BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİMDALI
MÜHENDİSLİK VE TEKNOLOJİ YÖNETİMİ TEZLİ YÜKSEK LİSANS PROGRAMI

ZAMAN SERİSİ TAHMİN MODELLERİNDE VERİ ANALİZİ VE MODEL SEÇİMİ

YÜKSEK LİSANS TEZİ

HAZIRLAYAN
SENA NUR GÖREN

ANKARA- 2020

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİMDALI
MÜHENDİSLİK VE TEKNOLOJİ YÖNETİMİ TEZLİ YÜKSEK LİSANS PROGRAMI

ZAMAN SERİSİ TAHMİN MODELLERİNDE VERİ ANALİZİ VE MODEL SEÇİMİ

YÜKSEK LİSANS TEZİ

HAZIRLAYAN
SENA NUR GÖREN

TEZ DANIŞMANI
DR. MEHMET GÜLŞEN

ANKARA-2020

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
YÜKSEK LİSANS TEZ ÇALIŞMASI ORİJİNALLİK RAPORU

Öğrencinin Adı, Soyadı: Sena Nur GÖREN

Öğrencinin Numarası: 21620239

Anabilim Dalı: Mühendislik Fakültesi Endüstri Mühendisliği

Programı: Mühendislik ve Teknoloji Yönetimi

Danışmanın Unvanı/Adı, Soyadı: Dr. Mehmet GÜLŞEN

Tez Başlığı: Zaman Serisi Tahmin Modellerinde Veri Analizi ve Model Seçimi

Yukarıda başlığı belirtilen Yüksek Lisans tez çalışmamın; Giriş, Ana Bölümler ve Sonuç Bölümünden oluşan, toplam 33 sayfalık kısmına ilişkin, 24 / 04 / 2020 tarihinde şahsım/tez danışmanım tarafından Başkent Üniversitesi İntihal Tespit Programı adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 16'dır.

Uygulanan filtrelemeler:

1. Kaynakça hariç
2. Alıntılar hariç
3. Beş (5) kelimeden daha az örtüşme içeren metin kısımları hariç "Başkent Üniversitesi Enstitüleri Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Usul ve Esaslarını" inceledim ve bu uygulama esaslarında belirtilen azami benzerlik oranlarına tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Öğrenci İmzası:

Onay
24/ 04 / 2020

Dr. Mehmet GÜLŞEN

TEŞEKKÜR

Bu çalışmanın gerçekleşmesindeki katkılarından ötürü,

Sayın Dr. Mehmet GÜLŞEN'e (tez danışmanı), çalışmanın sonuca ulaşmasında sağladığı tüm katkılar ve sürecin tamamındaki destekleri için,

Başta annem Ayla GÖREN olmak üzere bu süreçte desteklerini eksik etmeyen tüm aileme ve her zaman yanımda olan F. Merve Bağcı'ya,

İçtenlikle teşekkürlerimi sunarım.

ÖZET

Sena Nur GÖREN

ZAMAN SERİSİ TAHMİN MODELLERİNDE VERİ ANALİZİ VE MODEL SEÇİMİ

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Endüstri Mühendisliği Anabilim Dalı

2020

Zaman serileri bir değişkenin ardışık gözlem değerlerini içeren veri kümeleridir. Takdir edilecektir ki bu gözlemler zaman içerisinde çevresel veya sistematik etkiler nedeniyle değişim gösterebilmektedir. Bu nedenle zaman serisi modelleri ile tahmin gerçekleştirilirken bütün gözlem kümesi ile modelin eğitilmesi yerine tarihsel olarak sondan geriye doğru gidililerek gözlem verileri bölümlere ayrılabilir. Çalışmada, bu bölümlerden tahmin modeli açısından en alakalı dönemin tespit edilip, eğitim kümesi olarak kullanılması ile gerçeğe daha yakın sonuçlar elde edileceği savunulmuştur. Yapılan araştırmada, eğitim veri kümesinin elde edilmesi için değişim noktası analizi yöntemlerinden CUSUM algoritması kullanılmıştır. Öncelikle bu algoritma popüler tahmin modelleri olan ARIMA ve Holt's Winter yöntemleri ile entegre edilerek elde edilen veri kümesi ile tahmin yapılmıştır ve gerçek verilerden oluşan test kümesi ile performansı ölçülmüştür. Daha sonra aynı tahmin modelleri tüm gözlem kümesi ile eğitilerek gelecek değerler tahmin edilmiş ve test veri kümesi ile performansı ölçülmüştür. Ayrıca zaman serisindeki değişim noktalarının tespit edilmesinin önemini göstermek amacıyla çalışmada ek bir yöntem olarak sabit süreli zaman pencereleri ile eğitim veri kümeleri oluşturulmuş ve bu kümelerle tahminler gerçekleştirilip performansları ölçülmüştür.

CUSUM algoritması kullanılarak gözlem kümesinin tahmin için en *"doğru"* bölümü ile eğitilen modellerin MSE hata değerleri diğer iki yöntemden elde edilen tahmin sonuçlarına göre daha küçük olduğu yani gerçeğe daha yakın sonuçlar elde edildiği görülmüştür.

ANAHTAR KELİMELER: Zaman serileri, Değişim noktası analizi, Tahmin, CUSUM

ABSTRACT

Sena Nur GÖREN

INPUT DATA ANALYSIS AND MODEL SELECTION IN TIME SERIES FORECASTING

Başkent University Institute of Science and Engineering

Department of Industrial Engineering

2020

Time series are data sets that contain consecutive observation values of a variable. These observations may change over time due to environmental or systematic effects. Therefore, while estimating with time series models, instead of training the model with the whole set of observations, the data can be divided into sections starting from the very end, and the most relevant periods from these sections will be used in the forecasting model. In this study, the CUSUM algorithm, which is a change point analysis method, is used to determine the length of the training dataset. This algorithm is integrated with ARIMA and Holt's Winter methods, which are popular prediction models, and forecasts are generated and evaluated on the test data.For validation, the same prediction models are trained on the whole data set, and its performance is used as a benchmark to evaluate the proposed approach. Furthermore, to show the importance of determining the change points in the time series, training data sets were created with fixed-time time windows, and estimates were made with these sets, and their performances were measured.

It is observed that the models trained with the *"correct"* part of the time series have smaller MSE values as compared to the prediction results obtained from the other two methods, that is, more realistic results were obtained.

KEYWORDS: Time series, Forecasting, Change point analysis, CUSUM

İÇİNDEKİLER

Sayfa

TABLOLAR LİSTESİ

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR LİSTESİ

ARIMA	Otoregresif Entegre Hareketli Ortalama
AR	Otoregresif
MA	Hareketli Ortalama
I	Entegre
HW(A)	Holt's Winter Katılımlı
HW(M)	Holt's Winter Çarpımsal
CPD	Değişim Noktası Bulma
CUSUM	Kümüatif Toplam
MSE	Ortalama Hata Kareleri
AIC	Akaike Bilgi Kriteri
MAPE	Ortalama Mutlak Yüzde Hata

1. GİRİŞ

Zaman serisi; bir değişkenin, genelde sabit aralıklarla farklı zamanlarda aldığı değerlerle ilgili bir dizi gözlemi ifade eder. Bu aralıklar, zaman ifade eden dakika, saat, gün, hafta vb. olabilmektedir. Gözlemlenen değişkenin değerleri zamana bağlı olarak değiştiği için, gözlemlerin bağımsız olduğu doğrusal bir regresyon modelinin varsayımı bu veriler için geçerli olmayacaktır. Zaman serileri, artan veya azalan bir eğilim ile birlikte, çoğu zaman mevsimsellik eğilimine, yani belirli bir zaman dilimine özgü varyasyonlara sahiptir. Gelecekteki olası durumların bugün verilecek olan kararları belirlemesinden dolayı, bir değişken ile ilgili zaman serisini anlama ve öngörüde bulunma; satış tahmini, borsa tahmini, hava durumu tahmini ve daha pek çok gerçek uygulama için temel oluşturur. Bu nedenle, bir zaman serisi veri kümesini anlayarak diğer bir deyiş ile en son trendi takip ederek daha iyi tahminler yapılabilir.

Zaman serisi analizi tek bir değişkenin zaman içindeki değişimlerini gözlemliyorsa tek değişkenli zaman serisi (univariate) olarak, birden fazla değişkenin zaman içinde birlikte değişimini inceliyorsa çok değişkenli zaman serisi (multivariate) olarak adlandırılmaktadır. Bu tez kapsamında yapılan sayısal çalışmalarda tek değişkenli zaman serileri kullanılmıştır. İstatiksel teorilerin çoğu bağımsız rassal örneklem gözlemleridir fakat bunun yanında zaman serileri ardışık gözlemlerin zaman dizisini hesaba katmaktadır. Bu gerçek sayesinde, gelecekte alacakları değerleri geçmiş dönemlerdeki gözlemlerde öngörmek mümkün olmaktadır.

Zaman serisi yapıları stokastik ve deterministik olmak üzere iki gruba ayrılmaktadır: Stokastik seriler, gelecekte serinin alabileceği değerlerin kısmen geçmiş gözlemleri tarafından tanımlanabilmesi demektir. Stokastik serilerin tam öngörülerini yapmak mümkün olmamakta ancak gelecekteki değerler, geçmiş gözlemlerin bir bilgisiyle koşullandırılan bir olasılık dağılımına sahiptirler. Zaman serileri genellikle random (tesadüfi) değişkenler yani stokastik (olasılıklı) değişkenlerle çalışmaktadır. Eğer bir zaman serisi tam olarak öngörülebiliyorsa, deterministik (kesin) zaman serisi olarak ifade edilmektedir. Zaman serisinin deterministik özellikleri; sabit, trend ve mevsimselliğin varlığını ortaya koyarken, stokastik özellikleri ise, değişkenin durağanlığı ile ilgilenmektedir [10]. Burada durağanlık ile bahsedilen verideki değişimlerin istikrarlı olmasıdır, mevsimler değişimler, dönüşümler gibi [17].

Zaman serileri analizlerinin en önemli amaçlarından birisi öngörü yapmaktır. Bu öngörü için gelecekteki değerleri tahmin edilecek olan değerlerin gözlenmesi ve daha sonra bu verilerden sonuç çıkaracak tahmin modelinin seçilmesi gerekir. Zaman serileri üzerinde tahmin

yapmak için literatürde kullanılan birçok yöntem olmakla beraber bu çalışmada ARIMA modeli ve Holt's Winter modelinin iki farklı varyasyonu temel alınmıştır. En yaygın olarak kullanılan ve tanınan istatiksel zaman serisi tahmin modellerinden biri, Otoregresif Entegre Hareketli Ortalama (ARIMA) modelidir. Kısaltması ARIMA olan model, Otomatik Regresif Entegre Hareketli Ortalama (Auto-Regressive Integrated Moving Average) anlamına gelmektedir. ARIMA modeli, basitlik ile çeşitli zaman serilerini ve optimal model yapımı bakımından dikkate değer ölçüde tahmin doğruluğu ve verimliliği ile tanınmaktadır [7]. Bu modelin uygulanmasında temel varsayım, zaman serilerinin doğrusal olduğunu ve normal dağılım gibi istatiksel bir dağılım izlediğini varsaymaktır. ARIMA metodu temel olarak en az 40 tarihsel veri noktası ile gelecekteki değerleri yansıtmak için kullanılır. Gözlem sayısı az ise bu model ile çok iyi tahminler yapılamaz [7]. ARIMA algoritması, hem otoregresif (AR) bileşenlere hem de hareketli ortalama (MA) bileşenlere izin verir. AR bileşenleri, tahmin denklemindeki durağanlaştırılmış serilerin gecikmelerini yansıtırken, I entegre bileşeni, zaman serisinin durağan olmasına izin vermek için ham gözlemlerin farklılığını yansıtır ve veri değerleri ile önceki değerler arasındaki farkla değiştirilir, MA bileşeni ise bir gözlem ile gecikmeli gözlemlere uygulanan hareketli ortalama modelden kalan hata arasındaki bağımlılığı yansıtır. ARIMA metodolojisini uygulamanın ilk adımı durağanlığı kontrol etmektir. "Durağanlık", daha önce de bahsedildiği gibi serinin zaman içinde oldukça sabit bir seviyede kaldığını belirtir. Çoğu ekonomik veya piyasa verisinde olduğu gibi bir eğilim varsa veri sabit değildir. Veriler ayrıca zaman içindeki dalgalanmalarında sabit bir varyans göstermelidir. Mevsimsel ve daha hızlı büyüyen bir seri ile bu kolayca görülebilir. Böyle bir durumda, mevsimsellikteki iniş ve çıkışlar zaman içinde daha dramatik hale gelecektir. Bu durağanlık koşulları karşılanmadan, işlemle ilişkili hesaplamaların çoğu hesaplanamaz. ARIMA modelinin, diğer modellerden farkı çalışmadan önce sabit olmayan verileri sabit verilere dönüştürmesidir. Böylece, ARIMA tek değişkenli zaman serisi modeli tanımlama, parametre tahmini ve öngörmede büyük esneklik sunar [18].

