

T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
FEN FAKÜLTESİ
İSTATİSTİK BÖLÜMÜ

DEĞİŞİM NOKTASI BELİRLEME YÖNTEMLERİ VE UYGULAMALARI

Bitirme Projesi Raporu

Pelin PEKER
Merve AK
Edanur Binnaz DURSUN
Ahmet ÇALI

Mayıs 2024

Rapor Deęerlendirme

“DEęİŐİM NOKTASI BELİRLEME YÖNTEMLERİ VE UYGULAMALARI” baęlıklı bitirme projesi raporu tarafımdan okunmuş, kapsamı ve nitelięi aęısından bir Bitirme Projesi raporu olarak kabul edilmiştir.

Doę.Dr. Engin YILDIZTEPE

Teşekkür

Tüm çalışma süresince yönlendiriciliği, katkıları ve yardımları ile yanımızda olan danışmanımız Doç.Dr. Engin YILDIZTEPE 'ye ve böyle bir çalışmayı yapmamız için bize fırsat tanıyan Dokuz Eylül Üniversitesi Fen Fakültesi İstatistik Bölümüne teşekkür ederiz.

Pelin PEKER
Merve AK
Edanur Binnaz DURSUN
Ahmet ÇALI

Özet

Değişim noktası verilerde meydana gelen beklenmedik anlamlı değişiklikler olarak tanımlanabilir. Değişim noktası tespit yöntemleri bu noktaları istatistiksel tekniklerle bulmayı amaçlar. Değişim noktası analizi finans, kalite kontrol, ağ analizi gibi çok farklı alanlarda kullanılmaktadır. Bu çalışmada değişim noktası tespit yöntemleri incelenmiştir. Çalışma kapsamında AMOC, BinSeg, parçalı regresyon, PELT ve Prophet algoritmaları kullanılmıştır. Algoritmaların uygulamadaki performanslarını belirlemek amacıyla yirmi yapay, on bir gerçek veri ve on beş kademeli eşik egzersiz test verisi kullanılmıştır. Performanslar F1 puanı ve kapsama ölçütü kullanılarak değerlendirilmiştir. F1 puanına göre en başarılı sonuçlar yapay verilerde BinSeg ile, gerçek verilerde parçalı regresyon ile alınmıştır. Kapsama ölçütüne göre ise yapay verilerde BinSeg ve parçalı regresyon, gerçek verilerde parçalı regresyon ile en iyi sonuçlar elde edilmiştir. Değişim noktası içeren yapay veri üretmek ve bahsedilen algoritmaları uygulayabilmek amacıyla bir RShiny web uygulaması geliştirilmiştir. Çalışmada R ve Python programlama dilleri kullanılmıştır.

Anahtar Kelimeler: Değişim noktası, AMOC, BinSeg, PELT, Prophet, Parçalı Regresyon, R shiny

Abstract

Change points can be defined as unexpected significant changes occurring in data. Change point detection methods aim to identify these points using statistical techniques. Change point analysis is utilized in various fields such as finance, quality control, and network analysis. In this study, change point detection methods were examined. AMOC, BinSeg, Segmented Regression, PELT, and Prophet algorithms were used within the scope of the study. In the application, twenty simulated datasets, eleven real datasets and fifteen incremental ramp exercise test datasets were used to determine the performance of the algorithms. Performance was evaluated using the F1 score and cover metric. According to the F1 score, the most successful results were obtained with BinSeg on artificial data and with Segmented Regression on real data. Regarding the cover metric, BinSeg and Segmented Regression yielded the best results on artificial data, while Segmented Regression performed the best on real data. To generate artificial data containing change points and apply the mentioned algorithms, an R Shiny web application was developed. Both R and Python programming languages were used in the study.

Keywords: Changepoint, AMOC, BinSeg, PELT, Prophet, Segmented Regression, R shiny

İçindekiler

GİRİŞ	1
Bölüm 1: Değişim Noktası	3
1.1 Tek Değişim Noktası Tespiti	3
1.2 Birden Fazla Değişim Noktası Tespiti	4
1.2.1 İkili Bölümleme Algoritması	5
1.2.2 Prophet	6
1.2.3 PELT	7
1.2.4 Parçalı Regresyon	10
1.3 Kapsama Ölçütü	13
1.4 F1 Puanı	14
Bölüm 2: Uygulama	17
2.1 Veri	17
2.1.1 Yapay Veri	17
2.1.2 Gerçek Veri	18
2.2 Değişim Noktası Analizi	18
2.2.1 Yapay Veriler için Sonuçlar	19
2.2.2 Gerçek Veriler için Sonuçlar	22
2.2.3 Kademeli Egzersiz Test Verileri	25
2.3 R Shiny Uygulaması	28
2.3.1 R Shiny Nedir?	28
2.3.2 Kullanılan Paketler	28
2.3.3 Anasayfa	29
2.3.4 Yapay Veri Üretme	29
2.3.5 Veri Yükleme	30
Bölüm 3: Sonuç	33
Kaynaklar	35

