
	MICE Random Forest	95.062	91.177	87.006	93.761	97.291	88.073
0.25	Enterpolasyon	94.918	89.488	82.188	93.109	98.126	87.954
	Kalman	94.850	89.280	82.102	93.321	98.036	90.427
	LOCF	94.408	88.729	79.981	91.852	97.611	86.469
	MA	94.946	89.058	80.587	92.413	97.949	88.846
	MI AMELIA	94.583	91.108	87.055	95.717	97.470	86.825
	MICE Midastouch	90.515	84.104	75.531	84.981	92.523	80.398
	MICE PMM	94.356	92.236	91.003	96.177	96.953	84.017
	MICE Random Forest	92.277	85.421	76.168	88.872	95.951	79.901
0.5	Enterpolasyon	92.005	83.362	68.895	87.835	97.498	81.432
	Kalman	92.046	83.017	69.295	88.476	97.595	84.567
	LOCF	93.184	82.458	61.669	84.721	96.274	80.772
	MA	92.581	83.279	67.626	87.043	97.256	82.899
	MI AMELIA	91.045	83.047	73.627	91.835	96.537	79.876
	MICE Midastouch	88.580	76.430	58.541	76.495	90.726	72.341
	MICE PMM	90.671	87.709	85.180	93.136	95.774	79.090
	MICE Random Forest	90.770	77.159	56.130	79.234	93.407	74.903
0.75	Enterpolasyon	90.364	80.781	62.458	85.145	96.951	75.709
	Kalman	90.441	80.614	63.173	85.827	96.936	81.165
	LOCF	92.829	79.740	50.075	80.427	95.648	75.503
	MA	91.197	80.817	59.955	83.614	96.539	77.579
	MI AMELIA	88.188	76.945	64.303	89.657	95.604	71.898
	MICE Midastouch	88.394	73.732	47.434	67.075	86.869	64.330
	MICE PMM	89.812	85.425	82.553	91.805	94.181	72.481
	MICE Random Forest	89.517	73.997	45.934	74.615	92.096	69.692

Yöntemlere ait etkinliklerin arasındaki farklılıkları daha iyi incelemek için ANOVA testleri uygulandı. Tablo 7.6'da kayıp oranı 0.1 için yöntemlerin ANOVA ve

çoklu karşılaştırma testi sonuçları verildi. Ortalamaya göre en başarılı yöntemin MI AMELIA yöntemi olduğu görülse de çoklu karşılaştırma sonuçlarında istatistiksel olarak Enterpolasyon, Kalman, LOCF, MA, MI AMELIA ve MICE PMM yöntemleri arasında anlamlı farklılık görülmedi. Şekil 7.3'te kayıp oranı 0.1 için 100 iterasyona ait keman grafiği görülmektedir. MICE Midastouch ve MICE Random Forest yöntemlerinin etkinliklerinin 75'e kadar düşebildiği görülmektedir. MI AMELIA ve MICE PMM yöntemlerinin Enterpolasyon, Kalman, LOCF ve MA yöntemleri ile ortalamaları arasında anlamlı fark olmasa da standart sapmaları daha düşüktür. Şekil 7.3'te de görüldüğü gibi, düşük etkinlik seviyelerine inilmemiştir. Bu da 0.1 kayıp oranı için daha az riskli olduklarına bir işarettir.

Tablo 7.6. Kayıp oranı 0.1 için yöntemlerin etkinlik değleri ANOVA tablosu

	Yöntem	Ort.	SS	Test İst.	p	Grup
Etkinlik	Enterpolasyon	94.889	3.166	12.094	0.000	a
	Kalman	95.051	3.073			a
	LOCF	94.611	3.138			a
	MA	94.909	3.179			a
	MI AMELIA	**95.272**	2.250			a
	MICE Midastouch	91.959	4.681			b
	MICE PMM	95.031	2.361			a
	MICE Random Forest	92.062	4.122			b



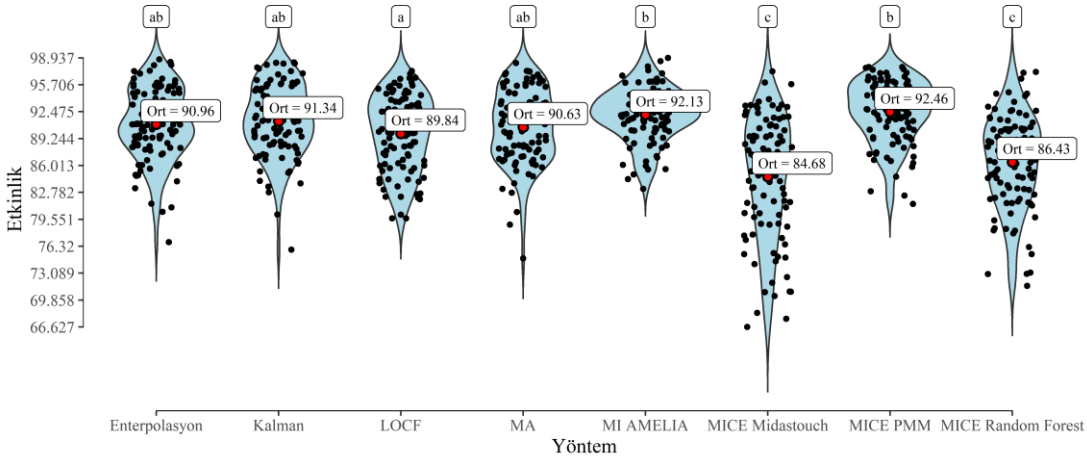
Şekil 7.3. Kayıp veri oranı 0.1 için etkinlik ölçülerinin 100 iterasyona ait keman grafiği