Artan veya azalan eğilimli ve mevsimsellik içermeyen bir katkı modeli kullanılarak tanımlanabilecek bir zaman dizisi varsa, kısa vadeli tahminler yapmak için Holt's Winter metodu kullanılabilir. Holt's Winter tahmin algoritması, bir zaman serisinin düzeltilmesine ve bu verilerin ilgili alanlarını tahmin etmek için kullanmasına olanak tanır. Holt's Winter modelinin temelini oluşturan üstel düzeltme (exponential smoothing) modeli, eski veriler için ağırlığın değerini azaltmak için geçmiş verilere karşı katlanarak ağırlıkları ve değerleri atar. Başka bir deyişle, daha yeni geçmiş verilere, tahminlerde eski sonuçlardan daha fazla ağırlık verilir. Holt's Winter, verideki trend ve mevsimsellik tanımlamasına dayanır. Bu yöntemin

mevsimsel bileşenin doğasında farklılık gösteren iki varyasyonu vardır, bunlar katkı yönetimi ve çarpımsal yöntemdir. Katkı yöntemi(additive) mevsimsel değişimler seri boyunca kabaca sabit olduğunda tercih edilirken, çarpımsal yöntem (multiplicative) mevsimsel değişimler serilerin seviyesiyle orantılı olarak değiştiğinde tercih edilir. Katkı yöntemi ile mevsimsel bileşen, gözlemlenen serinin ölçeğinde mutlak terimlerle ifade edilir ve seviye denkleminde seri, mevsimsel bileşen çıkarılarak mevsimsel olarak ayarlanır. Her yıl içinde mevsimsel bileşen yaklaşık olarak sıfır ekleyecektir. Çarpımsal yöntemiyle mevsimsel bileşen göreli terimlerle ifade edilir ve seri, mevsimsel bileşene bölünerek mevsimsel olarak ayarlanır [5].

Model seçimi zaman serisi araştırmalarında kritik bir aşamadır çünkü çoğu zaman aynı çıktıya sahip birden çok rakip modelle karşı karşıya kalınabilir. Modelleme gerçeğin yakınlaştırılmasıdır, bu nedenle model seçimi gerçeklikten uzak bir modeli reddetmek ve gerçeğe yakın olanı seçmek içindir [6]. Ölçülen hata metrikleri ile farklı modellerin performansı değerlendirilir ve belirli bir set için modeller arasından en iyisi seçilir. Veri kümesi için en uygun model seçimi yapılırken, üç hata parametresi hesaplanmıştır: AIC, MAPE, MSE. Metodoloji bölümünde detaylıca anlatılan Python diline ait Statsmodels kütüphanesi kullanılan tahmin modellerinin eğitim kümesine uygulandığında tahmin sonucu varlıkları içerisinde yapılan tahminin AIC değerini de vermektedir. AIC kriteri şeklindeki multimodel çıkarım, hangi modelin veri kümesine en uygun olduğunu belirmek için güçlü bir yöntemdir [8]. Gözlem sayısı küçük olduğunda AIC'nin çok fazla parametresi olan modelleri seçme olasılığı vardır. Bu nedenle AIC çalışmada karar verme metriği olarak kullanılmamıştır. Diğer bir popüler hata metriği olan MAPE ise, olumlu hatalara göre olumsuz hatalara fazla ağırlık verdiğinden tahminleri düşük olan bir yöntemi seçmesi muhtemel olduğu için sonuç kararı metriği olarak kullanılmamıştır [9]. MSE, büyük hatalarla daha fazla tahmin yapılmaması için çok uygundur çünkü hesaplamadaki kareleme ile hatalara daha fazla ağırlık verir. Bu nedenle nihai kararın verilmesinde MSE metriği kullanılmıştır.

Holt's Winter öngörü algoritması tahmin gerçekleştirirken verilerin güncelliğine göre ağırlıklandırma yapsa da veri kümesinden eskileşen veriler çıkarılmadığı için ortaya çıkan tahmini etkilemesi kaçınılmazdır. Bu noktada devreye veri kümesindeki değişim noktalarının analiz edilmesi ve kümeden çıkarılması girer. Veri kümesinin anlaşılması trend ve sezon bilgisinin belirlenmesinin yanında değişim noktalarının tespit edilmesini kapsar. Dolayısıyla yapılan çalışmada üzerinde durulan nokta, sadece veri kümesi üzerinde tahmin yapılırken kullanılan modelin parametrelerinin optimizasyonu değil aynı zamanda veri kümesi içerisindeki zaman içerisinde değer kaybedip eskileşen verileri ayıklayıp öngörü gerçekleştirmektir. Bunun için veri kümesinde sondan geriye doğru gidilerek değişim noktaları

analiz edilip ve tahmin modeli ile en alakalı bölüm tespit edilmelidir. Bu daha değerli olan veri kümesi kullanılarak nihai öngörü, belirlenmiş olan tahmin modelleri içerisinden gerçeğe en yakın sonuç veren ile yapılması amaçlanmıştır.

Zaman serisi verileri, zaman içinde sistemlerin davranışını tanımlayan ölçüm dizileridir. Bu davranışlar, dış olaylar ve/veya dahili sistematik değişiklikler nedeniyle zaman içinde değişebilir [1]. Tekrarlanmayan zaman serisi verilerinin yapılacak öngörülerde etkisi olmadığı gibi hatalı sonuçlar verebilmektedir. Bu nedenle değişim noktaları analizi verinin en güncel kümesi ile *"Doğru"* eğitim seti elde etmek amacıyla kullanılmıştır. Değişim noktası analizi, zaman içinde verilerin toplandığı herhangi bir özelliğe uygulanabilir (ortalama, standart sapma, aralık, hata seviyesi, vb.) ve zaman serisinin bir özelliği değiştiğinde verilerde ani değişiklikler bulma problemidir. Değişim noktası tahminlerinin odağı, bilinen değişimin niteliğini ve derecesini tanımlamaktır. CUSUM, literatürde kullanılan popüler bir anomali tespit metodu olarak bilinmektedir. Bu metot, uygulanan veri kümesindeki ani bir değişiklik, ani bir kayma veya ortalamadaki değişimi gösterir. Trendler ve değişiklikler için geçmiş verileri analiz etmek için kullanıldığında, bu değişiklik noktası analizi bir kontrol tablosundan çok daha faydalı bilgiler sağlamaktadır. Değişim noktası tespiti, küçük ve sürekli olan değişiklikleri tespit etmekte ve bunları karakterize etmekte güçlüdür.[12].

Değişim noktası analizi yinelemeli olarak verideki değişim noktalarını saptayan bir yöntemdir. Yani amaç sadece güncel veriyi alıp kullanmak değil en anlamlı olan güncel veriyi kullanmaktır. Bu yöntemi desteklemek amacıyla çalışma kapsamında ek olarak belirlenmiş sabit süreli zaman pencereleri kullanılarak eğitim kümeleri oluşturulmuş ve manuel olarak doğru veri kümesine nasıl karar verileceği ve sonucun CUSUM algoritmasından ne kadar farklı olacağı gösterilmiştir. Belirli katsayılarla gözlem kümesinin periyodu çarpılarak setlerin boyutları belirlenmiş ve eğitim kümesi olarak kullanılmıştır. Bu yöntemle CUSUM kullanılmasaydı bir zaman serisi modelinde eğitim veri kümesinin belirlenmesi için alternatif bir çalışma da sunulmuştur.

Bu bilgilere dayanılarak, yapılan tez çalışmasında örnek olarak alınan 15 gerçek zaman serisi üzerinde öncelikle değişim noktası tespiti gerçekleştirilmiş ve verinin gelecek dönemleri tahmin etmekte kullanılacak olan anlamlı bölümü tahmin yöntemlerine girdi olarak gönderilmiştir. Daha sonra bu zaman serisinin periyodunun üç, dört ve beş katı kadar güncel gözlem değerlerinden ayrılarak veri kümeleri oluşturulmuş. Bu üç veri kümesi için de tahmin modelleri çalıştırılmış ve her birinin performansı test için ayrılmış olan gerçek değerler ile karşılaştırılarak hata miktarları ölçülmüştür. Son olarak da bu verilerin değişim noktası analizi yapılmadan tamamı modelin girdisi olarak kullanıldığı zaman gelecek dönemler için nasıl

öngörülerde bulunduğu gözlemlenerek çok daha başarılı bir tahmin için istatiksel hata değerleri ölçülerek karşılaştırmalar yapılmıştır. Karşılaştırma sonucunda en iyi sonucu veren model ile değişkenin gelecek dönemde alacağı değerler tahmin edilmiştir. Çalışmanın algoritmasının bir uygulama haline getirilmesi PyCharm ortamında, Python dili ile gerçekleştirilmiştir. Tahmin modellerinin algoritmaları için Statsmodels kütüphanelerinden faydalanılmış, CUSUM ve Sabit Süreli Zaman Pencereleri yöntemlerinin algoritmaları ise çalışma kapsamında geliştirilmiştir.

2. ZAMAN SERİLERİNDE VERİ ANALİZİ VE MODEL SEÇİMİ

Herhangi bir zaman serisinde gözlenen değişiklikleri bulma probleminin, yapay zeka ve veri madenciliği alanında yapılan çalışmalarda fazlaca dikkat çektiği görülmektedir. Daha önce yapılmış çalışmalarda Yamada, Kimura, Naya ve Sawada tarafından özetlendiği gibi [19] örneklendirmek gerekirse, hücresel sistemlerde sahtekarlık tespiti, bilgisayar ağlarında izinsiz giriş tespiti, görme sistemlerinde düzensiz hareket algılama, müzik segmentasyonu ve twitter verilerinden duyarlılık analizi gibi çeşitli gerçek dünya uygulamaları vardır.

Değişim noktaları zaman serisi gözlemlenen ani değişimlerdir. Bu tür ani değişiklikler, durumlar arasında meydana gelen geçişleri temsil edebilir. Zaman serisinin güncel eğilimini tespit edip bunun üzerinden tahmin yapmakta faydalıdır.

Değişim noktası tespiti (CPD), zaman serisinin bir özelliği yapısal olarak değiştiğinde, verilerden istatistiksel olarak bu değişikliği tespit edebilme problemidir [2]. Değişken üzerinde gerçekleşen değişimler ile serideki güncel eğilim tespit edilebilir. Değişim noktası analizi ile, birden fazla değişikliğin tespiti yapılabilir, değişikliklerin zamanlaması ve uyum aralıklarına bakılarak değişiklikler daha iyi karakterize edilir [3].

Veri trendlerindeki bir değişikliği tespit etme fikri popüler bir konudur ve pratikteki ilk uygulamaları istatiksel süreç kontrolü alanlarında bulunabilir. Burr'da bildirildiği gibi [25], istatiksel süreç kontrolündeki kontrol grafikleri 1924 yılında Walter A. Stewart tarafından sunulmuştur. Süreç kontrol grafiğinin arkasındaki fikir, doğal rastgele varyasyonların ötesinde bir veri akışındaki varyasyonu yakalamaktır. Genel veri analizi bağlamında, veri örüntülerinde değişiklik bulma konusunu ele alan öncü çalışmalardan biri Taylor tarafından yapılmıştır [4]. Bir ilaç şirketinde kalite direktörü olan Taylor, toplam grafik (CUSUM) ve önyükleme temelli bir yaklaşım önermektedir. CUSUM, veri ortalamasından kümülatif sapmayı veren basit bir grafiktir. Sıfırdan başlar (aslında sıfır ilk sapma) ve tek veya birden fazla çıkış ve iniş yapabilir, ancak sıfırla biter. Maksimum mutlak sapmanın meydana geldiği nokta, örüntü değişikliği için bir aday noktası olarak alınır. Bu noktanın gerçekten bir yapısal değişikliği gösterip göstermediğini test etmek için, değişikliğin büyüklüğü önyükleme analizi ("bootstraping") kullanarak tahmin edilen bir tahminciyle karşılaştırılır. Bu analizde, çok sayıda önyükleme (yani orijinal verilerin rastgele yeniden sıralanması) üretilir ve her önyükleme için maksimum sapmanın büyüklüğü kaydedilir. Bir güven seviyesi, orijinal verilerinkinden daha büyük sapmaya sahip önyükleme yüzdesine göre hesaplanır. Oran yüzde 5 ile 10'un altında kalırsa, bir değişikliğin meydana geldiğine dair güçlü bir kanıt olarak kabul edilir.

Finans alanında, tarihsel verileri alt bölümlere ayırmayı düşünen birkaç çalışma vardır. Maheu ve McCurdy [33] verilerdeki kırılma noktalarını tanımlamaktadır. Her bölümün uzunluğu, gelecekteki getirilerin ortalaması ve varyansı olarak ifade edilen öngörücü içeriğe dayanmaktadır. İki kırılma noktası arasındaki her bölüm farklı bir alt model ile temsil edilir. Nihai tahmin, her bölümdeki tahminlerin olasılık ağırlıklı ortalamaya göre birleştirilmesiyle üretilir. Önerilen yaklaşımda, tarihin bazı bölümlerini atmak yerine, tüm tarih, tahmini bir olasılıkla da olsa, nihai öngörüye bir girdi içerir.

Daha doğru ve daha güçlü tahminler elde etmek için girdi verilerinin kullanılması açısından, sık kullanılan bir yöntem veri kümelemesidir ("aggregation"). Bu yöntem, ürün hiyerarşilerinin, lokasyonun ve zamanın kümeleme parametresi olarak kullanıldığı yaklaşımlar içerir.

Ürün hiyerarşisi, tek tek öğelerin hiyerarşik bir şekilde ilgili ürün kategorilerine gruplandırılmasını ifade eder. En düşük seviyede ürünler taneli düzeyde tanımlanır ve bazen her bir SKU (stock keeping unit) ürünü temsil eder. Aşağıdan yukarıya doğru hareket ederken, ürünler belirli özelliklere ve işlevlere göre gruplara ayrılır. Örneğin, giyim bağlamında, tişörtler daha sonra v yaka, polo ve diğerleri gibi alt kategorilere ayrılan bir üst kategoriyi temsil edebilir. Bu kategoriler ayrıca v yaka/ beyaz / boyut xl gibi daha alt kategorilere ayrılır. Alt kategoriler bazen SKU seviyesi kadar derin daha düşük seviyelere inebilir. Daha düşük seviyelerde, veriler daha ayrıştırılır ve düşük veya sıfır değerli gözlemlerle aralıklı bir seriye dönüşebilir. Bu tür seriler için tahmin modelleri oluşturmak kolay değildir ve birçok ciddi teknik zorluk içerir [31; 28; 22]. Bu seviyede, geleneksel doğruluk metrikleri performans tahmini için iyi bir gösterge olmayabilir [34; 29; 36].