Tablo Listesi

2.1	Yapay Veriler	17
2.2	Gerçek Veriler	18
2.3	F1 puanı (Varsayılan)	19
2.4	Kapsama Ölçütü (Varsayılan)	19
2.5	F1 Puanı (Ayarlanmış)	20
2.6	Kapsama Ölçütü (Ayarlanmış)	21
2.7	F1 puanı (Varsayılan)	23
2.8	Kapsama Ölçütü (Varsayılan)	23
2.9	F1 Puanı (Ayarlanmış)	24
2.10	Kapsama Ölçütü (Ayarlanmış)	24
2.11	Kapsama Ölçütü (Ayarlanmış)	26
2.12	F1 Puanı (Ayarlanmış)	27

Şekil Listesi

1.1	Kapsama Ölçütü	13
1.2	Yanılı matrisi	15
2.1	$V_E/P_{ET}CO_2$ - Power (W) (Ozkaya, Balci, As ve Yıldiztepe, 2021) . .	26
2.2	Anasayfa	29
2.3	Simüle Edilmiş Veri	29
2.4	Veri Yükleme	30

GİRİŞ

Değişim noktası, verilerde ani ve beklenmedik bir değişikliği ifade eden bir konum olarak tanımlanır. Bu noktalar genellikle bir desenin, trendin, varyansın veya diğer istatistiksel özelliklerin birdenbire ve belirgin bir şekilde değiştiği yerlerdir. Değişim noktası kestirimi; haberleşme, biyomedikal alanlar, konuşma sinyalleri işleme, sismik veri analizi, istatistiksel süreç kontrolü, finansal veri analizi gibi çeşitli alanlarda yaygın olarak kullanılan bir yöntemdir. Bu kestirim probleminin çözümü için çeşitli istatistiksel sinyal işleme teknikleri geliştirilmiştir. Veri dağılımının bilindiği ve bilinmediği durumlarda kullanılan yöntemler, parametrik ve parametrik olmayan olarak sınıflandırılır. Ölçümlere ait dağılım fonksiyonunun bilinmesi, genellikle parametrik değişim noktası kestirim yöntemleriyle zor problemlerde bile başarılı sonuçlar elde edilmesini sağlar. Ancak bu bilgi her zaman mevcut olmayabilir ve bu durumda parametrik olmayan yöntemlere başvurulur.

Örneğin, bir perakende satış verisinde, bir ürünün satışlarının aniden artması veya azalması bir değişim noktasını temsil edebilir. Endüstriyel süreçlerde, üretim hattındaki bir arıza nedeniyle üretimde ani bir düşüş de bir değişim noktası olabilir. Finansal piyasalarda, bir hisse senedinin değerinde ani bir değişiklik veya trendin tersine dönmesi de bir değişim noktasını işaret edebilir. Değişim noktalarını tespit etmek için istatistiksel analiz, zaman serisi analizi, makine öğrenimi gibi teknikler kullanılır. Bayes faktörü, kümülatif toplam, anomalilerin tespiti gibi istatistiksel kriterler, değişim noktalarını belirlemede kullanılan araçlardan bazılarıdır. Bu teknikler, veriyi parçalara ayırır ve her bir parçanın içindeki istatistiksel özellikleri değerlendirerek değişim noktalarını tanımlar.

Bu analizler, verilerdeki önemli değişiklikleri objektif bir şekilde belirleyerek, kullanıcılara olayları anlama ve gelecekteki eğilimleri tahmin etme konusunda yardımcı olabilir. Bu sayede, değişim noktalarının belirlenmesi, karar verme süreçlerini destekleyerek daha bilinçli ve stratejik adımlar atılmasına imkan tanır.

Bölüm 1

Değişim Noktası

1.1 Tek Değişim Noktası Tespiti

AMOC (At Most One Change), bir veride yalnızca bir değişim noktasının varlığını tespit etmeye odaklanan istatistiksel bir analiz yöntemidir(Hinkley (1970)). Bu tür analizler genellikle zaman serileri, süreç kontrolü, finansal veriler gibi çeşitli alanlarda kullanılır. Temel hedef, verilerdeki değişim noktasını tanımlamak ve bu noktada meydana gelen ani değişikliği belirlemektir.

Tek değişim noktası tespiti, bir zaman serisinde veya veri setinde belirli bir anın, örneğin bir trendin başlangıcı veya bir olayın etkisi gibi bir değişiklik noktasını belirlemek için kullanılır(Lykou, Tsaklidis ve Papadimitriou, 2020). Bu analiz genellikle istatistiksel yöntemler, matematiksel modeller veya makine öğrenimi algoritmaları kullanılarak gerçekleştirilir. Bu yöntemler, verilerdeki değişim noktasının istatistiksel olarak anlamlı olup olmadığını değerlendirir ve belirli bir ölçüye dayanarak değişim noktasını tanımlar. Bu tür analizler, anormal durumları tespit etmek, süreçlerdeki değişiklikleri anlamak veya zaman içindeki önemli olayları belirlemek gibi bir dizi uygulama alanında kullanılır. Örneğin, endüstriyel süreçlerde bir makinenin arızasının başlangıcını belirlemek veya finansal piyasalardaki bir trend değişikliğini saptamak gibi durumlar, tek değişim noktası tespiti analizine örnek olabilir. Bu yöntemler, veri analizi ve karar verme süreçlerinde bilinçli ve stratejik adımlar atılmasına yardımcı olabilir, çünkü belirli bir değişim noktasının tanımlanması, olayların anlaşılmasını ve gelecekteki trendlerin tahmin edilmesini destekleyebilir.