Tablo 7.7'de kayıp oranı 0.25 için yöntemlerin ANOVA ve çoklu karşılaştırma testi sonuçları verildi. Ortalamaya göre en başarılı yöntemin MI PMM yöntemi olduğu görülse de çoklu karşılaştırma sonuçlarında istatistiksel olarak Enterpolasyon, Kalman, MA, MI AMELIA ve MICE PMM yöntemleri arasında anlamlı farklılık görülmedi. Şekil 7.4'te kayıp oranı 0.25 için 100 iterasyona ait keman grafiği görülmektedir. MICE Midastouch ve MICE Random Forest yöntemlerinin

etkinliklerinin 66'ya kadar düşebildiği görülmektedir. Enterpolasyon, Kalman, MA, MI AMELIA ve MICE PMM yöntemleri arasında anlamlı fark bulunmasa da MI AMELIA yönteminin standart sapması diğer yöntemlerden daha azdır. Bu da 0.25 kayıp oranı için daha az riskli bir yöntem olduğuna işarettir.

Tablo 7.7. Kayıp oranı 0.25 için yöntemlerin etkinlik değleri ANOVA tablosu

Olcu	Grup	Ort.	SS	Test İst.	p	
Etkinlik	Enterpolasyon	90.964	4.335	24.870	0.000	ab
	Kalman	91.336	4.313			ab
	LOCF	89.842	4.504			a
	MA	90.633	4.471			ab
	MI AEMLIA	92.126	3.198			b
	MICE Midastouch	84.675	7.250			c
	MICE PMM	**92.457**	3.686			b
	MICE Random Forest	86.432	5.587			c

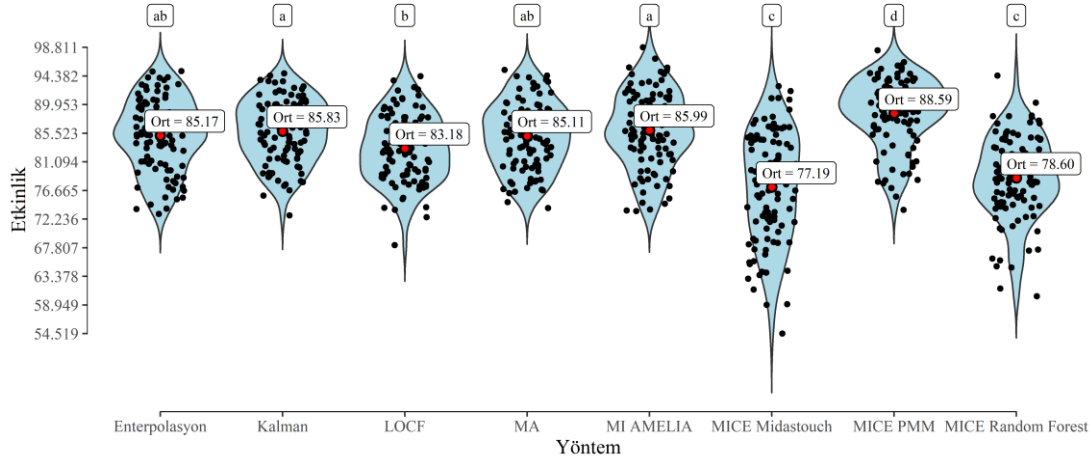


Şekil 7.4. Kayıp veri oranı 0.25 için etkinlik ölçülerinin 100 iterasyona ait keman grafiği

Tablo 7.8'de kayıp oranı 0.50 için yöntemlerin ANOVA ve çoklu karşılaştırma testi sonuçları verildi. Ortalamaya göre en başarılı yöntemin MI PMM yöntemi olduğu görüldü. Şekil 7.5'te kayıp oranı 0.50 için 100 iterasyona ait keman grafiği görülmektedir. Çoklu karşılaştırma sonuçlarında da MI PMM yönteminin diğer yöntemlerden anlamlı derecede farklı olduğu görüldü.

Tablo 7.8. Kayıp oranı 0.50 için yöntemlerin etkinlik değleri ANOVA tablosu

Olcu	Grup	Ort.	SS	Test İst.	p	Grup
Etkinlik	Enterpolasyon	85.171	5.566	34.100	0.000	ab
	Kalman	85.833	4.889			a
	LOCF	83.180	5.200			b
	MA	85.114	5.185			ab
	MI AMELIA	85.994	5.869			a
	MICE Midastouch	77.186	8.493			c
	MICE PMM	**88.593**	5.323			d
	MICE Random Forest	78.601	6.398			c