Hiyerarşide ürünleri bir araya getirmenin birçok yolu olmasına rağmen, bunlar iki yaklaşıma ayrılabilir. Aşağıdan yukarıya yöntemi, bir diziyi düşük düzeyde tahmin eder ve ardından en üst düzey tahmini elde etmek için bunları birleştirir. Tersine, yukarıdan aşağıya yaklaşım, daha sonra alt düzey serilere dağıtılan üst seviyede tahminler üreterek işleme başlar [38; 27; 20]. Performans açısından, bir yaklaşımın diğerinden üstün olduğu neticesine varmayı zorlaştıran karışık bir sonuç vardır [26; 37; 35].

Konum veya coğrafya, verileri toplamak için kullanılabilir. Fiziksel lokasyona ek olarak, dağıtım kanalları lokasyon parametreleri olarak düşünülebilir. Literatür, kümeleme için konum veya dağıtım parametrelerinin çeşitli kullanımını içerir [30; 39; 23].

Zaman bileşeninde kümeleme, verileri daha düşük zaman serilerine gruplamak anlamına gelir. Bu süreç genellikle zamansal regresyon olarak adlandırılır ve verileri daha yüksek sıklıktan daha düşük sıklık zaman serilerine (örneğin, aylık düzeyde toplanan günlük

veriler) toplamayı amaçlar. Bu tür toplamanın varyasyon katsayısını azaltması beklendiğinden daha iyi tahmin performanslarına yol açabilir.

Toplama yöntemi örtüşmeyen veya örtüşen olabilir. Çakışmayan yaklaşımda kümeleme süreleri, çakışma olmayan ayrık zaman parantezleridir. Örneğin, günlük veriler Ocak, Şubat, Mart vb. takvim aylarıyla birleştirilirse, bu örtüşmeyen bir toplamadır. Çakışan kümeleme için, sabit uzunluktaki hareketli bir pencere verileri toplamaktadır. 30 günlük örtüşen bir toplamda, son 30 güne ait bir toplamı içerir ve en yenisini toplama katarken en eskisini toplamdan düşürerek günlük olarak güncellenir [24; 21; 34; 32].

Zaman serileri tahmininde, geçmiş gözlemler toplanır ve serinin temelini oluşturan veri oluşturma sürecini tanımlayan uygun matematiksel modelin geliştirilmesini sağlamak için analiz edilir [13].

Zaman serileri tahmini, gözlemler arasında düzenli bir ilişkinin olduğu veriler üzerinde bir öngörü modeli geliştirmeyi ve kullanmayı içerir. Öngörü, geçmiş verilere uygun modellerin alınmasını ve gelecekteki gözlemleri tahmin etmektir [5].

Bir zaman serisi tahmin modelinin becerisi geleceği tahmin etme performansı ile belirlenir. Bu genellikle belirli bir tahminin neden yapıldığını, güven aralıklarını ve problemin altında yatan nedenleri daha iyi anlayabilme amacı ile gerçekleştirilir. Belirli bir zaman serisine uygun bir model ve parametreleri bilinen veri değerleri kullanılarak tahmin edilmesi prosedürü Zaman Serisi Analizi olarak adlandırılır [5].

İleriki dönemler için yapılacak öngörüler geliştirilen model üzerinden gerçekleştirilir. Değerli stratejik kararların alınmasında veya ihtiyati tedbirler zaman serileri tahminlerine göre yapıldığı durumlarda bu öngörülerin performansı daha çok önem kazanır. Örneğin, bir veri akışı ortalama, varyans gibi karakteristik girdi ölçeğinde ani bir değişime uğrayabilir; belirli bir noktadan sonraki davranışı daha önce olanlardan tamamen bağımsız hale gelebilir. Güçlü bir tahmin algoritması bu tür koşullar altında bile doğru tahminler yapabilmelidir [13].

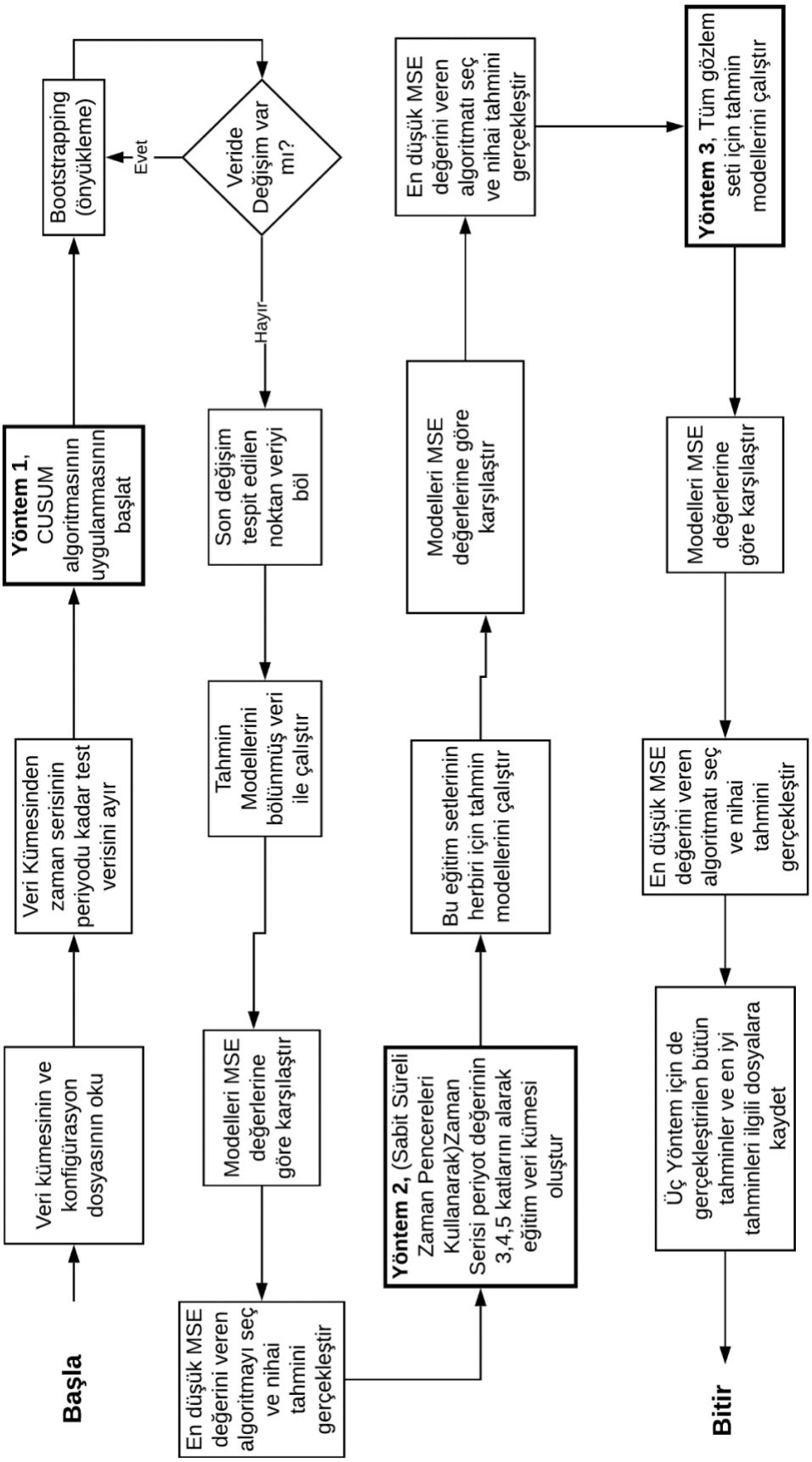
3. METODOLOJİ

Bu bölümde çalışma kapsamında önerilen algoritmanın zaman serisi problemlerinde Şekil 3.1'de akış diyagramı olarak ifade edilen çözüm yaklaşımının adımları anlatılmıştır. Kullanılan eğitim ve test veri kümeleri, tahmin modelleri parametreleri ve model sonuçlarının karşılaştırılması için kullanılan hata metrikleri detaylandırılmıştır.

3.1. Çözüm Yaklaşımı

Bir zaman serisi değişkeninin gelecek dönemlerdeki değerlerini tahmin etmek için kullanılan pek çok tahmin modeli vardır. Bu modeller varsayılan durumda, bir veri kümesinin tamamını eğitim kümesi olarak kullanır ve modelde kullanılacak olan parametrelerin optimum değerlerinin hesaplanması ile tahmininin gerçekleştirilmesini sağlar. Bu çalışmada amaç gerçeğe daha yakın tahminlerin yapılması için sadece modellerin parametre optimizasyonlarını yapmak değil aynı zamanda kullanılan eğitim kümesinin de en anlamlı bölümünü bulmaktır. Zaman serisi değişkenleri için yapılan tahminlerin daha başarılı sonuçlar vermesi için önerilen algoritma adımları Şekil 3.1'de gösterilen akış şemasındaki gibidir. Algoritmanın akışında belirtilen üç yöntemde, veri analizi ve veri seçimi adımları farklılık göstermekte olup tahmin modellerinin uygulanması ve performansların ölçümü benzer şekillerde gerçekleştirilmiştir. Çalışmada veri seçimi kapsamında "Doğru Eğitim Kümesi"nin bulunması amacıyla verideki değişim noktalarını tespit eden CUSUM algoritmasının kullanılması önerilmiştir. Bu nedenle Şekil 3.1'de akıştaki yöntemlerden birincisi bu algoritma kullanılarak yapılan tahmin adımlarını içermektedir. İkinci bir yöntem olarak ise Sabit Süreli Zaman Pencereleri kullanılarak eğitim veri kümelerinin oluşturulması ve bu kümelerin tahmin modellerinde kullanılması gerçekleştirilmiştir. Burada veri setinin periyot değerinin 3,4 ve 5 katı kadar veri, gözlem kümesinin en güncel değerinden itibaren alınmış ve her biri için yöntem ikiden itibaren bahsedilen adımlar tek tek gerçekleştirilmiş ve değerlendirilmiştir. CUSUM yönteminin gerekliliği değerlendirilirken alternatif bir metoda ihtiyaç duyulması ve varsayılan yöntemle kıyaslamaların daha doğru değerlendirilmesi için bu alternatif yöntem çalışmaya dahil edilmiştir. CUSUM algoritması ile yapılmak istenen veri kümesi bölümlere ayrılırken zaman serisinin anlaşılması ve değişim noktalarının tespit edilip eskileşen verilerden arınmış doğru kümenin kullanılmasıdır yani sadece güncel olan kısmın kullanılması değildir. Sabit Süreli

Zaman Pencereleri yöntemiyle sadece verinin zaman boyutuna bakarak sabit boyutta kümeler oluşturulduğu için değişim noktası analizinin önemi bu yöntemle tekrar desteklenmeye çalışılmıştır. En son olarak da varsayılan yöntem olan bütün zaman serisi gözlem değerlerinin, test veri kümesi hariç, tahmin modelini eğitmek için kullanılması ve elde edilen sonuçların değerlendirilmesi gerçekleştirilmiştir. Yöntemlerin hepsinde test veri kümesi serinin en sonundan tahmin edilecek dönem kadar ayrılmış olan gerçekleşmiş gözlem değerleridir. Çözüm yaklaşımının beklenen sonucu CUSUM algoritması uygulanarak elde edilen eğitim setinin modellerde kullanılmasının gerçeğe daha yakın tahminler yapılmasını sağladığı kanısına varmaktır. Bunun için Şekil 5.1'de gösterilen akıştaki gibi üç yönteme ait adımların hepsi gerçekleştirilmiş ve bu yöntemlerin hepsi için ölçülen hata metrikleri kıyaslanmıştır. Yöntemlerden elde edilen sonuçların hepsi ilgili dosyalarda daha sonra kullanılabilecek şekilde kayıt edilmiştir. Önerilen model algoritması çalışma kapsamında Python dili ile yazılıp projelendirilmiştir. Kullanılan tahmin modellerinin alt yapısı ve modellerin her birindeki parametre optimizasyonları için Python'da bulunan *Statsmodels* kütüphanesindeki metotlardan faydalanılmış, CUSUM ve Sabit Süreli Zaman Pencereleri algoritmaları çalışma kapsamında geliştirilmiştir.



Şekil 3.1. Metodoloji Akış Diyagramı

3.2. Veri Analizi ve Eğitim Verisinin Seçimi

3.2.1. Birinci yöntem

Çalışmanın yapılmasındaki temel amacı kapsayan bu yöntemde, tahmin modellerinin eğitilmesi için kullanılacak olan veri kümesinin seçimi CUSUM algoritması ile gerçekleştirilmiş ve tahmin modellerinde bu küme kullanılmıştır. Yapılan tahminler sonucunda bu kümenin diğer yöntemlerden daha iyi sonuç verip vermediğinin karşılaştırılması için hata metrikleri hesaplanmıştır.