Tek bir değişim noktasının tespiti için hipotez testi bir olasılık oranı tabanlı yaklaşımla formüle edilebilir. Burada, H_0 boş hipotezi, değişim noktasının olmadığına ($m = 0$) karşılık gelirken, alternatif hipotez H_1 tek bir değişim noktasına ($m = 1$) karşılık gelir.

Bu hipotezi test etmek için olasılık oranı test istatistiği, genel bir hipotez tabanlı yaklaşımı kullanır. İlk olarak, Hinkley (1970) tarafından önerilen bu yöntem, asimptotik dağılımı türetilen bir olasılık tabanlı yaklaşımı benimser. Bu yaklaşım, normal olarak dağılmış gözlemler içindeki ortalama değişikliği için olasılık oranı test istatistiğini hesaplar.

Jen ve Gupta (1987) tarafından yapılan genişletme ile bu olasılık tabanlı yaklaşım,

normal olarak dağılmış gözlemler içindeki varyans değişiklikleri için de geçerli bir test istatistiği sağlar.

Bu yöntemler, belirli bir değişim noktasının varlığını istatistiksel olarak değerlendirmek ve bu değişim noktasının ne zaman gerçekleştiğini belirlemek için kullanılır. Bu analizler, verilerdeki belirli bir noktadaki değişikliklerin anlamlılığını değerlendirerek, değişim noktasının varlığını istatistiksel olarak doğrulamaya yöneliktir.

Boş hipotezi için maksimum log-olabilirlik, $\log p(y_{1:n}|\hat{\theta})$ şeklinde ifade edilir, burada $p(\cdot)$ verilerin dağılımıyla ilişkilendirilen olasılık yoğunluk fonksiyonudur ve $\hat{\theta}$ parametrelerin maksimum olabilirlik tahminidir.

Alternatif hipotez altında, τ_1 ile bir değişim noktası içeren bir modeli ele alalım, burada $\tau_1 \in 1, 2, \dots, n-1$. Bu durumda, belirli bir τ_1 için maksimum log-olabilirlik şu şekildedir: $ML(\tau_1) = \log p(y_{1:\tau_1}|\hat{\theta}_1) + \log p(y_{(\tau_1+1):n}|\hat{\theta}_2)$. Değişim noktasının doğası göz önüne alındığında, alternatif hipotez altındaki maksimum log-olabilirlik değeri basitçe $\max \tau_1 ML(\tau_1)$ olarak ifade edilir, burada maksimum tüm olası değişim noktası konumları için alınır. Test istatistiği şu şekildedir: $\lambda = 2\tau_1 ML(\tau_1) - \log p(y_{1:n}|\hat{\theta})_{\max}$.

Bu test, $\lambda > c$ ise boş hipotezi reddedilir şeklinde bir eşik değeri c seçerek yapılır. Eğer boş hipotezi reddedilirse, yani bir değişim noktası algılanırsa, onun pozisyonunu $\hat{\tau}_1$ olarak tahmin ederiz, bu değer $ML(\tau_1)$ 'i maksimize eden τ_1 değeridir. Bu parametre için uygun değer c henüz açık bir araştırma sorusudur ve farklı değişim tipleri altında p değerleri ve diğer bilgi kriterleri oluşturan birkaç yazar bulunmaktadır (Birgé ve Massart, 2007; Chen, Gupta ve Gupta, 2000; Guyon ve Yao, 1999; Lavielle, 2005).

Açıkça görülmektedir ki, olabilirlik test istatistiği basitçe m parçaların her biri için olasılığı toplamak suretiyle birden fazla değişime genişletilebilir. Sorun, tüm olası $\tau_{1:m}$ kombinasyonları üzerinde $ML(\tau_{1:m})$ 'in maksimumunu belirlemeye dönüşür.

1.2 Birden Fazla Değişim Noktası Tespiti

Birden fazla değişim noktasının (Multi change points) belirlenmesi, bir veride meydana gelen yapısal değişiklikleri tespit etme sürecini ifade eder. Bu analiz, verilerin farklı parçalara bölünmesi ve her bir parçadaki değişim noktalarının tanımlanması yoluyla gerçekleştirilir. Değişim noktaları, verinin genel özelliklerinde veya desenlerinde beklenmeyen değişiklikleri temsil eder.

Özellikle zaman serilerinde birden fazla değişim noktasının belirlenmesi, farklı dönemlerde farklı trendlerin veya desenlerin varlığını anlamak açısından önemlidir. Bu tür analizler, endüstriyel süreçlerde, finansal piyasalarda veya epidemiyolojik veriler gibi farklı alanlarda meydana gelen önemli olayları veya dönemleri belirlemek için kullanılabilir.