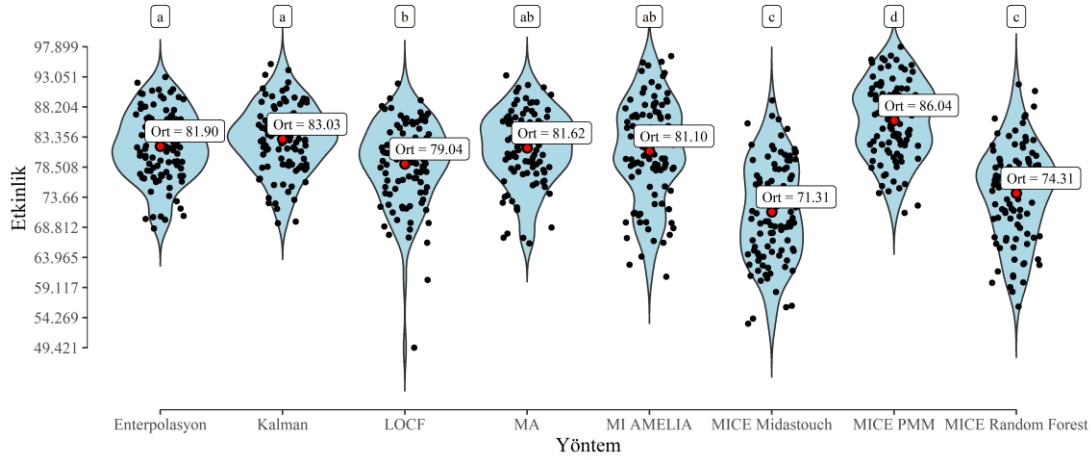


Şekil 7.5. Kayıp veri oranı 0.50 için etkinlik ölçülerinin 100 iterasyona ait keman grafiği

Tablo 7.9'de kayıp oranı 0.75 için yöntemlerin ANOVA ve çoklu karşılaştırma testi sonuçları verildi. Ortalamaya göre en başarılı yöntemin MI PMM yöntemi olduğu görüldü. Şekil 7.6'de kayıp oranı 0.75 için 100 iterasyona ait keman grafiği görülmektedir. Çoklu karşılaştırma sonuçlarında da MI PMM yönteminin diğer yöntemlerden anlamlı derecede farklı olduğu görüldü.

Tablo 7.9. Kayıp oranı 0.75 için yöntemlerin etkinlik değleri ANOVA tablosu

Olcu	Grup	Ort.	SS	Test İst.	p	
Etkinlik	Enterpolasyon	81.901	5.621	44.248	0.000	a
	Kalman	83.026	5.418			a
	LOCF	79.037	6.760			b
	MA	81.617	5.938			ab
	MI AMELIA	81.099	8.121			ab
	MICE Midastouch	71.306	7.989			c
	MICE PMM	**86.043**	6.199			d
	MICE Random Forest	74.308	7.686			c

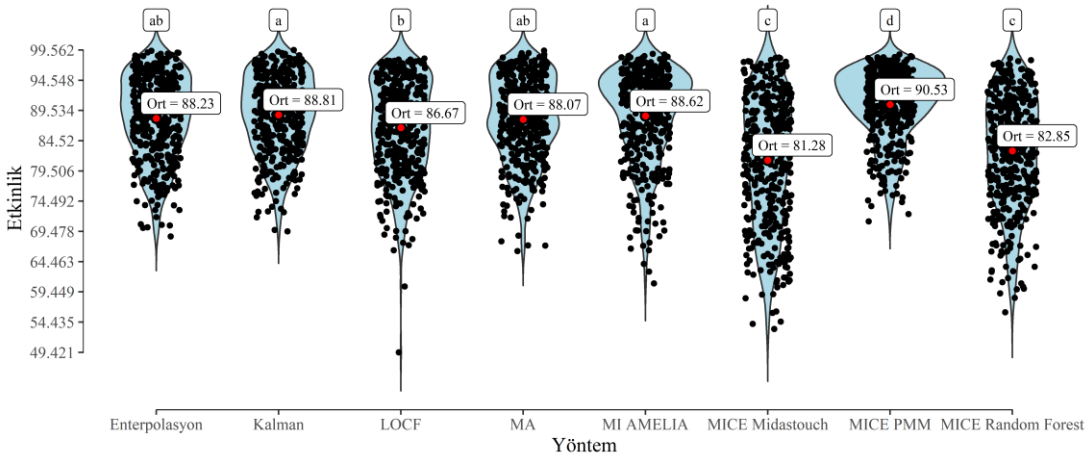


Şekil 7.6. Kayıp veri oranı 0.75 için etkinlik ölçülerinin 100 iterasyona ait keman grafiği

Tablo 7.10'da tüm kayıp oranları için yöntemlerin ANOVA ve çoklu karşılaştırma testi sonuçları verildi. Ortalamaya göre en başarılı yöntemin MI PMM yöntemi olduğu görüldü. Şekil 7.7'de tüm kayıp oranları için 100 iterasyona ait keman grafiği görülmektedir. Çoklu karşılaştırma sonuçlarında da MI PMM yönteminin diğer yöntemlerden anlamlı derecede farklı olduğu görüldü. MI PMM yöntemi en yüksek etkinlik ortalamasına sahip olduğu gibi en düşük varyansa da sahip yöntemdir. En kötü performans gösteren yöntemler ise MICE Midastouch ve MICE Random Forest yöntemleridir.