3.2.1.1. CUSUM algoritması

Bu yöntemin temel adımı, artıkların toplamını hesaplamak ve istatistiksel özellikleri tespit etmek için referans durumdan sapmaları belirlemek adına tekrarlanan işlemlerdir. Bu yöntem çok basit ve gerçekçi bir yöntemdir, çünkü zaman serisinin herhangi bir işlevsel formunu varsaymadan değişimini algılar [3]. CUSUM prosedürü, biriken kalıntıları ölçülen bir değişken arasında çizmek ve kalan zaman serilerinin durağan veya homojen olmadığını kanıtlamak için kullanılır. Kümülatif toplamlar değerlerin kümülatif toplamları değil, değerler ve ortalama arasındaki farkların kümülatif toplamıdır. Bu farklar sıfıra toplanır böylece kümülatif toplam daima sıfır ile biter.

X_1, X_2, …, X_n değerleri n tane veri seti elemanını temsil ediyor, bu değerlerden hesaplanan kümülatif toplamlar ise S_1, S_2, …, S_n şeklinde gösteriliyor. Bu durumda kümülatif toplam algoritması adımları şu şekilde gösterilebilir:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.1}$$

$$S_0 = 0 \tag{3.2}$$

$$S_i = S_{i-1} + (x_i - \overline{x}) \tag{3.3}$$

$$\text{Değişim Noktası} \rightarrow |S_m| = \max_{i=0,1,\ldots,n} |S_i| \tag{3.4}$$

Eşitlik 3.1'de öncelikle serinin ortalaması hesaplanır daha sonra kümülatif toplamların ilk değeri sıfır olarak atanır (3.2) ve serideki her değerin ortalama ile farkı alınarak kümülatif toplamlar hesaplanır (3.3). $|S_m|$, CUSUM'da sıfırdan en uzak olan noktadır. m noktası değişimden bir önceki noktayı, m+1 noktası değişimden bir sonraki noktayı ifade eder. Bu yöntemin doğru yorumlanması için değişiklik zamanlamasını ölçmek için bir önyükleme yaklaşımı ile güven seviyesi (3.5) kullanılmıştır. Var olan eğitim verisinden rastgele olacak şekilde 10000 tane önyükleme listesi oluşturulmuş ve bu listelerde CUSUM algoritması uygulanarak S_i değeri hesaplanmıştır. Bulunan S_i değerleri ile (S_{diff} = max S_i – min S_i) hesaplanarak değerlerin dağılımının nerede olduğunu belirlemek için önyükleme sonuçları (3.5) değerlendirilmiştir. Bu değerlendirme aşağıda formülü verilen *Güven Seviyesi* (Confidential Level) üzerinden gerçekleştirilmiştir.

$$CL = \frac{R}{N} * 100 \qquad (3.5)$$

R eğitim setinden oluşturulan rastgele listelere CUSUM algoritması uygulandığında $S_{diff} > S_{diff}^{bootstrap_i}$ karşılaştırması doğru olan veri seti sayısını gösterirken, N toplam önyükleme listesi sayısını göstermektedir [4]. Bu çalışmada önyükleme sayısı 10000 olarak ve veri kümesinde değişiklik olduğunu belirten güven seviyesinin %99 olması gerektiği belirlenmiştir. CUSUM algoritması adımları veride değişiklik tespit edildikçe özyinelemeli olarak tekrar edilmiştir. Tespit edilen her değişim noktasından veri bölünerek kalan kısım için CUSUM algoritması tekrar çalıştırılmıştır. Güven seviyesini sağlayan bir değişim tespit edilmediği anda en son değişim gözlenen noktadan itibaren yani S_m değerini veren m (3.4) noktasından zaman serisi bölünerek nihai eğitim kümesi elde edilmiştir.

3.2.2. İkinci yöntem

3.2.2.1. Sabit süreli zaman pencereleri

Şekil 3.1'deki akışta bahsedilen ikinci yöntemde veri seçimi, sabit katsayılarla serinin periyodunun çarpımı kadar boyutlarda gözlem değerinin serinin sonundan ayrılması ile yapılması şeklindedir. Çalışma kapsamında sabit katsayılar 3,4 ve 5 olarak belirlenmiştir. Dolayısıyla bu yöntem kapsamında 3 farklı eğitim veri kümesi oluşturulmuş ve her biri için tahmin modellerinin hepsi ile tahminler gerçekleştirilmiştir. Buradaki ana amaç değişim noktalarına bakılmaksızın, verideki değişimlerin analizi yapılmaksızın en güncel veriden en eski veriye doğru gidilerek verinin bölümlere ayrılarak seçilmesi ile alınan sonuçların CUSUM algoritmasının kullanımını desteklemektir. Bunun için farklı katsayılarla üç farklı küme oluşturulup daha fazla örneklendirme yapılmaya çalışılmıştır. Bu çalışmanın tabi ki bir yandan da CUSUM gibi özyinelemeli çalışan bir veri analizi otomasyon algoritması kullanılmasaydı alternatif yöntem ne olabilirdi sorusunu da cevaplayabilecek faydası olmuştur. Zaman serisindeki gereksiz verilerden arınarak daha küçük bir veri kümesi ile çalışıldığında nasıl sonuçlar elde edildiği gözlemlenebilmiştir. Eğer veri kümesi değişkeni belirli bir noktadan itibaren aynı değişimleri gösteriyor yani daha sabit bir küme ise bu yöntemin iyi sonuçlar vermesi muhtemeldir.

3.2.3. Üçüncü yöntem

Veri seçiminin üçüncü yöntemi aslında değişim noktası analizi yapılarak veya farklı bir yöntemle veride seçime gidilmeden varsayılan serinin tamamının kullanılmasıdır. Çalışmada amaçlanan bu yönteme alternatif yöntemler sunmaktır. Bir zaman serisi bir değişkenin aldığı değerleri ifade eder ve çoğu zaman dış parametrelere bağlı değişimler gösterir. Çok uzun periyotlarda 10 yıl, 20 yıl gibi eskileşen ve anlamını yitiren verilerin oluşması kaçınılmaz duruma gelir dolayısıyla gelecekle ilgili yapılan tahminlerin daha gerçekçi olması için anlamlı kısmın bulunması önemli hale gelir. Bunun yanında kısa zaman aralığında toplanan gözlem verilerinden tahmin yapılırken bahsedilen bu durum tam tersi olacaktır verinin bölünmeye çalışması ile elde edilecek olan tahmin muhtemelen daha kötü sonuç verecektir.

3.3. Tahmin Modeli Seçimi

Algoritmada her üç yöntem ile elde edilen veri kümeleri kullanılarak sırası ile ARIMA ve Holt's Winter modelleri kullanılarak tahminler gerçekleştirilmiştir. Tahmin modeli seçimi yapılmadan önce zaman serisi içerisinden tahmin sonuçlarının doğruluğunun ölçülebilmesi için test veri kümesi ayrılmıştır. Bu test veri kümesi daha önce de bahsedildiği gibi tahmin edilecek dönem sayısı kadar yani bu da çalışmada serinin periyodu kadardır ve gerçek gözlem değerlerinden oluşmaktadır. Model seçiminde karar verici metrik MSE'nin hesaplanmasında her bir tahmin sonucu ve bu test kümesi kullanılmıştır.

3.3.1. Holt's Winter yöntemi

Bu yöntem için bu çalışma kapsamında Python diline ait *Statsmodels.tsa.holtswinter* kütüphanesinin *ExponentialSmoothing* tahmin metodundan faydalanılmıştır. Bu metoda ait iki farklı yöntem aşağıdaki gibi kullanılmıştır:

Holt's Winter (katkılı) modeli için *Statsmodels* kütüphanesinin *ExponentialSmoothing(train, trend="add", seasonal="add", seasonal_periods = s_period)* şeklinde ifade edilen metodu kullanılmıştır. Burada *train* modelde kullanılacak eğitim setini belirtirken, *trend* modelin trend(eğilim) bileşeninin türünü yani katkılı mı çarpımlı olacağını, *seasonal* modelin sezon bileşeninin türünü, *seasonal_periods* parametresi ise verinin tam bir mevsimsel döngüdeki dönem sayısını belirtir. Bu model verinin trend, seviye ve sezon bilgilerini parametre olarak kullanarak gelecek dönemler için tahmin gerçekleştirir. Modelin verimli bir sonuç elde etmesi için bu parametrelerin kullanılan eğitim veri kümesi için optimum değeri hesaplanmalı ve bu değerler kullanılarak tahmin yapılmalıdır yani parametrelerin optimizasyonu zaman serisi tahmin sürecinin bir parçasıdır. Kullanılan *Statsmodels* kütüphanesi bu optimizasyonun gerçekleştirilmesini modelin fit edilmesi olarak tanımlamış ve daha önce bahsedilmiş olan *ExponentialSmoothing()* metodu sonucunda oluşacak modelin üzerine *fit()* metodunun çağırılması ile gerçekleştirilmektedir. Algoritmayı doğrulamak için yapılan sayısal örneklerde veri kümelerinin sezon periyotları bilinmektedir bunlar parametre olarak kullanılan metoda gönderilmiştir, seviye ve eğilim parametreleri kütüphanenin metoduna bırakılmıştır. En son optimum parametrelerle belirlenen eğitim veri kümesi üzerinden tahmin gerçekleştirmek için kütüphanenin *forecast(s_period)* metodu kullanılarak belirtilen *s_period* parametresi kadar öngörü verisi oluşturması beklenmiştir.

Holt's Winter (çarpımlı) modelinde ise katkılı varyasyonun modeli oluşturmak için kullanılan *ExponentialSmoothing(train, trend="mul", seasonal="mul", seasonal_periods = s_period)* metodunun *trend* ve *seasonal* parametrelerinin değerleri değiştirilmiş ve katkılı modeldeki gibi tanımlanan modelin parametre optimizasyonunun yapılması için *fit()* metodu kullanılmıştır. Bu parametrelerle gelecek dönemlerin tahmin edilmesi için de katkılı modelde de yapıldığı gibi *forecast(s_period)* metodundan faydalanılmıştır.

3.3.2. ARIMA yöntemi

Bu yöntem için çalışma kapsamında Python diline ait *Statsmodels.tsa.arima.model* kütüphanesinin *ARIMA* metodu kullanılmıştır. ARIMA yönteminin modelini oluşturmak için kütüphanenin *ARIMA (train, order(p,d,q))* metodundan faydalanılmıştır. Bu metotta *train* parametresi diğer metotlardaki gibi eğitim veri kümesini temsil ederken, *order* kullanılacak AR parametresi sayısı, farklılıkları ve MA parametresi sayısını ifade eden modelin *(p,d,q)* sıralamasını göstermektedir. Modelde kullanılan parametrelerin optimum değerlerinin hesaplanması için oluşturulan model üzerinden kütüphanenin *fit(trend='nc', transparams=false, seasonal_periods=s_perod)* metodu kullanılmıştır. Burada *transparams* verilerdeki durağanlığın sağlanması için parametreleri dönüştürmeyi veya dönüştürmemeyi, trend parametresi bir sabitin modele dahil edilip edilmeyeceğini ('c': sabit içerir, 'nc' sabir içermez), *seasonal_periods* ise verinin tam bir mevsimsel döngüdeki dönem sayısını ifade eder. Daha önce de bahsedildiği gibi durağanlık, verideki değişimlerin istikrarlı olmasıdır, mevsimlsel değişimler, dönüşümler gibi [17]. Sayısal çalışmalarda kullanılan veri kümelerinin sezon bilgileri bilinmektedir ARIMA modelinde bu bilgi parametre olarak kullanılmıştır. Verinin mevsimsel dönem sayısı kadar optimum parametreler kullanılarak modelden elde edilecek tahminin gerçekleştirilmesi için ise kütüphanenin forecast(s_period) metodundan faydalanılmıştır.

3.3.3. Modellerin performanslarının karşılaştırılması

Hangi modelin veri setinde daha iyi tahmin sonuçları verdiğine karar verebilmek için doğrulama yapılması gerekmektedir. Bu doğrulama eğitim veri kümesinden test için ayrılan gerçek değer ile modellerden bu dönem için çıkan tahmin sonuçları arasında

gerçekleştirilmiştir. Test seti ve tahmin setlerinin doğruluğu eşitlik 3.6'da gösterildiği gibi hesaplanan MSE (ortalama hatanın karesi) hata ölçüm parametresi ile gerçekleştirilmiştir.

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2 \tag{3.6}$$

MSE, hata ölçüm değerleri içinde yaygın kullanıma sahiptir ve tahmin hata değerlerinin karesini alarak büyük hatalara daha fazla ağırlık verir. Çok büyük veya aykırı tahmin hatalarının karesini aldığı için, daha büyük ortalama kare hata değeri ile sonuçlanır. Dolayısıyla büyük hata yapan modellerin performansını daha düşük olarak ölçer. Ölçülen değer sıfıra ne kadar yakınsa model performansı o kadar iyidir yorumu yapılabilir.

3.3.4. Yöntemlerin karşılaştırılması

Şekil 3.1'deki akışta gösterildiği gibi üç farklı yöntemde, farklı veri seçme metotları ile eğitim veri kümleri oluşturulmuş daha sonra bu veri kümelerinin her biri belirlenmiş olan tahmin modellerini eğitmek için kullanılmıştır. Yöntemlerin hepsinde bütün modeller için ayrı ayrı tahmin sonuçları elde edilmiştir. Şekil 3.1'deki akışta da belirtildiği gibi her bir yöntemin kendi içinde en iyi tahmini yapan model belirlenmiş daha sonra bütün yöntemler arasında en iyi tahmin sonucu veren yönteme karar verilmiştir. Bu şekilde çalışmada önerilen CUSUM yönteminin zaman serilerinin gelecek dönemlerini tahmin etmek için kullanılacak eğitim kümesinin oluşturulmasında kullanılmasının faydası değerlendirilmiştir.