Bu tür analizler genellikle istatistiksel yöntemler, matematiksel modeller veya makine öğrenimi algoritmaları kullanılarak gerçekleştirilir. Birden fazla değişim noktası tespiti, verilerdeki karmaşık yapısal değişiklikleri belirleme ve anlama amacı taşır. Bu da kullanıcılara önemli olayları tespit etme ve verilerdeki farklı bölümlerindeki değişiklikleri anlama imkânı sağlar. Bu analizler, verilerin farklı parçalara ayrılmasını sağlayarak, her bir parçadaki farklı özellikleri ve eğilimleri anlamak için bir yol sunar. Bu da karar verme süreçlerinde daha bilinçli ve stratejik adımlar atılmasına yardımcı

olabilir.

Belirli bir zaman serisi veya sinyal akışındaki birden fazla değişim noktasını verimli ve doğru bir şekilde belirleyebilmek için literatürde yaygın olarak kullanılan bir yöntem, maliyet fonksiyonunu (C) minimize ederek birden çok değişim noktasının konumunu belirlemektir. Bu yöntemde, aşırı uyumu önlemek için bir ceza terimi $\beta f(m)$ ile birlikte maliyet fonksiyonunun toplamı minimize edilmeye çalışılır. Formülü şu şekilde ifade edebiliriz: $\sum_{i=1}^{m+1} C(y(\tau_{i-1} + 1) : \tau_i) + \beta f(m)$

Bu denklem, değişiklik noktaları ($y(\tau_{i-1} + 1)$ ile τ_i arasındaki parçalar) için maliyet fonksiyonunun toplamını ve aşırı uyumu önlemek için ceza terimini içerir. Bu yöntem, verileri birden çok bölüme bölmek (maliyet fonksiyonu tarafından belirlenen) ile aşırı karmaşıklık veya fazla uyumu önlemeye yönelik ceza terimi arasında bir denge kurarak birden fazla değişim noktasını etkili bir şekilde bulmayı amaçlar.

Literatürde birden fazla değişim noktasını belirleme konusunda en yaygın yöntem, bir parça için bir maliyet fonksiyonunu (genellikle negatif log olasılık gibi) minimize etmek ve aşırı uyumlanmayı engellemek için bir ceza terimi (c 'nin birden fazla değişim noktası versiyonu olan $\beta f(m)$) kullanmaktır. Bu, aynı zamanda benimsediğimiz ve eşlik eden pakette kullandığımız yaklaşımdır. Bu minimize işlemini gerçekleştirmek için bir kaba kuvvet yöntemi, $2n - 1$ çözümü düşünerek, m bilindiğinde $n - 1$ m'ye indirgenir.

1.2.1 İkili Bölümleme Algoritması

İkili bölümleme algoritması (BinSeg), değişim noktası literatüründe kullanılan en köklü arama yöntemidir. İkili bölümleme arama algoritmasının erken uygulamaları arasında (Scott ve Knott, 1974; Sen ve Srivastava, 1975) bulunmaktadır.

İkili bölümleme, herhangi bir tek değişim noktası yöntemini ardışık olarak farklı verilerde tekrarlayarak çoklu değişim noktalarına genişletmek için kullanılabilir. İlk olarak, tek bir değişim noktası test istatistiğini tüm veride uygular. y_1, y_2, \dots, y_n şeklindeki veriler üzerinde bir başlangıç noktası belirlenir. Bu başlangıç noktası, verinin ortalaması, medyanı veya başka bir özelliği olabilir. Belirlenen başlangıç noktasında bir değişim noktası testi yapılır. Bu test, verilerin iki alt küme olarak bölündüğünde, oluşan alt kümelerin toplam maliyetinin belirli bir kritere göre düşük olup olmadığını kontrol eder, yani bir τ 'nin aşağıdaki koşulu sağlayıp sağlamadığını test eder:

$$C(y_{1:\tau}) + C(y_{(\tau+1):n}) + \beta < C(y_{1:n})$$

Burada:

- C : Bir parça için maliyet fonksiyonu
- $y_{1:\tau}$: Başlangıçtan değişim noktasına kadar olan veri
- $y_{(\tau+1):n}$: Değişim noktasından sona kadar olan veri
- β : Aşırı uyum karşısında koruma sağlayan ceza terimi

Eğer bu koşul sağlanmıyorsa, o zaman herhangi bir değişim noktası tespit edilememiştir ve algoritma durur. Aksi takdirde veri, belirlenen değişim noktasından önce ve sonra olmak üzere iki parçaya bölünür. Tek değişim noktası tespit yöntemi, değişiklikten önce ve sonra iki yeni parçaya da tekrarlanır. Her iki parçada da değişim noktaları belirlenirse, bunları yeni belirlenen değişim noktasında daha fazla parçalara böler ve her yeni parçaya değişim noktası tespit yöntemini uygular. Bu süreç, verinin herhangi bir bölümünde değişim noktası bulunamayana kadar devam eder (Killick, Fearnhead ve Eckley (2012)).

Çoklu değişim noktalarını belirlemek için yaygın olarak kullanılan bir yaklaşım, aşağıdaki ifadeyi minimize etmektir:

$$\sum_{i=1}^{m+1} C(y(\tau_{i-1} + 1) : \tau_i) + \beta f(m)$$

Burada, C bir parça için bir maliyet fonksiyonu ve $\beta f(m)$ aşırı uyum karşısında koruma sağlayan bir ceza terimidir.