Tablo 7.10. Tüm kayıp oranları için yöntemlerin etkinlik değleri ANOVA tablosu

Olcu	Grup	Ort.	SS	Test İst.	p	Grup
Etkinlik	Enterpolasyon	88.231	6.932	54.350	0.000	ab
	Kalman	88.812	6.492			a
	LOCF	86.667	7.843			b
	MA	88.068	6.992			ab
	MI AMELIA	88.623	7.667			a
	MICE Midastouch	81.281	10.622			c
	MICE PMM	**90.531**	5.772			d
	MICE Random Forest	82.851	9.169			c



Şekil 7.7. Tüm kayıp oranları için etkinlik ölçülerinin 100 iterasyona ait keman grafiği

Tablo 7.5'teki etkinlikler kullanılarak Tablo 7.11'deki rank değerleri elde edilmiştir. Burada amacımız değişkenlerin katsayısını hesaplarken hangi yöntemin daha başarılı olduğunu görmektir. Kırmızı renkli hücreler, diğer yöntemlere kıyasla başarısız olan yöntemleri, beyaz hücreler ise başarılı hücreleri göstermektedir. Yani rank numaraları başarılıdan başarısıza, küçükten büyüğe olacak şekilde kodlanmıştır. Ortalamalara bakıldığında en iyi yöntem olarak görülen MICE PMM yönteminin X_1, X_5 ve X_6 değişkenlerinin katsayıları konusunda diğer yöntemlere kıyasla başarısız

olduğu görülmektedir. Enterpolasyon ve özellikle Kalman yönteminin diğer yöntemlere göre daha genellenebildiği görülmektedir.

Tablo 7.11. Rank tablosu

Kayıp Oranı	Yöntem	X_1	X_2	X_3	X_4	X_5	X_6
0.1	Enterpolasyon	3	5	5	5	2	3
	Kalman	1	3	4	4	1	1
	LOCF	5	6	3	3	4	4
	MA	2	4	6	6	3	2
	MI AMELIA	4	1	2	1	5	5
	MICE Midastouch	8	7	7	7	8	8
	MICE PMM	6	2	1	2	6	6
	MICE Random Forest	7	8	8	8	7	7
0.25	Enterpolasyon	2	3	3	4	1	3
	Kalman	3	4	4	3	2	1
	LOCF	5	6	6	6	4	5
	MA	1	5	5	5	3	2
	MI AMELIA	4	2	2	2	5	4
	MICE Midastouch	8	8	8	8	8	7
	MICE PMM	6	1	1	1	6	6
	MICE Random Forest	7	7	7	7	7	8
0.5	Enterpolasyon	4	2	4	4	2	3
	Kalman	3	5	3	3	1	1
	LOCF	1	6	6	6	5	4
	MA	2	3	5	5	3	2
	MI AMELIA	5	4	2	2	4	5
	MICE Midastouch	8	8	7	8	8	8
	MICE PMM	7	1	1	1	6	6
	MICE Random Forest	6	7	8	7	7	7
0.75	Enterpolasyon	4	3	4	4	1	3
	Kalman	3	4	3	3	2	1
	LOCF	1	5	6	6	4	4
	MA	2	2	5	5	3	2
	MI AMELIA	8	6	2	2	5	6
	MICE Midastouch	7	8	7	8	8	8
	MICE PMM	5	1	1	1	6	5
	MICE Random Forest	6	7	8	7	7	7

Tablo 7.11'i daha iyi yorumlayabilmek için bu değerlerin yöntemler ve kayıp oranlarına göre ortalamaları hesaplanarak Tablo 7.12'de verilmiştir. Önceki karşılaştırmalar parametrik sonuçları gösteriyor iken bu değerler parametrik olmayan sonuçları gösteriyor denilebilir. Tablo 7.11'de sezilen sonuçların doğru olduğu Tablo 7.12'de görülmektedir. Kalman yönteminin rank ortalamaları tüm kayıp oranları için en yüksek başarıyı göstermiştir. MICE PMM yönteminin genel rank ortalaması

bakımından sekiz yöntem arasında dördüncü olduğu dikkat çekmektedir. Bunun nedeni MICE PMM yönteminin başarılı performans gösterdiği değişkenlerde çok etkili, başarısız olduğu değişkenlerde ise çok başarısız olmasıdır. Bu da genel rank ortalamasında geride kalmasına neden olmaktadır. En başarısız yöntem ise 7.708 rank ortalaması ile MICE Midastouch yöntemidir.

Tablo 7.12. Yöntemlere ve kayıp oranlarına göre rank ortalamaları

Yöntem	Kayıp oranı				
	0.1	0.25	0.5	0.75	Ort
Enterpolasyon	3.833	**2.667**	3.167	3.167	3.208
Kalman	**2.333**	2.833	**2.667**	**2.667**	**2.625**
LOCF	4.167	5.333	4.667	4.333	4.625
MA	3.833	3.500	3.333	3.167	3.458
MI AMELIA	3.000	3.167	3.667	4.833	3.667
MICE Midastouch	7.500	7.833	7.833	7.667	7.708
MICE PMM	3.833	3.500	3.667	3.167	3.542
MICE Random Forest	7.500	7.167	7.000	7.000	7.167

8. SONUÇ VE ÇIKARIMLAR

Sayı verileri, zaman serilerinde sıklıkla karşılaşılan bir veri türüdür. Sayı verileri kesikli dağılıma sahip olarak görülebilir. Bu nedenle sayı verileri için model kurarken, genellikle Poisson dağılımı gibi dağılımlardan faydalanılır. Bu yönteme Poisson zaman serisi yöntemi denir.