4. SAYISAL ÇALIŞMA

Tüm zaman serileri tahminleri Şekil 3.1'de bahsedildiği gibi üç farklı akış ile gerçekleştirilmiştir. Tahmin modellerinde girdi olarak kullanılacak verinin seçimi aşamasında bu akışlar değişiklik göstermektedir. Birinci olarak, zaman serisinin tahmin için en anlamlı bölümünü belirlemeyi amaçlayan CUSUM algoritması kullanılarak tahmin sonuçları elde edilip gerçek verilere göre MSE hata değerleri hesaplanmıştır. İkinci olarak Sabit Süreli Zaman pencereleri kullanılarak oluşturulan eğitim kümeleri ile tahminler gerçekleştirilmiş ve MSE hata değerleri hesaplanmıştır. Son olarak da varsayılan yöntem olarak ifade edilebilecek olan tüm zaman serisinin tahmin modelini eğitmek için kullanılmış ve tahminin doğruluğunu ölçen MSE değerleri hesaplanmıştır. Bu üç algoritmanın detayları bir örnek üzerinden açıklanmıştır. Çalışmada kullanılan tüm veri setleri için özet sonuçlar Tablo 4.5'te verilmektedir Her bir yaklaşım MSE hata ölçütü kullanılarak değerlendirilmiştir. İlk aşamada, yukarıda bahsedilen metotlar kullanılarak seçilen veri, "eğitim" ve "test" verisi olarak ikiye bölünmüştür. Bölünme tahmin edilecek olan dönem kadar gerçek gözlem verisinin serinin en sonundan test için ayrılması esasına göre yapılmaktadır. Akış diyagramında da bahsedildiği gibi tahmin modellerinde girdi olarak kullanılacak *doğru* eğitim veri kümesinin oluşturulması için birinci yöntemde, örnek zaman serisi kümesine tez çalışmasının temelini oluşturan CUSUM algoritması uygulanmıştır.

CUSUM algoritması ile veri kümesindeki tüm değişim noktaları özyinelemeli olarak tespit edilmiş ve bulunan son değişim noktasından itibaren veri kümesi ayrılmıştır. Bu ayrılmış veri kümesi, tahmin modellerinde eğitim veri kümesi olarak kullanılmış ve tahminlerin gerçeğe yakınlığını ölçmek için test veri kümesi elemanları kullanılarak MSE değerleri hesaplanmıştır. Test veri kümesi daha önce de bahsedildiği gibi zaman serisinin içerisinden tahminlerin performansını ölçmek için ayrılmış olan gerçek gözlem değerlerinden oluşmaktadır. Test veri kümesinin boyutu gerçekleştirilecek tahmin dönemi kadar olmalıdır. Çalışma kapsamında belirlenmiş olan tahmin dönem boyutu zaman serisinin periyodu kadar olacak şekildedir. Zaman serisinde bu bahsedilen periyot kavramı, serideki mevsimsel veya döngüsel dalgalanmanın süresi olarak belirtilebilir

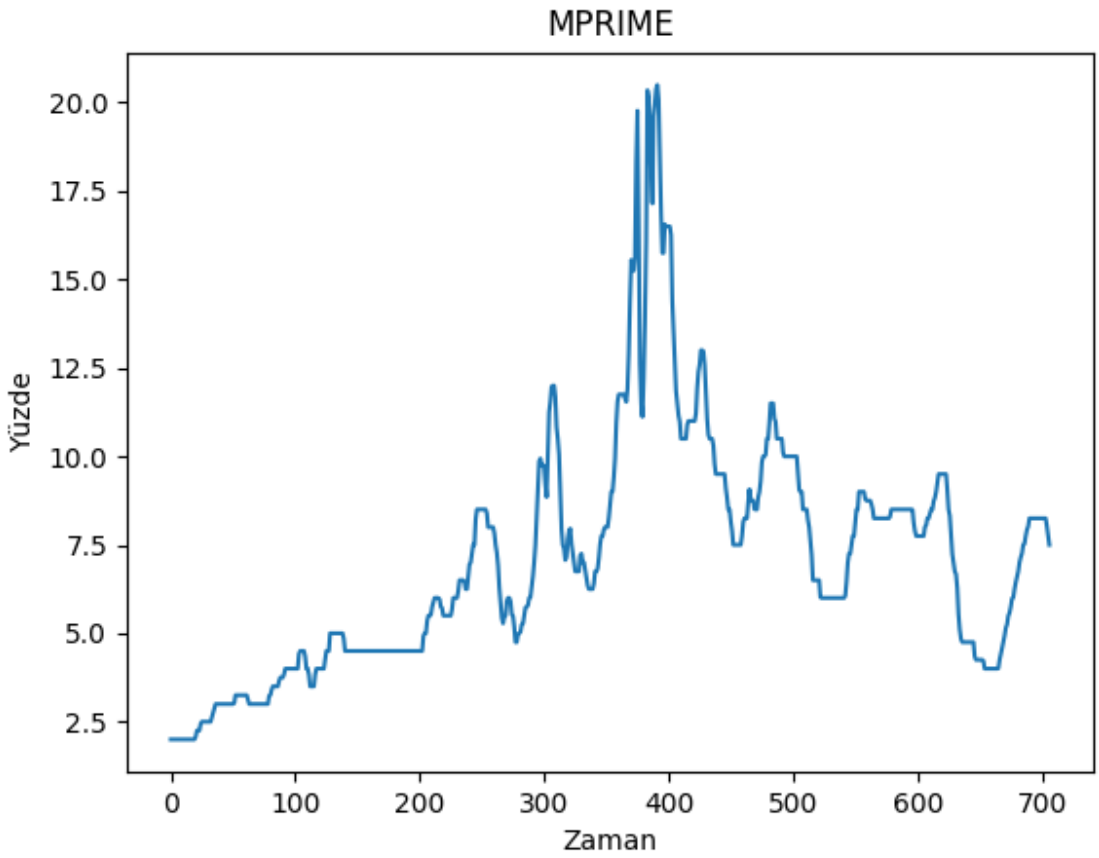
Sabit zaman pencereli veri kümeleri olarak ifade edilebilecek olan ikinci yöntemde ise, zaman serisinin periyodunun üç, dört ve beş katı kadar gözlem verisi ile eğitim kümeleri oluşturulmuştur. Bu alternatif yöntemin çalışmaya dahil edilmesindeki motivasyonlardan birisi, CUSUM algoritması kullanılmasaydı zaman serisi veri kümesinin *en doğru* eğitim

veri kümesi nasıl elde edilebilirdi sorusuna cevap bulunmaya çalışılmasıdır. Diğer bir motivasyon ise eğitim veri kümesinin boyutu değiştikçe yapılan tahminin performansının daha iyi yorumlanabilmesinin sağlanmak istenmesidir. Buna ek olarak bahsedilen alternatif yöntem zaman serisindeki aktüeliteden çok değişim noktalarının tespit edilmesinin tahmin sonuçlarını iyileştirdiğini doğrulamak için CUSUM algoritmasının kullanılmasının önemini kanıtlamak amacıyla da kullanılmıştır. Bu sabit katsayılarla oluşturulan eğitim veri kümeleri belirlenmiş olan tahmin modellerini eğitmek için kullanılmış ve diğer yöntemlerde de kullanılan test veri kümesi kullanılarak performans ölçüt değeri olan MSE değerleri hesaplanmıştır.

En son yöntem olarak da zaman serisinin tamamı tahmin modellerinde eğitim veri kümesi olarak kullanılmış ve sonuç olarak elde edilen öngörü seti ile test veri kümesi verileri arasındaki MSE değerleri hesaplanmıştır. İlk iki yöntemde her zaman verinin daha güncel olan ve "Doğru Eğitim Veri Kümesi" olarak isimlendirilebilecek olan kısmı bulunmaya çalışılmış ve bu setlerden elde edilen tahminlerin tüm zaman serisine göre daha iyi sonuçlar vereceği savunulmuştur. "ICMC- USP Time Series Prediction Repository" kaynağından alınan gerçek zaman serilerinin hepsi için Şekil 3.1'deki diyagramın adımları gerçekleştirilmiş ve serilerin her biri için "Tahmin Model Sonuç Tablosu" oluşturulmuştur. Bahsedilen tüm işlemlerde kullanılan zaman serilerinin periyot değerleri "ICMC- USP Time Series Prediction Repository" kaynağında seri için belirtilmiş olan değerlerdir.

4.1. MPRIME Veri Kümesi

Çalışmada kullanılan metodoloji daha iyi gösterebilmek amacıyla örnek bir veri üzerinden yaklaşımın ayrıntılı ara aşamaları verilmektedir. Örnek olarak seçilen, MPRIME veri kümesi 707 gözlemden oluşmaktadır. Aylık banka kredi oranlarından oluşan MPRIME zaman serisi Şekil 4.1'de grafik olarak gösterildiği gibidir. Bu zaman serisi üzerinde, bahsedilen üç yöntem sırasıyla gerçekleştirilmiş ve her bir yöntemin performansları ölçülmüştür.



Şekil 4.1. MPRIME Zaman Serisi Grafiği

Zaman serisinin "ICMC- USP Time Series Prediction Repository" kaynağında belirtilen periyot değeri 12dir. Dolayısıyla veri kümesinin son 12 gözlem değeri gerçekleştirilen bütün yöntemlerde tahmin modellerinin doğrulunun ölçülmesi için test verisi olarak ayrılmıştır. Çalışmaların sonucu Tablo 4.4'te gösterildiği gibidir. Bütün yöntemlerin detaylı anlatımı devam eden bölümlerdedir.

4.1.1. CUSUM yöntemi

Öncelikli olarak çalışmada tahmin modellerine entegre olarak kullanılması önerilen, CUSUM algoritması tahmin modellerinde kullanılacak olan doğru eğitim veri kümesinin bulunması amacı ile zaman serisine uygulanmıştır. Bu algoritma modelinin detaylı olarak anlatıldığı 3.2.1 bölümünde bahsedilen "bootstraping" işlemindeki tekrar sayısı 10000 olarak, güven seviyesi değeri ise 9990 olarak belirlenmiştir. CUSUM algoritması bootsraping içerisinde yapılan tekrarlarla verideki bütün değişim noktalarını tespit etmeyi

amaçlamaktadır. Burada yapılan tekrar sayısının çokluğu doğru sonuca ulaşma olasılığını arttırmaktadır. Güven seviyesinin büyüklüğü de en ufak değişikliklerin bile göz önünde tutulmasını sağlamaktadır. Parametrelerin uygulamadaki karşılığını daha detaylı anlatmak gerekirse, orijinal seride yer alan gözlem noktaları rastgele bir şekilde yeniden sıralanarak yeni bir seri elde edilmiştir. Bu işlem 10000 defa tekrarlanarak orijinal seriden, 10000 farklı seri oluşturulmuştur. Bu serilerin her biri için CUSUM yöntemi uygulanmıştır. Bu 10000 tekrarın sonucunda veri kümesinde değişiklik tespit edildiği sürece yani 3.2.1 bölümünde bahsedilen eşitlik 3.5'in sonucunun güven seviyesinden büyük çıkması durumunda yinelemelere devam edilmiştir. Her yinelemede değişim noktası tespit edildiği için tespit edilen değişim noktasından itibaren veri kümesi ayrılıp kalan kısım üzerindeki yinelemeye devam edilmiştir. Gerçekleştirilen tekrarlar sonucunda verideki son değişim noktası 653. gözlem verisi olarak belirlenmiş ve bu noktadan itibaren veri kümesi ayrılarak 53 gözlem verisi içeren eğitim kümesi elde edilmiştir. Bu yöntem ile oluşturulmuş olan eğitim veri kümesi çalışmada belirlenmiş olan Holt's Winter (Additive ve Multiplicative) ve ARIMA modellerinin eğitilmesinde kullanılıp sonuç olarak test veri kümesi kadar öngörüler gerçekleştirilmiştir. Gerçekleştirilen tahminler Tablo 4.1'de gösterildiği gibidir ve test için ayrılmış olan gerçek veriler kıyaslandığında en iyi sonucu 0.025 ile ARIMA modeli vermiştir.