İkili bölümlleme, herhangi bir değişim noktasının konumu önceden belirlenmiş değişim noktalarına bağlı olduğu için ($f(m) = m$ olarak) yukarıdaki denklemin yaklaşık bir minimize edilmesidir. Algoritmanın her adımı, bu denklemi azaltıyorsa ek bir değişim noktası eklemeye çalışır. İkili bölümlleme algoritmasının avantajı, n 'nin veri uzunluğu olduğu durumda $O(n)$ hesaplama maliyeti ile uygulanabilen hızlı bir algoritma olmasıdır. Ancak, C 'yi uygun bir şekilde seçmek zor olabilir ve farklı C seçimleri, değişim noktalarının sayısının tahmininde önemli farklara neden olabilir.

1.2.2 Prophet

Facebook tarafından 2017'de önerilen Prophet tahmin modeli (Taylor ve Letham, 2018), aynı anda birden fazla mevsimsellik dönemini modelleyebilir. Güçlü tahmin yeteneklerinin yanı sıra, Prophet aynı zamanda büyük aykırı değerlere ve trendlerdeki kaymalara sahip günlük periyodik verileri işlemede de başarılıdır. (Zhao, Liu, Vanos ve Cao, 2018).

Prophet algoritması, uzun dönem zaman serilerinde belirli bir konumda değişen eğilimi bulmak için kullanılır. Değişim noktalarının tespit edilmesinde iyi tahmin edilen sonuçlar -gerçek değişim noktasına yakın veya eşit olunan durum - potansiyel değişim noktaları olarak belirlenir. Bu durumu takiben zaman boyutu kullanılarak potansiyel değişim noktaları tespit edilir ve gerçek değişim noktası bulunur.

Prophet algoritması, zaman serisi hakkında alan bilgisine sahip analistler tarafından sezgisel olarak ayarlanabilen, yorumlanabilir parametrelere sahip modüler bir regresyon modelidir. Her bir zaman serisini trend, mevsimsellik ve tatiller olmak üzere üç ana bileşene ayırır, aşağıdaki denklemde gösterildiği gibi:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

Burada, $g(t)$ zaman serisinin değerindeki periyodik olmayan değişimleri modelleyen trend fonksiyonunu, $s(t)$ periyodik değişimleri (örneğin, haftalık ve yıllık mevsimsellik) ve $h(t)$ potansiyel olarak düzensiz zamanlarda gerçekleşen tatillerin etkilerini temsil

eder. Hata terimi ϵ_t , model tarafından karşılanmayan herhangi bir özgün değişimi temsil eder ve normal olarak dağıldığı varsayılır.

Prophet algoritması kullanıldığında, potansiyel değişim noktaları otomatik olarak tespit edilebilir. Prophet algoritması çok küçük değişiklikleri tespit edebilir. Zaman boyutundaki potansiyel değişim noktalarını ayırt eder ve gerçek değişim noktasıyla, potansiyel değişim arasında minimum değişim oranına sahip olan noktaları bulur.

1.2.3 PELT

PELT (Pruned Exact Linear Time) algoritması (Killick ve diğerleri, 2012) tarafından geliştirilmiştir. Bu algoritma, zaman serisindeki değişim noktalarını tespit etmektedir ve bu değişim noktalarını tespit ederken belirli bir maliyet fonksiyonunu minimize etmeyi amaçlar.

Değişim noktası analizi, genel anlamda, bir veri içerisindeki istatistiksel özelliklerin değiştiği noktaların belirlenmesidir. Daha resmi olarak, $y_{1:n} = (y_1, \dots, y_n)$ şeklinde sıralı bir veri dizisine sahiptir.

Model, konumlarıyla birlikte bir dizi değişim noktasına (m) sahip olacaktır: $\tau_{1:m} = (t_1, \dots, t_m)$. Her değişim noktası konumu 1 ile $n - 1$ dahil arasında bir tam sayıdır.

$$\tau_0 = 0$$

ve $\tau_{m+1} = n$ 'yi tanımlıyor ve değişim noktalarının $\tau_i < \tau_j$ ancak ve ancak $i < j$ olacak şekilde sıralandığı varsayılmaktadır. Sonuç olarak m değişim noktası, verileri $m + 1$ parçaya bölecek ve i 'inci parça $y(\tau_{i-1} + 1) : \tau_i$ 'yi içerecektir.

$$m + 1 \sum_{i=1} [C(y(\tau_{i-1} + 1) : \tau_i)] + \beta f(m).$$

Burada C , bir parça için bir maliyet fonksiyonu ve $\beta f(m)$, aşırı uyuma (overfitting) karşı koruma amaçlı bir ceza terimidir.

Başlangıçta, tüm zaman noktaları aday değişim noktaları olarak seçilir. Genellikle, zaman serisinin başlangıcı ve sonu da bu aday değişim noktaları arasına dahil edilir. Aday değişim noktaları sırayla ele alınır. Her bir aday değişim noktası için bir budama koşulu kontrol edilir. Bu koşul, değişim noktasının gelecekteki bir zamanda optimal bir değişim noktası olma olasılığını değerlendirir. Koşulu sağlamayan aday değişim noktaları budanır, yani gelecekteki değerlendirmelerden çıkarılır.

Budamanın özü, Algoritma 1'in her yinelemede gerçekleştirilen minimizasyondan asla minimum olamayacak olan t değerlerini çıkarmaktır.