Poisson zaman serisi, bağımsız değişkenler kullanılarak daha verimli hale getirilebilir. Fakat, sayı verisini etkileyen değişkenleri toplayabilmek her zaman mümkün olamamaktadır. Bu durumlarda genellikle, ya kayıp verili örnekler veri setinden çıkarılmakta veya fazla kayıp veri oranına sahip değişken modelden atılmaktadır. Poisson zaman serisi için kayıp veri doldurma yöntemleri kullanılarak daha verimli sonuçlar elde edilebilir.

Çalışmada sayım verisi, zaman serisi ve kayıp veri kavramları tanıtılmış ve bu kavramlarla ilgili yöntemlere değinilmiştir. Sayım verilerinde zaman serisi yöntemlerinden birisi olan Poisson zaman serisine uygun bir veri seti ele alınarak zaman serilerinde kullanılan kayıp veri doldurma yöntemleri karşılaştırılmıştır. Bu karşılaştırma için gerçek bir veri seti alınmış ve 0.1, 0.25, 0.50, 0.75 kayıp oranları ile kayıp verili veri setleri oluşturulmuştur. Sekiz farklı kayıp gözlem doldurma yöntemlerine (Enterpolasyon, Kalman, LOCF, MA, MI AMELIA, MICE Midastouch, MICE PMM ve MICE Random Forest) ait etkinlik ortalamalarına göre sonuçlar şöyledir;

- 0.10 kayıp oranı için yapılan istatistiksel test sonucunda ortalama etkinlik değerleri arasında anlamlı farklılık bulunmuştur. Etkinlik ortalamalarına göre en iyi yöntem Enterpolasyon, Kalman, LOCF, MA, MI AMELIA ve MICE PMM yöntemidir. En başarısız yöntemler ise MICE Midastouch ve MICE Random Forest yöntemidir.
- 0.25 kayıp oranı için yapılan istatistiksel test sonucunda ortalama etkinlik değerleri arasında anlamlı farklılık bulunmuştur. Etkinlik ortalamalarına göre en iyi yöntem Enterpolasyon, Kalman, MA, MI AMELIA ve MICE PMM yöntemidir. En başarısız yöntemler ise MICE Midastouch ve MICE Random Forest yöntemidir.
- 0.50 kayıp oranı için yapılan istatistiksel test sonucunda ortalama etkinlik değerleri arasında anlamlı farklılık bulunmuştur. Etkinlik ortalamalarına göre

en iyi yöntem MICE PMM yöntemidir. En başarısız yöntemler ise MICE Midastouch ve MICE Random Forest yöntemidir.

- 0.75 kayıp oranı için yapılan istatistiksel test sonucunda ortalama etkinlik değerleri arasında anlamlı farklılık bulunmuştur. Etkinlik ortalamalarına göre en iyi yöntem MICE PMM yöntemidir. En başarısız yöntemler ise MICE Midastouch ve MICE Random Forest yöntemidir.
- Tüm kayıp oranları için yapılan istatistiksel test sonucunda ortalama etkinlik değerleri arasında anlamlı farklılık bulunmuştur. Etkinlik ortalamalarına göre en iyi yöntem MICE PMM yöntemidir. En başarısız yöntemler ise MICE Midastouch ve MICE Random Forest yöntemidir.

Sekiz farklı kayıp gözlem doldurma yöntemlerine ait rank ortalamalarına göre sonuçlar şöyledir;

- 0.10 kayıp oranı için en iyi yöntem Kalman yöntemi, en kötü yöntemler ise MICE Midastouch ve MICE Random Forest yöntemidir.
- 0.25 kayıp oranı için en iyi yöntem Enterpolasyon yöntemi, en kötü yöntemler ise MICE Midastouch yöntemidir.
- 0.50 kayıp oranı için en iyi yöntem Kalman yöntemi, en kötü yöntemler ise MICE Midastouch yöntemidir.
- 0.75 kayıp oranı için en iyi yöntem Kalman yöntemi, en kötü yöntemler ise MICE Midastouch yöntemidir.
- Tüm kayıp oranları için en iyi yöntem Kalman yöntemi, en kötü yöntemler ise MICE Midastouch yöntemidir.

Hem etkinlik ortalaması hem de sıra ortalamalarına göre MICE Midastouch ve MICE Random Forest yöntemlerinin kötü performans sergiledikleri gözükmektedir. Poisson zaman serisinde bağımsız değişkenlerdeki kayıp verilerin doldurulmasında bu iki yöntemin kullanılması önerilmemektedir. Alternatif yöntemler daha başarılı performans göstermektedir. Etkinlik ortalamaları ve rank ortalamalarına bakıldığında MICE PMM ve Kalman yöntemlerinin başarısı ortaya çıkmıştır. Bu çalışmadaki sonuçlara bakılarak Poisson zaman serisinde kayıp gözlem durumunda MICE PMM ve Kalman kayıp veri doldurma yöntemlerini önermektedir.

KAYNAKLAR

Allan, F. E., & Wishart, J. 1930. A method of estimating the yield of a missing plot in field experimental work. *The Journal of Agricultural Science*, 20:3, 399-406.

Andreasen, M. M. 2008. Non-linear DSGE models, the central difference Kalman filter, and the mean shifted particle filter. *CREATES Research Paper*, 33.

Barnard, J., & Meng, X. L. 1999. Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical methods in medical research*, 8:1, 17-36.

Benmouzi K, Cheknane A. Small-scale solar radiation forecasting using ARMA and nonlinear autoregressive neural network models. *Theor Appl Climatol* 2016; 124:945–584

Bortkiewicz, L. V. 1898. Die Grenznutzentheorie als Grundlage einer ultraliberalen Wirtschaftspolitik. *Schmollers Jahrbuch*, 22, 1177-216.