Tablo 4.1. CUSUM Yöntemi Tahmin Sonuçları Tablosu

	MSE		
	ARIMA	**HW(A)**	**HW(M)**
CUSUM	0.025	0.696	0.354

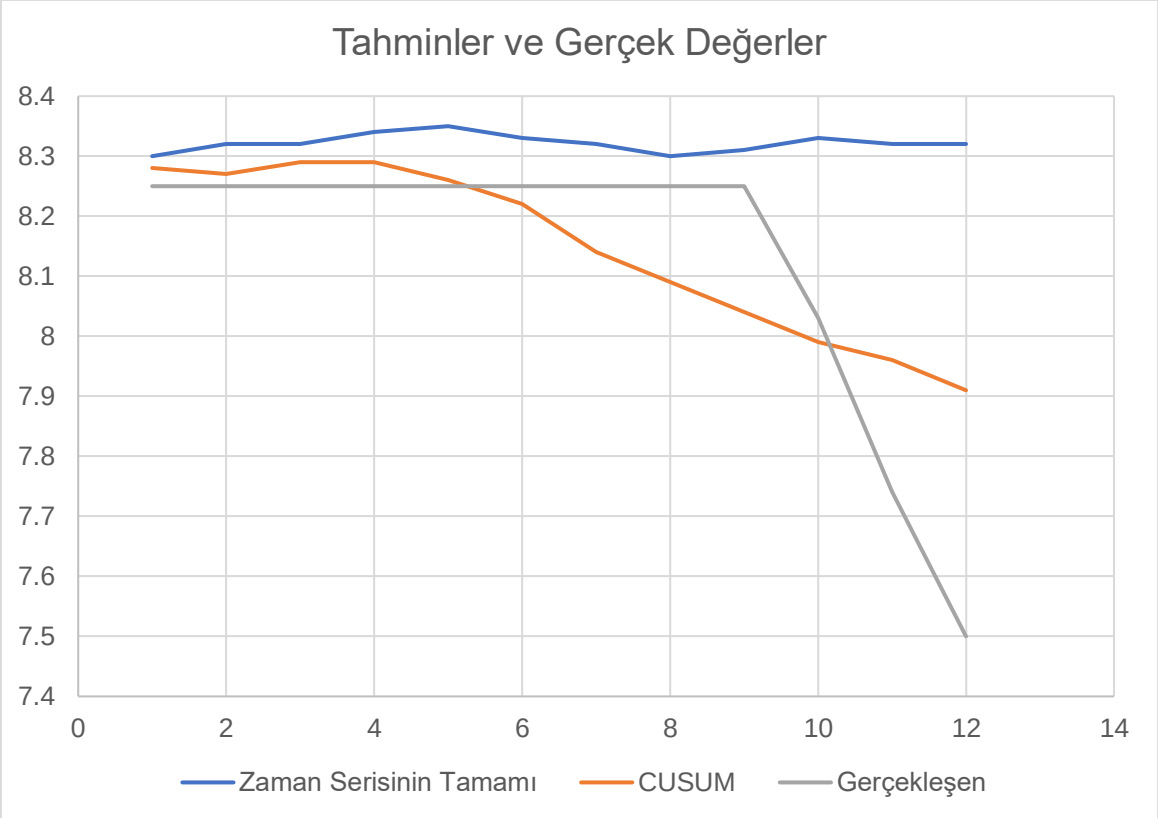
Tablo 4.2. Varsayılan Yöntem Tahmin Sonuçları Tablosu

	MSE		
	ARIMA	**HW(A)**	**HW(M)**
Tüm Veri	0,095	0,223	1,382

Zaman serilerinde tahmin yapılırken varsayılan yöntem tüm zaman serisinin eğitim kümesi olarak kullanılmasıdır. Bu çalışmada önerilmiş olan CUSUM algoritması ile varsayılan yöntemden ayrı ayrı elde edilmiş olan tahminler ve bunların gerçek gözlem

verilerinden oluşan test verisi ile karşılaştırılması Şekil 4.2'de gösterilmiş olan grafikteki gibidir. Grafikte Y ekseni tahmin değerlerini, X ekseni ise zamanı göstermektedir. İki veri kümesi için de 12 aylık tahminler yapılmış ve bu tahminler grafiğe yansıtılmıştır. Şekil 4.2'de "Zaman Serisinin Tamamı" etiketi ile varsayılan yöntemden elde edilen tahmin sonuçları, "CUSUM" etiketi ile de çalışmada önerilen veri seçimi ile gerçekleşen tahmin sonuçları gösterilmektedir. Bu iki sonucun gerçek verilere olan yakınlığını ifade etmek için ise "Gerçekleşen" etiketi ile grafiğe eklenen gerçek gözlem değerlerinden oluşan test veri kümesi gösterilmiştir.

Grafikten de anlaşılacağı gibi CUSUM algoritması ile yapılan en iyi tahmin 0.025 MSE hata değeri ile tahmin yaparken, zaman serisinin tamamı kullanıldığında 0.095 MSE hata değerinde tahmin yapılmıştır.



Şekil 4.2. MPRIME Zaman Serisi Tahmin Sonuçları

4.1.2. Sabit süreli zaman pencereleri yöntemi

Diğer bir yöntem olarak, CUSUM algoritmasına alternatif yapılan verinin değişimlerini gözetmeksizin sadece güncelliğine göre belirlenmiş olan sabit katsayılara göre ayrılan gözlem değerlerinden boyutları 36(12*3), 48(12*4) ve 60(12*5) olacak şekilde eğitim veri kümeleri oluşturulmuştur. Bu eğitim veri kümelerinin her biri için Holt's Winter modelleri ve ARIMA modeli ile tahminler gerçekleştirilmiştir. Gerçekleştirilen öngörüler başta ayrılmış olan test verileri ile MSE hesaplaması yapılmış ve performansları ölçülmüştür. Verinin 3 katı tutulduğunda en iyi sonucu 0.150 MSE değeri ile ARIMA modelinin verdiği, 4 katına kadar güncel veri eğitim kümesi olarak kullanıldığında 0.106 MSE değeri ile yine ARIMA modelinin en iyi sonucu verdiği görülmüştür. Son olarak verinin 5 katı kadar güncel gözlemlerle tahmin yapıldığında ise CUSUM algoritması ile elde edilen değere en yakın olan 0.031 MSE değeri ile Holt's Winter (Additive) modelinin en gerçek sonucu verdiği gözlenmiştir. Tüm zaman serisine tahmin modelleri uygulandığında ise 0.095 MSE hata değeri ile en iyi sonucu ARIMA modelinin verdiği Tablo 4.3'te görülmektedir.

Tablo 4.3. Sabit Zaman Pencereli Yöntemi Tahmin Sonuçları Tablosu

	MSE		
	ARIMA	**HW(A)**	**HW(M)**
Periyot*3	0,150	1,191	1,019
Periyot*4	0,106	0,814	1,594
Periyot*5	0,086	0,031	1,008

4.1.3. Yöntemlerin kıyaslanması

MPRIME zaman serisi için çalışmada savunulduğu gibi CUSUM ile verinin değişim noktaları gözetilerek oluşturulan eğitim veri kümesi en iyi tahmin sonucu verdiği Tablo 4.4'te gösterilmiştir. Çalışmada CUSUM metodunun kullanılmasının asıl amacı olan "Doğru Eğitim Veri Kümesini Bulma" hedefini gerçekleştirdiği yine aynı tablodan anlaşılmaktadır. Sabit katsayılı zaman çerçeveleri ile ayrılan kümeler ise CUSUM gibi otomasyona çevrilmiş bir algoritma uygulanmadan verinin en güncel en anlamlı kısmının manuel olarak

bulunmasına yardımcı olduğu ve değişen katsayılarda CUSUM'ın oluşturduğu eğitim veri kümesine en yakın olan eğitim kümesinin daha iyi sonucu vermesiyle doğrulanmıştır. Sadece verinin eskileşmesi göz önünde bulundurulduğunda elde edilecek sonucun Tablo 4.4'teki *PERİYOT*3* ile etiketlenen satırında belirtildiği gibi gerçekten en uzak tahmini yaptığı görülmüştür. Amaç burada sadece en güncel veri kümesi ile tahmin gerçekleştirmek değil zaman serisinin gelecek dönemlerini tahmin ederken en anlamlı kısmını elde etmektir ve bu tabloda gösterildiği gibi elde edilen sonuçlarla da doğrulanmıştır.

Tablo 4.4. MPRIME Tahmin Sonuçları Tablosu

	EĞİTİM KÜMESİ BOYUTU	**MSE**		
		ARIMA	**HW(A)**	**HW(M)**
CUSUM	53	0,025	0,696	0,354
PERİYOT * 3	36	0,150	1,191	1,019
PERİYOT * 4	48	0,106	0,814	1,594
PERİYOT * 5	60	0,086	0,031	1,008
TÜM VERİ	700	0,095	0,223	1,382

4.1.4. Yaklaşımın gerçek veri setlerine uygulanması

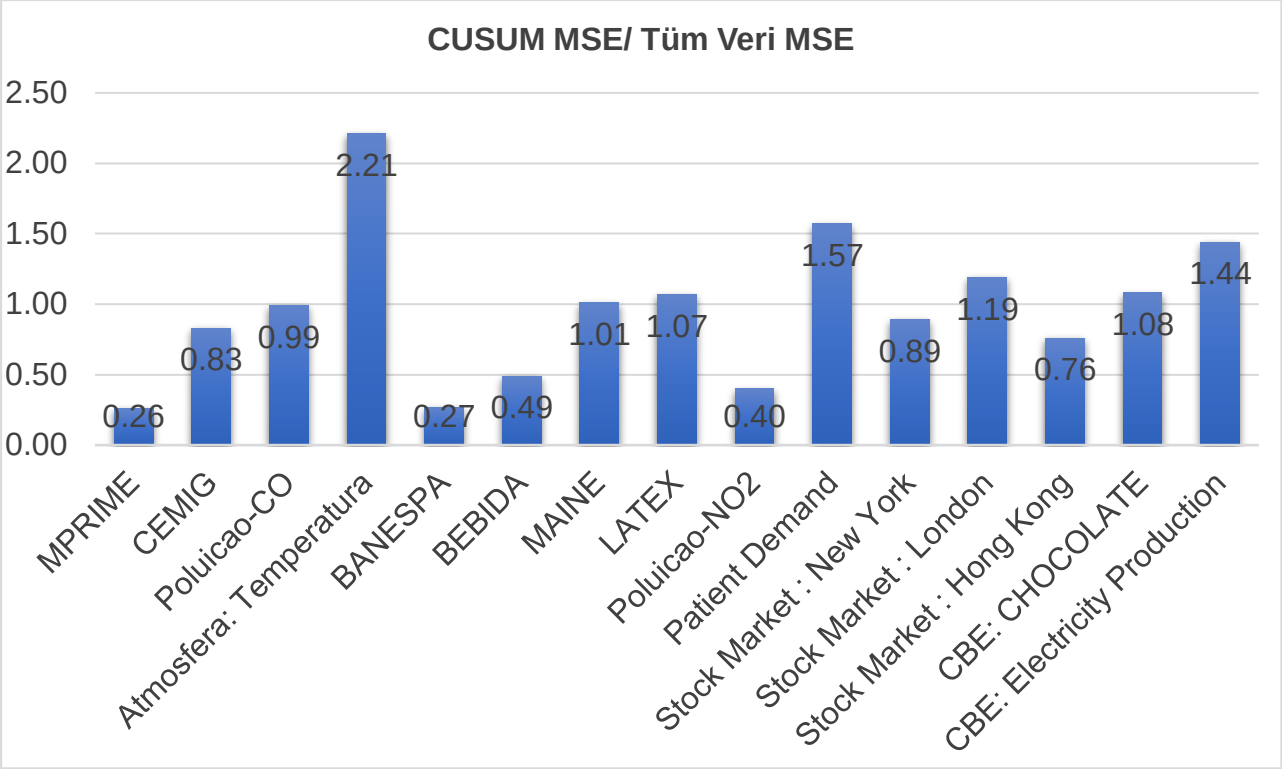
"ICMC-USP Time Series Prediction Repository" kaynağındaki 15 gerçek zaman serisi verisi için çalışma kapsamında belirlenen akış diyagramı adımları tek tek uygulanarak bu verilerle ilgili tahminler gerçekleştirilmiştir ve elde edilen sonuçlar Tablo 4.5'te gösterilmiştir. Tabloda VS olarak gösterilen alanlar eğitim veri kümesi olarak kullanılan gözlemlerden oluşan "Veri Sayısı" miktarını ifade etmektedir. Tabloda bütün veri kümeleri için uygulanmış olan yöntemlerden (ARIMA, Holt's Winter (Additive), Holt's Winter (Multiplicative) en iyi tahmin sonucunu veren metot ve yapılan öngörünün gerçeğe ne kadar yakın olduğunu gösteren MSE değerleri gösterilmiştir. Tablo 4.5'teki verilerden de görüldüğü üzere yapılan sayısal çalışmalarda çoğunlukla CUSUM algoritması ile zaman serisinin tahmin için en doğru veya anlamlı bölümünün tanımlanıp eğitim kümesi olarak kullanılması daha iyi tahminlerin elde edilmesini sağlamıştır. Tablo 4.5'in özeti olarak Şekil 4.3'te CUSUM algoritması entegreli tahmin modellerin MSE hata ölçüm değerlerinin,

zaman serisinin tamamını kullanan tahmin modellerinin MSE hata ölçüm değerlerine oranı grafik olarak yansıtılmıştır. Bu oranın 1'den küçük olması CUSUM algoritması entegreli tahmin modelinin daha küçük hata oranı verdiğini göstermektedir. Şekil 4.3'teki grafikten de anlaşılabileceği gibi yapılan 15 çalışmada 4 zaman serisi kümesinin 1'in üzerinde sonuçlar verdiği görülmüştür. Bunlardan örnek olarak "Patient Demand" ve "CBE: Electricity Production" veri kümelerinin bu sonucu elde etme nedenlerinin anlaşılması için CUSUM algoritması içerisinde çalışma başında belirlenmiş olan "Yineleme (Bootstraping) Sayısı" ve "Güven Aralığı Değeri" parametreleri değiştirilerek tahminler yenilenmiştir.