Aşağıdaki teorem bu tür budamanın yapılabileceği basit bir koşulu verir.

Teorem 1: Bir gözlem dizisine bir değişim noktası tanıtıldığında, dizinin maliyeti C azalmaktadır. Daha formal bir ifadeyle, $t < s < T$ için tüm t değerleri için bir K sabiti var kabul edilmektedir.

Optimal Bölütleme

Girdi: $y_i \in \mathbb{R}$ olmak üzere (y_1, y_2, \dots, y_n) formundaki bir veri kümesi. Uyumun ölçüsü olarak $C(\cdot)$ ve değişim noktalarının sayısına veya konumuna bağlı olmayan bir ceza

sabiti β kullanılır.

$n =$ verinin uzunluğu olsun ve $F(0) = -\beta$, $cp(0) = Null$ olarak ayarlanır.

Yineleme için $T^* = 1, \dots, n$

1. $F(T^*) = \min_{0 \leq s < T^*} [F(s) + C(y_{s+1} : T^*) + \beta]$.
2. $\tau' = \arg \min_{0 \leq s < T^*} [F(s) + C(y_{s+1} : T^*) + \beta]$ olsun.
3. $cp(T^*) = (cp(\tau'), T^*)$ olarak ayarlanır.

$cp(n)$ 'de kaydedilen değişim noktalarını çıktı olarak verir.

Aşağıdaki teorem bu tür budamanın yapılabileceği basit bir koşulu verir.

Teorem 1: Bir gözlem dizisine bir değişim noktası tanıtıldığında, dizinin maliyeti \mathcal{C} azalmaktadır. Daha formal bir ifadeyle, $t < s < T$ için tüm t değerleri için bir K sabiti var kabul edilmektedir.

$$\mathcal{C}(y_{1:t}) + \mathcal{C}(y_{t+1:T}) + K \leq \mathcal{C}(y_{1:T}).$$

Eğer

$$F(t) + \mathcal{C}(y_{t+1:a}) + K \geq F(s)$$

PELT Yöntemi

$y_i \in \mathbb{R}$ olmak üzere (y_1, y_2, \dots, y_n) formundaki bir veri kümesi. Uyumun ölçüsü olarak $C(\cdot)$ ve değişim noktalarının sayısına veya konumuna bağlı olmayan bir ceza sabiti β kullanılır.

$n =$ verinin uzunluğu olsun ve $F(0) = -\beta$, $cp(0) = Null$, $R_1 = \{0\}$ olarak ayarlanır.

Yineleme için $T^* = 1, \dots, n$

1. $F(T^*) = \min_{T \in R_{T^*-1}} [F(T) + C(y_{T+1} : T^*) + \beta]$.
2. $\tau^+ = \arg \min_{T \in R_{T^*-1}} [F(T) + C(y_{T+1} : T^*) + \beta]$ olsun.
3. $cp(T^*) = [cp(\tau^+), T^+]$ olarak ayarlanır.
4. $R_{T^*+1} = \{T \in R_{T^*-1} - \{T^*\} : F(T) + C(y_{T+1} : T^*) + K \leq F(T^*)\}$ olarak ayarlanır.

$cp(n)$ 'de kaydedilen değişim noktalarını çıktı olarak verir.

PELT yönteminin teorik hesaplama maliyeti incelendiğinde, değişim noktası modelleri ve cezaların en önemli sınıfına odaklanılır ve yöntemin veri noktalarının sayısına göre doğrusal bir hesaplama maliyetine sahip olması için yeterli koşullar sağlanır.

Odaklanılan durum, bölüm parametrelerinin bölümler arasında bağımsız olduğu ve bir bölüm için maliyet fonksiyonunun o bölümdeki verinin maksimum log-olabilirlik değerinin eksi olarak alındığı modeller kümesidir.

Daha formal olarak, sonuçlar yöntemin beklenen hesaplama maliyeti ile ilgilidir ve bu maliyetin analiz edilen veri noktalarının sayısına nasıl bağlı olduğunu ele almaktadır.

Bu amaçla, veri oluşturma süreci için temel bir stokastik model tanımlanır. Özellikle, bu süreci pozitif tamsayı zaman noktaları üzerine tanımlanır ve ardından bu süreç tarafından üretilen ilk n veri noktasının analizini ele alınır.

Elde edilen sonuç, belirli bir bölümle ilişkilendirilmiş parametrelerin $\pi(\theta)$ yoğunluk fonksiyonuna sahip olduğunu varsayar. Gösterim basitliği açısından, bir bölümün parametresi θ verildiğinde, bölüm içindeki veri noktalarının $\pi(\theta)$ yoğunluk fonksiyonuna sahip olduğunu varsayılır, ancak bir bölüm içindeki bağımlılıklara genişletmeler basittir. Son olarak, önceki belirtildiği gibi maliyet fonksiyonu, maksimum log-olabilirliğin eksi değerine dayanacaktır.

$$C(y_{t+1:T}) = - \max_{\theta} \sum_{i=t+1}^T \log f(y_i|\theta).$$

(5)

Bu kayıp fonksiyonu için (4) içinde $K = 0$ dikkate alınmalıdır. Bu nedenle, PELT'teki budama, sadece ceza sabiti β 'nın seçimine bağlı olacaktır.