Burton, A., & Altman, D. G. 2004. Missing covariate data within cancer prognostic studies: a review of current reporting ve proposed guidelines. *British journal of cancer*, 91:1, 48.

Buuren, S. V., & Groothuis-Oudshoorn, K. 2010. mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 1-68.

Cameron, A. C., & Trivedi, P. K. 2013. *Regression analysis of count data* . Cambridge university press.

CHAPRA, S., & Canale, R. P. 1998. Numerical Methods for Engineers: with Programing and Software Applications.

Chou, Y. L. 1975. section 17:9, Statistical Analysis. *Holt International*.

David, M., Little, R. J., Samuhel, M. E., & Triest, R. K. 1986. Alternative methods for CPS income imputation. *Journal of the American Statistical Association*, 81:393, 29-41.

De Silva, N. R., Brooker, S., Hotez, P. J., Montresor, A., Engels, D., & Savioli, L. 2003. Soil transmitted helminth infections: updating the global picture. *Trends in parasitology*, 19:12, 547-551.

Dempster, A. P., & Rubin, D. B. 1983. Rounding error in regression: The appropriateness of Sheppard's corrections. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45:1, 51-59.

Dempster, A. P., N. M. Laird, ve D. B. Rubin. 1977. "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm (with Discussion)." *Journal of the Royal Statistical Society B* 39:1: 1–38.

Diaz-Ordaz, K., Kenward, M. G., Cohen, A., Coleman, C. L., & Eldridge, S. 2014. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clinical Trials*, 11:5, 590-600.

Dunsmuir, W. T., & Scott, D. J. 2015. The glarma package for observation-driven time series regression of counts. *Journal of Statistical Software*, 67:7, 1-36.

Durbin, J. and Koopman, S. J. 2001. *Time Series Analysis by State Space Methods*. Oxford University Press.

Eggenberger, F., & Pólya, G. 1923. Über die statistik verketteter vorgänge. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 3:4, 279-289.

Gelfand, A. E., & Smith, A. F. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85:410, 398-409.

Gelman, A., & Hill, J. 2006. *Data analysis using regression and multilevel/hierarchical*

models. Cambridge university press.

Gourieroux, C., Monfort, A., & Trognon, A. 1984. Pseudo maximum likelihood methods: Theory. *Econometrica: journal of the Econometric Society*, 681-700.

Greenwood, M., & Yule, G. U. 1920. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal statistical society*, 83:2, 255-279.

Hamer, R. M., & Simpson, P. M. 2009. Last observation carried forward versus mixed models in the analysis of psychiatric clinical trials.

Hansun, S. 2013. A new approach of moving average method in time series analysis. In *2013 conference on new media studies (CoNMedia)* pp. 1-4. IEEE.

Harvey, A.C. 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, pp. 519–523.

Hausman, J. A., Hall, B. H., & Griliches, Z. 1984. Econometric models for count data with an application to the patents-R&D relationship.

Humpherys, J., Redd, P., & West, J. 2012. A fresh look at the Kalman filter. *SIAM review*, 54:4, 801-823.

Hunter, J. S. 1986. The exponentially weighted moving average. *Journal of quality technology*, 18:4, 203-210.

Hussain S, Al Alili A. Day ahead hourly forecast of solar irradiance for Abu Dhabi, UAE. In: *The 4th IEEE international conference on smart energy grid engineering*; 2016. p. 68–71.

Ishihara, J. Y., Terra, M. H., & Campos, J. C. 2006. Robust Kalman filter for descriptor systems. *IEEE Transactions on Automatic Control*, 51:8, 1354-1354.

Jeličić, H., Phelps, E., & Lerner, R. M. 2009. Use of missing data methods in longitudinal studies: the persistence of bad practices in developmental psychology. *Developmental psychology*, 45:4, 1195.

Kalman, R. E. 1960. A new approach to linear filtering and prediction problems.

Kalman, R. E., & Bucy, R. S. 1961. New results in linear filtering and prediction theory.

Kalton, G. 1986. The treatment of missing survey data. *Survey methodology*, 12, 1-16.

Kelly, A. 1994. *A 3D state space formulation of a navigation Kalman filter for autonomous vehicles*. carnegıe-mellon unıv pıttsburgh pa robotıcs ınst.

Klebanoff, M. A., & Cole, S. R. 2008. Use of multiple imputation in the epidemiologic literature. *American journal of epidemiology*, 168:4, 355-357.

Lachin, J. M. 2016. Fallacies of last observation carried forward analyses. *Clinical trials*, 13:2, 161-168.

Little, R. J., & Rubin, D. B. 2019. *Statistical analysis with missing data*. John Wiley & Sons.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., ... & Stern, H. 2012. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367:14, 1355-1360.

MacKinnon, J. G. 2010. *Critical values for cointegration tests*. Queen's Economics Department Working Paper.

Maddala, G. S. 1983. Methods of estimation for models of markets with bounded price variation. *International Economic Review*, 361-378.

McCullagh, P., ve J. A. Nelder. 1989. Generalized Linear Models. 2nd ed. New York: Chapman & Hall.

Molenberghs, G., & Kenward, M. 2007. *Missing data in clinical studies* (Vol. 61). John Wiley & Sons.