Tablo 4.5. Tüm Zaman Serilerinin Tahmin Sonuç Tablosu

		CUSUM			3'LÜ PERİYOT			4'LÜ PERİYOT			5'Lİ PERİYOT			TÜM VERİ		
		VS	MSE	METOT	VS	MSE	METOT	VS	MSE	METOT	VS	MSE	METOT	VS	MSE	METOT
1	MPRIME	53	0,025	ARIMA	36	0,150	ARIMA	48	0,106	ARIMA	60	0,031	HW(A)	700	0,095	ARIMA
2	CEMIG	45	0,406	HW(M)	21	0,315	ARIMA	28	0,287	ARIMA	35	0,440	HW(M)	1492	0,491	ARIMA
3	Poluicao-CO	98	0,549	ARIMA	21	1,070	ARIMA	28	1,374	ARIMA	35	1,409	HW(A)	357	0,557	HW(M)
4	Atmosfera: Temperatura	56	1,552	HW(M)	21	-	-	28	2,550	HW(M)	35	0,737	ARIMA	358	0,703	HW(M)
5	BANESPA	21	0,890	HW(A)	21	0,890	HW(A)	28	2,243	HW(M)	35	0,598	ARIMA	1493	3,324	ARIMA
6	BEBIDA	71	68,890	ARIMA	36	353,047	HW(M)	48	124,914	HW(A)	60	99,322	HW(A)	181	140,832	ARIMA
7	MAINE	72	0,142	ARIMA	36	0,933	HW(M)	48	0,111	ARIMA	60	0,106	ARIMA	122	0,141	ARIMA
8	LATEX	60	12,239	IIW(A)	36	14,27	IIW(M)	48	12,397	IIW(A)	60	12,239	IIW(A)	192	11,463	IIW(A)

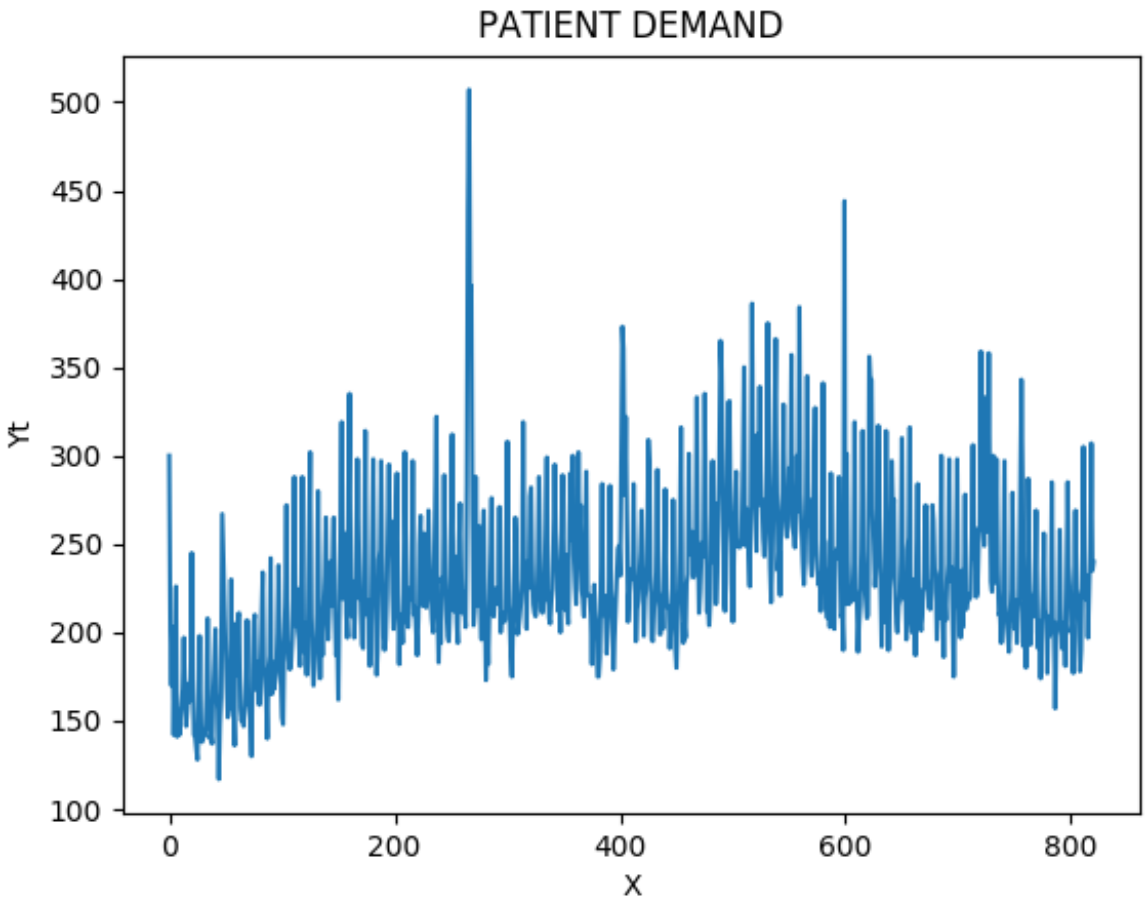
		CUSUM			3'LÜ PERİYOT			4'LÜ PERİYOT			5'Lİ PERİYOT			TÜM VERİ		
		VS	MSE	METOT	VS	MSE	METOT	VS	MSE	METOT	VS	MSE	METOT	VS	MSE	METOT
9	Poluicao-NO2	111	465,199	HW(A)	21	-	-	28	1.968,334	ARIMA	35	3.309,010	HW(M)	357	1.152,476	HW(A)
10	Patient Demand	189	389,816	HW(A)	21	480,237	HW(A)	28	812,622	HW(A)	35	357,045	HW(A)	821	248,875	HW(A)
11	Stock Market: New York	29	226,432	HW(a)	21	363,621	HW(M)	28	226,432	HW(A)	35	300,475	ARIMA	3128	253,520	ARIMA
12	Stock Market: London	49	5.321,718	HW(M)	21	-	-	28	1.968,306	HW(A)	35	3.066,324	ARIMA	3128	4.474,743	HW(M)
13	Stock Market: Hong Kong	52	83.468,789	ARIMA	21	-	-	28	56.919,629	HW(A)	35	59.763,739	HW(M)	3128	110.436,760	HW(A)
14	CBE: CHOCOLATE	48	901.692,680	HW(A)	36	1.427.453,140	HW(A)	48	901.692,680	HW(A)	60	1.019.722,060	HW(A)	396	832.089,860	HW(A)
15	CBE: Electricity Production	70	94.822,130	HW(A)	36	4.582.685,110	HW(M)	48	179.666,240	HW(A)	60	92.799,740	HW(A)	396	65.816,000	HW(A)



Şekil 4.3. MPRIME CUSUM MSE/ Tüm Veri Kümesi MSE Oran Grafiği

Çalışmanın temel önerisine göre beklenen performansları göstermeyen veri kümelerinden biri olan Patient Demand zaman serisi Şekil 4.4'te gösterildiği gibidir. Zaman serisi istisnai durumlar dışında CUSUM algoritmasının kullanılmasını gerektirecek olan değişimlere sahip olmadığı Şekil 4.4'ten de anlaşılmaktadır. CUSUM algoritmasında kullanılan "Yineleme Sayısı" değerinin 10000 olduğu ve "Güven Seviyesi Değeri"nin yani veride değişim olup olmamasının kararının verildiği değerin 9990 olduğu durumda elde edilen sonuçlar Tablo 4.6'da, "Yineleme Sayısı" değerinin 100000 olduğu ve "Güven Seviyesi Değeri"nin 99000 olduğu durumda elde edilen sonuçlar Tablo 4.7'de gösterildiği gibidir. Sonuçlardan da anlaşılacağı gibi tekrar sayısı artıkça elde edilen eğitim kümesi azalmış ve tahmin sonucunun performansı da kötüleşmiştir.

Eğitim veri kümesi tekrar sayısı arttıkça verideki en ufak değişikliklere bile hassasiyetin artması nedeniyle olsa da bu veri kümesindeki değişimler kendini tekrarlayan aslında serideki değişkenin örüntüsü olan değişimlerdir. Dolayısıyla Şekil 4.4'te gösterildiği gibi veri kümesinde istisnai olan belki zaman boyutu dışında başka parametrelere bağlı değişimler dışında bir değişim olmaması ve bu nedenle zaman serisi kümesinin tamamının tahmin modeli için anlamlı veri kümesi haline gelmiştir.



Şekil 4.4. Patient Demand Zaman Serisi Grafiği

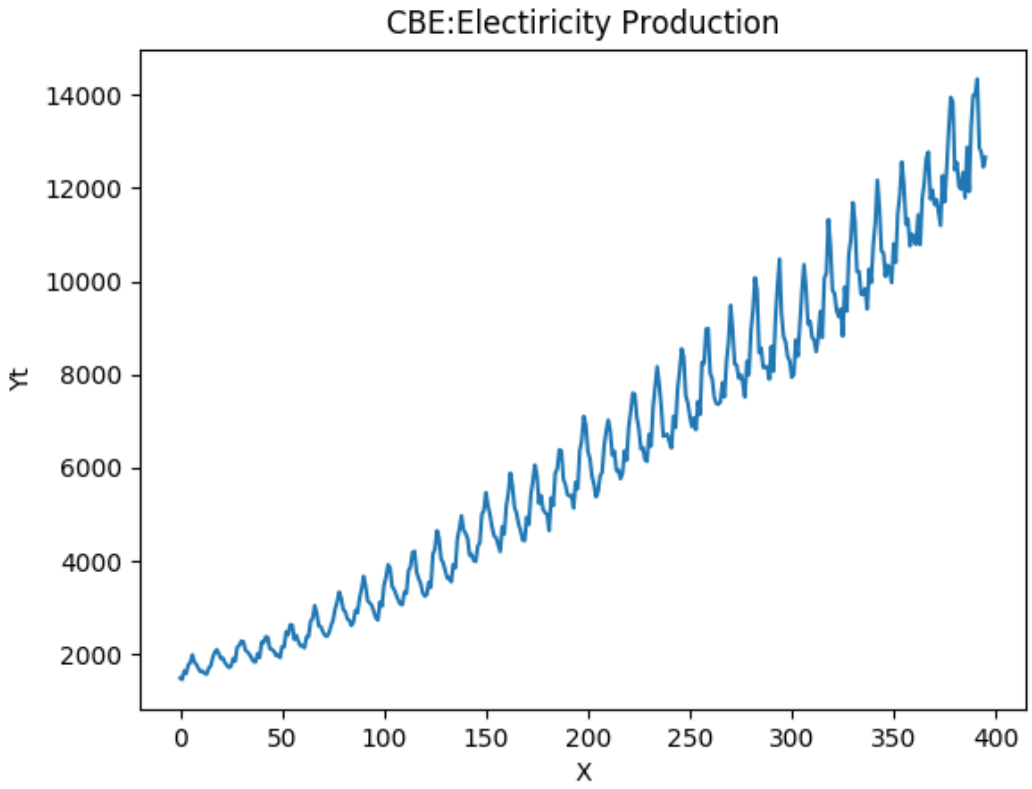
Tablo 4.6. Patient Demand Zaman Serisi İlk Tahmin Sonuçları

Yineleme = 10000 GL = 9990	**EĞİTİM KÜMESİ BOYUTU**	**MSE**	**EN İYİ TAHMİN METODU**
CUSUM	189	389,816	HW(A)
PERİYOT * 3	21	480,237	HW(A)
PERİYOT * 4	28	812,622	HW(A)
PERİYOT * 5	35	357,045	HW(A)
TÜM VERİ	821	248,875	HW(A)

Tablo 4.6. Patient Demand Zaman Serisi Yinelenen Tahmin Sonuçları

Yineleme = 100000 GL = 99000	**EĞİTİM KÜMESİ BOYUTU**	**MSE**	**EN İYİ TAHMİN METODU**
CUSUM	77	558,816	HW(A)
PERİYOT * 3	21	480,237	HW(A)
PERİYOT * 4	28	812,622	HW(A)
PERİYOT * 5	35	357,045	HW(A)
TÜM VERİ	821	248,875	HW(A)

Çalışmada CUSUM algoritması ile oluşturulan eğitim kümesinin daha düşük performanslı sonuç verdiği bir diğer zaman serisi kümesi ise "CBE: Electricity Production" olmuştur. Zaman serisi Şekil 4.5'te gösterildiği gibidir. Bu zaman serisi kümesinde CUSUM algoritması çalışmada belirlenmiş olan parametrelerle uygulandığında Tablo 4.7'deki MSE değerleri elde edilmiştir ve CUSUM algoritması tüm verinin kullanılmasından daha düşük performanslı bir tahmin gerçekleştirmiştir. Bunun üzerine CUSUM algoritması parametreleri değiştirilip yeniden çalıştırıldığında Tablo 4.8'de gösterilmiş olan sonuçlar elde edilmiştir burada elde edilen sonuçlarda MSE değerinde Tablo 4.7'dekine göre azalma görülse de tüm verinin kullanılması ile elde edilen tahmin daha gerçeğe yakın olmuştur. Bu da Şekil 4.5'te görüleceği gibi zaman serisinde artan bir trend vardır ve gözlem değerlerindeki değişimler de belirli bir paterne sahiptir. Dolayısıyla veriyi bölmek bu trendin kaybına sebep olmaktadır. Zaman serisinin tamamı kullanılarak gözlem değerlerinin anlamı kaybedilmeden tahmin yapılması gerekmektedir.