- Ayrıca, her bir bölümün uzunluğu için bir model olarak bir değişim noktasının konumu için stokastik bir model gereklidir. Değişim noktası pozisyonları τ_1, τ_2, \dots ise, bölüm uzunluklarını $S_1 = \tau_1$ olarak tanımlanırsa ve $i = 2, 3, \dots$ için $S_i = \tau_i - \tau_{i-1}$.
- S_i , bağımsız ve aynı dağılıma(iid) sahip rastgele değişken S 'in bağımsız ve aynı dağılıma sahip kopyaları olarak kabul edilir. Ayrıca, S_1, S_2, \dots , bölümlerle ilişkilendirilmiş parametrelerle bağımsızdır.

Teorem 3.2: Beklenen log-olabilirliği maksimize eden değeri tanımlayalım:

$$\theta^* = \arg \max \int \int f(y|\theta) f(y|\theta_0) dy \pi(\theta_0) d\theta_0.$$

θ_i gerçek parametre olsun ve bu parametreyi içeren bölümle ilişkilendirilsin. $\hat{\theta}_n$, verilen veri $y_{1:n}$ ve tek bir bölüm varsayımına dayanarak θ için en büyük olabilirlik tahminidir:

$$\hat{\theta}_n = \arg \max_{\theta} \sum_{i=1}^n \log f(y_i|\theta).$$

$B_n = \sum_{i=1}^n |\log f(y_i|\theta_n) - \log f(y_i|\theta^*)|$ ifadesine sahip olarak, $E(B_n) = o(n)$ ve $E[(B_n - E(B_n))^4] = O(n^2)$ olmaktadır.

(A2)

$$\mathbb{E} \left(\log(f(Y|\theta)) - \log(f(Y|\theta')) \right) < \infty;$$

(A3)

$$E(S^4) < \infty; \text{ ve}$$

(A4)

$$E(\log f(Y|A) - \log f(Y|O')) > \frac{\beta}{E(S')}$$

Burada, S beklenen bölüm uzunluğudur, n veri noktasını analiz etmek için PELT'in beklenen CPU maliyeti L_n için bazı sabit $L < \infty$ ile yukarıdan sınırlıdır.

Θ_i , y'_i yi içeren bölümle ilişkilendirilmiş gerçek parametre olsun ve $\hat{\Theta}_n$, veri $y_{1:n}$ ve tek bir bölüm varsayımıyla verilen Θ için maksimum olabilirlik tahminidir (q_i).

S 'nin beklenen bölüm uzunluğu olduğu durumda, n veri noktasını analiz etmek için PELT'in beklenen CPU maliyeti, L sabiti için yukarıda L_n ile sınırlanır.

Teorem 3.2'nin Şartları (A1) ve (A2), zayıf teknik şartlardır. Örneğin, maksimum olabilirlik tahminleri içi n genel asimptotik sonuçlar, $B_n = Op(1)$ verecektir, ve (A1), B_n 'nin $O(n^{1/2})$ veya daha büyük değerleri alma olasılığını kontrol eden biraz daha güçlü bir koşuldur.

Diğer iki koşul daha önemlidir. Koşul (A3), büyük bölümlerin olasılığını kontrol etmek için gereklidir. (A3)'ün önemli bir sonucu, değişim noktalarının beklenen sayısının n ile lineer olarak artacak olmasıdır. Son olarak, (A4) koşulu doğaldır çünkü gerçek değişim noktası ve parametre değerleri ile elde edilen beklenen cezalandırılmış olabilirlik değerinin, veriye tek bir bölüm uydurulduğunda ve bölüm parametresi Θ^* ile elde edildiğinde beklenen cezalandırılmış olabilirlik değerinden daha büyük olması gerekir. Tüm durumlarda, algoritmanın en kötü durum karmaşıklığı, budama yapılmadığı durumdur ve hesaplama maliyeti $O(n^2)$ olacaktır.

1.2.4 Parçalı Regresyon

Parçalı veya kesik çizgili modeller, yanıt ile bir veya daha fazla açıklayıcı değişken arasındaki ilişkilerin parçalı doğrusal olduğu, yani iki veya daha fazla değişkenle temsil edildiği regresyon modelleridir. Bu ilişkiler, genellikle bilinmeyen değerlerde birleştirilen iki veya daha fazla düz çizgi tarafından temsil edilir, bu değerlere genellikle kırılma noktaları, değişim noktaları veya birleşim noktaları denir. Bu yöntemde bağımsız değişken, aralıklara bölünür ve her bir aralığa ayrı bir çizgi bölümü uyarlanır. Parçalı regresyon analizi ayrıca çeşitli bağımsız değişkenlerle yapılan çok değişkenli veriler üzerinde de gerçekleştirilebilir. Parçalı regresyon analizi, bağımsız değişkenlerin belirli gruplara ayrıldığı durumlarda, bu gruplardaki değişkenler arasındaki ilişkilerin farklı olduğuna inanıldığında kullanışlıdır. Bu parçalar arasındaki sınırlar, değişim noktaları olarak adlandırılır.