Nelder, J. A., & Wedderburn, R. W. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135:3, 370-384.

Noor, M. N., Yahaya, A. S., Ramli, N. A., & Al Bakri, A. M. M. 2014. *Filling missing data using interpolation methods: Study on the effect of fitting distribution* (Vol. 594, pp. 889-895). Trans Tech Publications Ltd.

Peugh, J. L., & Enders, C. K. 2004. Missing data in educational research: A review of reporting practices ve suggestions for improvement. *Review of educational research*, 74:4, 525-556.

Rijnhart, J. J., Twisk, J. W., Eekhout, I., & Heymans, M. W. 2019. Comparison of logistic-regression based methods for simple mediation analysis with a dichotomous outcome variable. *BMC medical research methodology*, 19:1, 1-10.

Roberts, G. 1998. Competitive altruism: from reciprocity to the handicap principle. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265:1394, 427-431.

Poisson, S. D. 1837. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*. Bachelier.

Powney, M., Williamson, P., Kirkham, J., & Kolamunnage-Dona, R. 2014. A review of the handling of missing longitudinal outcome data in clinical trials. *Trials*, 15:1, 1-11.

Rubin, D. B. 1976. Inference ve missing data. *Biometrika*, 63:3, 581-592.

Rubin, D. B. 1996. Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91:434, 473-489.

Rubin, J. (Ed.). 1987. *Learner strategies in language learning*. Macmillan College.

Rubin, Z. 1970. Measurement of romantic love. *Journal of personality and social psychology*,16:2, 265.

Pudney, S. 1989. *Modelling individual choice*. Basil Blackwell, Oxford.

Schafer, J. L., & Graham, J. W. 2002. Missing data: our view of the state of the art.

Scheuren, F. 2005. Multiple imputation: How it began and continues. *The American Statistician*, 59:4, 315-319. Psychological methods, 7:2, 147.

Seker, S. E. Zaman Serisi Analizi 2015 (Time Series Analysis).

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. 2014. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American journal of epidemiology*, 179:6, 764-774.

Strid, I., & Walentin, K. 2009. Block Kalman filtering for large-scale DSGE models. *Computational Economics*, 277-304.

Swerling, P. 1958. *A proposed stagewise differential correction procedure for satellite tracking and prediciton*. Rand Corporation.

Terra, M. H., Cerri, J. P., & Ishihara, J. Y. 2014. Optimal robust linear quadratic regulator for systems subject to uncertainties. *IEEE Transactions on Automatic Control*, 2586-2591.

Van Buuren, S. 2018. *Flexible imputation of missing data.* CRC press.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. 2006. Fully

conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 1049-1064.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, ve the Gauss—Newton method. *Biometrika*, 439-447.

Wood, A. M., White, I. R., & Thompson, S. G. 2004. Are missing outcome data adequately hveled? A review of published rveomized controlled trials in major medical journals. *Clinical trials*, 368-376.

Yates, F. 1933. "The Analysis of Replicated Experiments When the Field Results Are Incomplete." *Empirical Journal of Experimental Agriculture*

Zhang, Y., Crittenden, J. C., Hve, D. W., & Perram, D. L. 1994. Fixed-bed photocatalysts for solar decontamination of water. *Environmental science & technology*, 435-442.