Şekil 4.5. CBE: Electiricity Production Zaman Serisi Grafiği

Tablo 4.8. CBE: Electiricity Production İlk Tahmin Sonuçlar

Yineleme = 10000 GL = 9990	**EĞİTİM KÜMESİ BOYUTU**	**MSE**	**EN İYİ TAHMİN METODU**
CUSUM	70	94822,13	HW(A)
PERİYOT * 3	36	4582685,11	HW(M)
PERİYOT * 4	48	179666,24	HW(A)
PERİYOT * 5	60	92799,74	HW(A)
TÜM VERİ	396	65816,00	HW(A)

Tablo 4.9. CBE: Electiricity Production Yinelenen Tahmin Sonuçları

Yineleme = 100000 GL = 99000	**EĞİTİM KÜMESİ BOYUTU**	**MSE**	**EN İYİ TAHMİN METODU**
CUSUM	60	92799,74	HW(A)
PERİYOT * 3	36	4582685,11	HW(M)
PERİYOT * 4	48	179666,24	HW(A)
PERİYOT * 5	60	92799,74	HW(A)
TÜM VERİ	396	65816,00	HW(A)

5. SONUÇ VE ÖNERİLER

Zaman serileri bir değişkenin tarihsel süreçteki gözlenen değerleridir. Dolayısıyla bu veriler kullanılarak yapılacak olan tahminlerin daha doğru sonuçlar vermesi için bu değişimlerin analiz edilip kronolojik olarak verinin tahmin modeli için anlamlı olan son kısmı bulunması ve sadece bunun üzerinden tahmin gerçekleştirilmesi bu çalışmanın temelini oluşturmuştur. Serideki anlamlı kısmın bulunması ile güncel trendin ne olduğu tespit edilmeye çalışılmıştır. Veri kümesinin anlamlı son bölümünün belirlenmesinde kullanılmak üzere değişim noktası analizi algoritmalarından CUSUM algoritması seçilmiştir. Önerilen algoritmada uygulanan metodoloji CUSUM yöntemi ile tahmin modellerinin entegrasyonunu kapsamaktadır. Belirlenen tahmin modelleri olan ARIMA ve Holt's Winter yöntemlerinin ihtiyacı olan girdi veri kümesi CUSUM algoritması ile belirlenip gerçeğe daha yakın gelecek öngörüleri elde edilmeye çalışılmıştır. Zaman serisindeki değişim noktalarının tespit edilmesinin önemini vurgulamak için alternatif olarak Sabit Süreli Zaman Pencereleri ile veri seçim yöntemi de çalışmaya dahil edilmiştir. Bu alternatif yöntemde veriler sadece zaman boyutundaki güncelliğine göre seçilmiştir. Belirlenmiş sabit katsayılar ile serinin periyodunun katları kadar veri tahmin modellerini eğitmek için ayrılmıştır. Bu veri seçiminde verinin değişim noktaları göz önünde tutulmamıştır. Bu iki yöntem tahmin modellerinde tüm zaman serisinin kullanılması ile karşılaştırılmıştır. Sayısal çalışmalar kapsamında "ICMC-USP Time Series Prediction Repository"den alınmış olan 15 gerçek zaman serisi ile metodolojideki algoritma çalıştırılmıştır. Yapılan çalışmaların büyük oranında bu tez kapsamında önerildiği gibi CUSUM algoritması kullanılarak oluşturulan eğitim kümelerinden gerçeğe daha yakın sonuçlar elde edildiği görülmüştür. Bunun matematiksel olarak kanıtı çalışmada önerilen metodolojinin hata ölçüm metriği olan MSE değerlerinin alternatif yönteme göre daha küçük değerlerde sonuç vermesi ile yapılmıştır. CUSUM algoritması ile serinin daha iyi anlaşılması ve tahmin yapılmasının daha doğru sonuçlar verdiğinin doğrulaması ise Sabit Süreli Zaman Pencereleri yöntemi ile yapılmıştır. Bu alternatif yöntemde amaç zaman serisini analiz etmek değil sadece güncel verilerle tahmin gerçekleştirmektir. Buradan elde edilen sonuçlarda yüksek oranda CUSUM algoritmasından daha kötü tahminler gerçekleştirildiği görülmüştür. Yapılan çalışmada amacın sadece verinin son periyodunu takip etmek olmadığı, Sabit Süreli Zaman Pencereleri yönteminde serinin periyodunun üç katı alınarak oluşturulan eğitim kümesinin kullanılmasından elde edilen tahmin sonucunun MSE değerlerinin neredeyse hepsinde

CUSUM MSE değerinden fazla olmasından da anlaşılmaktadır. Çalışmada kanıtlanmaya çalışılan zaman serisinde daha iyi sonuçlar elde etmek için serinin son trendinin bulunup buna göre tahmin yapılması gerektiğidir. Serinin güncel trendinin tespit edilmesi de CUSUM algoritması ile veri analizi yapılması sayesinde gerçekleştirilebileceği örneklerle gösterilmiştir.

Tezde önerildiği gibi CUSUM algoritması yapılan 15 sayısal çalışmanın büyük oranında beklendiği gibi daha başarılı sonuçlar vermiştir. Fakat genellikle verideki değişimlerin sabit olduğu tüm zaman serisinin başarılı olduğu örneklerde alternatif ne gibi yaklaşımlar yapılabileceği başka araştırmaların konusu olmaya adaydır.

KAYNAKLAR

[1] G.D. Montanez, S. Amizadeh, N. Laptev, Inertial hidden Markov models: modeling change in multivariate time series, *AAAI. Conference on Artificial Intelligence*, Şubat, 2015.

[2] Y. Kawahara, M. Sugiyama, Sequential change-point detection based on direct density-ratio estimation, *SIAM. International Conference on Data Mining*, vol.5, Nisan,2009, doi ; 10.1002/sam.10124.

[3] İ. Doğan, N. Doğan, A review of some change-point detection methods, *Türkiye Klinikleri Journal of Biostatistics*, Şubat, 2018, doi ; 10.5336/biostatic.2018-61056.

[4] W.A. Taylor, Change-point analysis: a powerful new tool for detection changes, *Libertyville: Taylor Enterprises*, s.19, Mart, 2000.

[5] R. Adhikar, R. K. Agrawal, An introductory study on time series modeling and forecasting, *LAP. Lambert Academic Publishing*, Germany, Şubat, 2013, doi;10.13140/2.1.2771.8084.

[6] S.O. Ongbali, A. C. Igboanugo, S.A Afolalu, M.O Udo, I.P Okokpujie, Model selection process in time series analysis of production system with random output, *IOP. Conference Series: Materials Science and Engineering*, vol.413., Ekim, 2018. [Çevrimiçi] . Erişilebilir: https://doi.org/10.1088/1757-899X/413/1/012057

[7] P. Chen, A. Niu, D. Liu, W. Jiang, B. Ma, Time series forecasting of temperatures using SARIMA: an example from nanjing, *IOP. Conf. Series: Materials Science an Engineering*, vol.319, no. 5, 2018. [Çevrimiçi]. Erişilebilir: https://doi.org/10.1088/1757-899X/394/5/052024

[8] M. Snipes, D.C. Taylor, Model selection and Akaike information criteria: an example from wine ratings and prices, *Wine Econ. Policy*, vol.3, no.1, s.3-9, Şubat, 2014. [Çevrimiçi]. Erişilebilir : https://doi.org/10.1016/j.wep.2014.03.001

[9] C. Tofallis, A better measure of relative prediction accuracy for model selection and model estimation, *Journal of the Operational Research Society*, vol.66, Mart, 2015, doi; 10.1057/jors.2014.124.

[10] S. Makridakis, S. Wheelwright, R.J. Hyndman, "The Forecasting Perspective" in *Forecasting: Methods and Applications*, Willey, 1998, 3rd Edition.

[11] Veri kümeleri , http://sites.labic.icmc.usp.br/icmc_tspr/index.php/datasets

[12] S. Aminikhanghahi, D.J. Cook, A survey of methods for time series change point detection, *Knowledge and Information Systems*, Eylül, 2016, doi ;10.1007/s10115-016-0987-z.

[13] R. Garnett, M.A. Osborne, S. Reece, A. Rogers, S.J. Roberts, Sequential bayesian prediction in the presence of changepoints and faults, *The Computer Journal*, vol.53, no. 9, s.1430–1446, Eylül, 2010, doi ; 10.1093/comjnl/bxq003.

[14] G. Comert, A. Bezuglov, An online change-point-based model for traffic parameter prediction, *IEEE. Transactions on Intelligent Transportation Systems*, vol.14, no.3, s. 1360-1369, Ekim, 2013, doi; 10.1109/TITS.2013.2260540.

[15] M. Cetin, G. Comert, Short-term trafic flow prediction with regime switching models, *Transp. Res. Rec. J. Transp. Res. Broad*, vol.1965, no.1, s. 22-31, Şubat, 2006, doi; 10.3141/1965-03.

[16] M. Steyvers, B. Scott, Prediction and change detection, *Advances in Neural Information Processing Systems*, vol.18, Şubat, 2006.

[17] S. TRAIANI, "Seasonal stability in time series of zooplankton abundance", Doktora Tezi, Department of Mathematics and Statistics University of Strathclyde, 2010.

[18] P. Mondal, L. Shit, S. Goswami, Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices, *International Journal of Computer Science, Engineering and Applications IJCSEA.*, vol.4, no.2, Nisan, 2014, doi ; 10.5121/ijcsea.2014.4202.

[19] M. Yamada, A. Kimura, F. Naya, H. Sawada, Change-point detection with feature selection in high-dimensional time series, *Twenty-Third Internetional Joint Conference on Artificial Intelligence*, Şubat, 2013.

[20] G. Athanasopoulos, R.A. Ahmed, R.J. Hyndman, Hierarchical forecasts for Australian fomestic tourism, *International Journal of Forecasting*, vol.25, no.1, s.146-166, Şubat-Mart, 2009, doi; 10.1016/j.ijforecast.2008.07.004.

[21] M.Z. Babai, M.M. Ali, K. Nikolopoulos, Impact of temporal aggregation on stock control performance of intermittent demand estimators: empirical analysis, *Omega-International Journal of Management Science*, vol.40, no.6, s.713–721, Aralık, 2012, doi; 10.1016/j.omega.2011.09.004.

[22] A. Bacchetti, N. Saccani, Spare parts classification and demand forecasting for stock control: investigating the gap between research and practice, *Omega- International Journal of Management Science*, vol.40, no.6, s.722–737, Aralık, 2012, doi;10.1016/j.omega.2011.06.008.

[23] B.P. Bhattarai, K.S. Myers, B. Bak-Jensen, I. Mendaza, R. J. Turk, and J.P. Gentle, Optimum aggregation of geographically distributed flexible resources in strategic smart-grid/microgrid locations, *International Journal of Electrical Power & Energy Systems*, vol.92, s.193–201, Kasım, 2017, doi; 10.1016/j.ijepes.2017.05.005.

[24] J.E. Boylan, M.Z. Babai, On the performance of overlapping and non-overlapping temporal demand aggregation approaches, *International Journal of Production Economics*, vol.181, b. A, s.136– 144, Kasım, 2016, doi; 10.1016/j.ijpe.2016.04.003.

[25] I.W. Burr, Statistical quality control methods, vol. 16. CRC Press, *Taylor and Francis Group, Boca Rato* FL., 1976.

[26] B.J. Dangerfield, J.S. Morrıs, Top-down or bottom-up-aggregate versus disaggregate extrapolatıons, *International Journal of Forecasting*, vol.8, no.2, s.233–241, Eylül, 1992, doi; 10.1016/0169-2070(92)90121-O.

[27] G. Fliedner, An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation, *Computers & Operations*

Research, vol.26, no.10-11, s.1133–1149, Ekim, 1999, doi; 10.1016/S0305-0548(99)00017-9.

[28] R.S. Gutierrez, A.O. Solis, S. Mukhopadhyay, Lumpy demand forecasting using neural networks, *International Journal of Production Economics*, vol.111, no.2, s. 409–420, Şubat, 2008, doi; 10.1016/j.ijpe.2007.01.007.

[29] R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, *International Journal of Forecasting*, vol.22, no.4, s.679–688, Eylül- Aralık, 2006, doi; 10.1016/j.ijforecast.2006.03.001.

[30] Y.H. Jin, B.D. Williams, M.A. Waller, A.R. Hofer, Masking the bullwhip effect in retail: the influence of data aggregation, *International Journal of Physıcal Distribution & Logistics Management*, vol.45, no.8, s.814–830, 2015, doi; 10.1108/IJPDLM-11-2014-0264.

[31] N. Kourentzes, On intermittent demand model optimisation and selection, *International Journal of Producion Economics*, vol.156, s.180–190, Eylül, 2014, doi; 10.1016/j.ijpe.2014.06.007.

[32] N. Kourentzes, Petropoulos F., Forecasting with multivariate temporal aggregation: The case of promotional modelling, *International Journal of Production Economics*, vol.181, b. A, s.145–153, Ekim, 2016, doi; 10.1016/j.ijpe.2015.09.01.

[33] J.M. Maheu, T.H. McCurdy, How useful are historical data for forecasting the long-run equity return distribution, *Journal of Business & Economic Statistics*, vol. 27, no.1, s.95–112, Şubat, 2009, doi : 10.2139/ssm.996696.

[34] A.A. Syntetos, Forecasting by temporal aggregation, *Foresight*, vol.34, s.6–11, 2014.

[35] A.A. Syntetos, Z. Babai, J.E. Boylan, S. Kolassa, K. Nikolopoulos, Supply chain forecasting: theory, practice, *Their Gap and The Future, European Journal of Operational Research*, vol.252, no.1, s1–26, Şubat, 2016, doi; 10.1016/j.ejor.2015.11.010.

[36] P. Wallstromand, E. Segerstedt, Evaluation of forecasting error measurements and techniques for intermittent demand, I*nternational Journal of Production Economics*, vol.128, no. 2, s.625–636, Aralık, 2010, doi; 10.1016/j.ijpe.2010.07.013.

[37] H. Widiarta, S. Viswanathan, R. Piplani, Forecasting aggregate demand: an analytical evaluation of top-down versus bottom-up forecasting in a production planning framework, *International Journal of Production Economics*, vol.118, no.1, s.87–94, Mart, 2009, doi: 10.1016/j.ijpe.2008.08.013.

[38] A. Zellner, J. Tobias, A note on aggregation, Disaggregation and Forecasting Performance, *Journal of Forecasting*, vol.19, no.5, s.457–465, Ekim, 2000, doi; 10.1002/1099-131X(200009)19:5<457::AID-FOR761>3.0.CO;2-6.

[39] G. Zotteri, M. Kalchschmidt, A model for selecting the appropriate level of Aggregation in forecasting processes, *International Journal of Production Economics*, vol.108, no.1-2, s.74–83, Şubat, 2007, doi: 10.1016/j.ijpe.2006.12.030.