EKLER

EK 1. Uygulama R Kodları

```
##-------------------------------------------------------------
##                          paketler                          -
##-------------------------------------------------------------
library(tscount)
library(missCompare)
library(mice)
library(forecast)
library(Amelia)
library(imputeTS)
library(glarma)
##-------------------------------------------------------------

##-------------------------------------------------------------
##                            veri                            -
##-------------------------------------------------------------
data <- Seatbelts
y <- data[,"DriversKilled"]
x <- data[,c("drivers","front","rear","kms","PetrolPrice","VanKilled")]
n <- nrow(x)
p <- ncol(x)

var_names <- colnames(x)
##-------------------------------------------------------------


##-------------------------------------------------------------
##                    mevsimsellik grafigi                    -
##-------------------------------------------------------------
seasonality_plot <- function(y, n_seasonality){
  n <- length(y)
  n_plots <- ceiling(n/n_seasonality)

  start <- 1
  y_temp <- y[1:n_seasonality] - min(y[1:n_seasonality])
  plot(y_temp, col = "white", ylim = c(min(y_temp), max(y_temp)))
  while (Inf) {
    lines(y[start:(start + n_seasonality - 1)] - min(y[start:(start +
n_seasonality - 1)]))
    start <- start + n_seasonality
    if (start > n) {
      break
    }
  }
}
```

```
44.
45. seasonality_plot(y = y, n_seasonality = 12)
46. ##-----------------------------------------------------------------
47.
48. ##-----------------------------------------------------------------
49. ##                          Full model                            -
50. ##-----------------------------------------------------------------
51. model_tspoisson_full <- glarma(y = y, X = x, type = "Poi")
52. coefs_full <- data.frame(Variables = attr(model_tspoisson_full$delta,"names"),
53.                          Coef = model_tspoisson_full$delta)
54.
55. coefs_full <- c(model_tspoisson_full$delta)
56. names(coefs_full) <- attr(model_tspoisson_full$delta,"names")
57.
58. write.csv(x = coefs_full, file = "coefficients full model.csv")
59. ##-----------------------------------------------------------------
60.
61. ##-----------------------------------------------------------------
62. ##                       Calisma Frameworku                       -
63. ##-----------------------------------------------------------------
64. missing_rates <- c(0.1, 0.25, 0.5, 0.75)
65. imputation_methods_names <- c("MICE random forest",
66.                                "MICE Midastouch",
67.                                "MICE PMM",
68.                                "MI AMELIA",
69.                                "Decomposed interpolation",
70.                                "Decomposed LOCF",
71.                                "Decomposed kalman smoothing",
72.                                "Decomposed MA",
73.                                "Splitted interpolation",
74.                                "Splitted LOCF",
75.                                "Splitted kalman smoothing",
76.                                "Splitted MA")
77. iter <- 100
78. ##-----------------------------------------------------------------
79.
80.
81. ##-----------------------------------------------------------------
82. ##                          Fonksiyonlar                          -
83. ##-----------------------------------------------------------------
84. imputation_functions <- list(
85.   function(x_missing){
86.     complete(mice(data = x_missing, m = 10, method = "rf", printFlag = FALSE))
87.   },
88.   function(x_missing){
89.     complete(mice(data = x_missing, m = 10, method = "midastouch", printFlag =
    FALSE))
90.   },
```

```
91.   function(x_missing){
92.     complete(mice(data = x_missing, m = 10, method = "pmm", printFlag =
    FALSE))
93.   },
94.   function(x_missing){
95.     asd <- amelia(x = as.data.frame(x_missing), p2s = 0)
96.     asd <- Reduce("+", asd$imputations)/length(asd$imputations)
97.     return(asd)
98.   },
99.   function(x_missing){
100.          imputeTS::na_seadec(x = x_missing, algorithm = "interpolation")
101.        },
102.        function(x_missing){
103.          imputeTS::na_seadec(x = x_missing, algorithm = "locf")
104.        },
105.        function(x_missing){
106.          imputeTS::na_seadec(x = x_missing, algorithm = "kalman")
107.        },
108.        function(x_missing){
109.          imputeTS::na_seadec(x = x_missing, algorithm = "ma")
110.        },
111.        function(x_missing){
112.          imputeTS::na_seasplit(x = x_missing, algorithm = "interpolation")
113.        },
114.        function(x_missing){
115.          imputeTS::na_seasplit(x = x_missing, algorithm = "locf")
116.        },
117.        function(x_missing){
118.          imputeTS::na_seasplit(x = x_missing, algorithm = "kalman")
119.        },
120.        function(x_missing){
121.          imputeTS::na_seasplit(x = x_missing, algorithm = "ma")
122.        }
123.      )
124.      names(imputation_functions) <- imputation_methods_names
125.      ##-------------------------------------------------------------------
126.
127.
128.      ##-------------------------------------------------------------------
129.      ##                              grid                               -
130.      ##-------------------------------------------------------------------
131.      grid <- expand.grid(imputation_methods = imputation_methods_names,
132.                          iter = 1:100,
133.                          missing_rates = missing_rates)
134.      ##-------------------------------------------------------------------
135.
136.      ##-------------------------------------------------------------------
137.      ##                        all missing datasets                       -
```

```
##-------------------------------------------------------------------
x_missingler <- lapply(missing_rates,
                       function(m) lapply(1:iter,
                                          function(m2) {
                                            cat(m, m2, "\r")
                                            mice::ampute(data = x,
                                                         prop = m)$amp
                                            }))
##-------------------------------------------------------------------


##-------------------------------------------------------------------
##                              START                               -
##-------------------------------------------------------------------
coefs_estimations_grid <- data.frame(matrix(NA, ncol = p, nrow =
nrow(grid)))
colnames(coefs_estimations_grid) <- var_names

for (i in 1:nrow(grid)) {
  grid_selected <- grid[i,]
  x_selected <- x_missingler[[which(missing_rates ==
grid_selected$missing_rates)]][[grid_selected$iter]]

  x_imputed <-
imputation_functions[[grid_selected$imputation_methods]](x_missing =
x_selected)
  x_imputed <- ts(x_imputed, start = 1969, frequency = 12)

  model <- glarma(y = y, X = x_imputed, type = "Poi")
  coefs_estimations_grid[i,] <- c(model$delta)

  process <- i/nrow(grid)
  cat("\r",
      " ",
      "|",
      rep("-",floor(50*process)),
      rep("=",ceiling(50 - 50*process)),
      "|    ",
      " %",
      formatC(process*100, digits = 3, format = "f"),
      sep = "")
}
##-------------------------------------------------------------------

write.csv(x = coefs_estimations_grid, file = "furkan tez seatbelt
katsayilar.csv")
write.csv(x = grid, file = "grid.csv")

```

```
181.        library(xlsx)
182.
183.        sonuclar <- read.xlsx(file = "furkan tez seatbelt katsayilar.xlsx",
    sheetIndex = 1, rowIndex = 1)
184.
185.        Sonuclar
```

ÖZ GEÇMİŞ

Adı ve Soyadı	: Furkan KOÇAL
E-Posta	: iletisim.fk@gmail.com
Yabancı Dili	: İngilizce

Eğitim Durmu

Lise	: Şehit Üsteğmen İbrahim Abanoz Lisesi
Lisans	: Ondokuz Mayıs Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü
Yüksek Lisans	: Ondokuz Mayıs Üniversitesi Lisansüstü Eğitim Enstitüsü İstatistik Anabilim Dalı

Çalıştığı Kurum

Samsun Üniversitesi