Matematiksel olarak, model şu şekilde ifade edilir:

$$y = \beta_{0i} + \beta_{1i}x + \epsilon$$

Bu denklemde, β_{0i} ve β_{1i} sırasıyla i -inci bölümün kesişim noktası ve eğimini temsil eder.

Parçalı regresyon, ekonomi, biyoloji, çevre bilimleri, epidemiyoloji gibi çeşitli alanlarda kullanılır. Kalite iyileştirme müdahaleleriyle ilgili çalışmalarda parçalı regresyon analizlerinin kullanımına dair birçok örnek yayınlanmıştır.

Temel bir parçalı regresyon analizinde zaman periyodu müdahale öncesi ve sonrası parçalara bölünür ve her parçada ayrı ayrı kesişim noktaları ve eğimler tahmin edilir. Müdahaleden önce ve sonra kesişmelerde ve eğimlerdeki değişiklikleri test

etmek için istatistiksel testler gerçekleştirilir. Veriler ve model spesifikasyonunda bazı basit değişikliklerle parçalı regresyon analizi, genellikle standart istatistiksel yazılım paketlerinde kolayca uygulanabilir. Genellikle, zaman içinde alınan gözlemler ilişkilidir, bu nedenle otokorelasyonu hesaba katmak için genellikle ek bir düzeltme yapılması gerekir.

Parçalı doğrusal regresyon, parçalı regresyonun doğrusal regresyon kullanılarak elde edilen bir alt türüdür. İki parçalı doğrusal regresyon, bir değişim noktası ile ayrılmış iki parçayla, değişken bir etkenin (x) yanıt fonksiyonunun (Y_r) ani bir değişikliğini nicelendirmek için kullanışlı olabilir. Değişim noktası, yanıt fonksiyonunun kritik, güvenli veya eşik değeri olarak yorumlanabilir; bu değerin ötesinde veya altında (istenmeyen) etkiler meydana gelebilir. Değişim noktası, karar verme süreçlerinde önemlidir.

Her bir parça için ayrı ayrı uygulanan en küçük kareler yöntemi, iki regresyon çizgisini veriye mümkün olduğunca iyi uyacak şekilde oluştururken, gözlemlenen (y) ve hesaplanan (Y_r) bağımlı değişken değerleri arasındaki farkın karesini en aza indirir. Bu yöntemle şu iki denklem elde edilir:

$$Y_r = A1.x + K1 < BP \text{ (değişim noktası için)}$$

$$Y_r = A2.x + K2 > BP \text{ (değişim noktası için)}$$

Burada:

- Y_r , x 'in belirli bir değeri için beklenen (tahmin edilen) y değeridir;
- $A1$ ve $A2$ regresyon katsayılarıdır (çizgi bölümünün eğimini gösterir);
- $K1$ ve $K2$ regresyon sabitleridir (y -ekseninde kesişimi gösterir).

Bu yöntem aynı zamanda iki korelasyon katsayısı (R) da üretir:

$$R_1^2 = 1 - \frac{\sum (y - Y_r)^2}{\sum (y - Y_{a1})^2} \text{ } x < BP \text{ (değişim noktası için)}$$

ve

$$R_2^2 = 1 - \frac{\sum (y - Y_r)^2}{\sum (y - Y_{a2})^2} \text{ } x > BP \text{ (değişim noktası için)}$$

Burada:

- $\sum (y - Y_r)^2$: Her bir parça için minimize edilmiş SSD'yi temsil eder.
- Y_{a1} ve Y_{a2} : İlgili parçalarda y 'nin ortalamasıdır.

En uygun eğilimi belirlemede, bu eğilimin güvenilir (anlamli) olduğundan emin olmak için istatistiksel testler gerçekleştirilmelidir.

Eğer anlamlı bir değişim noktası tespit edilemezse, değişim noktası olmadan bir regresyona geçilmelidir. Aşağıdaki istatistiksel testler, eğilim türünü belirlemek için kullanılır:

- **A1 ve A2'nin Anlamlılığı:** $A1$ ve $A2$ 'nin anlamlılığı, $A1$ ve $A2$ 'nin standart hata SE 'si ve Student'ün t -distribution'ı kullanılarak belirlenir.
- **A1 ve A2'nin Farkının Anlamlılığı:** $A1$ ve $A2$ 'nin farkının anlamlılığı, farklarının standart hatası SE ve Student'ün t -distribution'ı kullanılarak belirlenir.

- **Y1 ve Y2'nin Farkının Anlamlılığı:** Y1 ve Y2'nin farkının anlamlılığı, farklarının standart hatası SE ve Student'ün t-distribution'ı kullanılarak belirlenir.

- **Değişim Noktasının Varlığını Test Etme:** Sözde skor testi, parçalı çizginin tahminini gerektirmez ve kırılma noktasının varlığını test etmek için daha formal bir istatistiksel yaklaşımdır.

Ayrıca, tüm verilerin korelasyon katsayısı (R_a), belirleme katsayısı veya açıklama katsayısı, regresyon fonksiyonlarının güven aralıkları ve ANOVA analizi kullanılmaktadır.

C_d katsayısı, tüm veri seti için belirlenen şartlar altında maksimize edilmesi gereken bir belirleme katsayısıdır ve şu formülle hesaplanır:

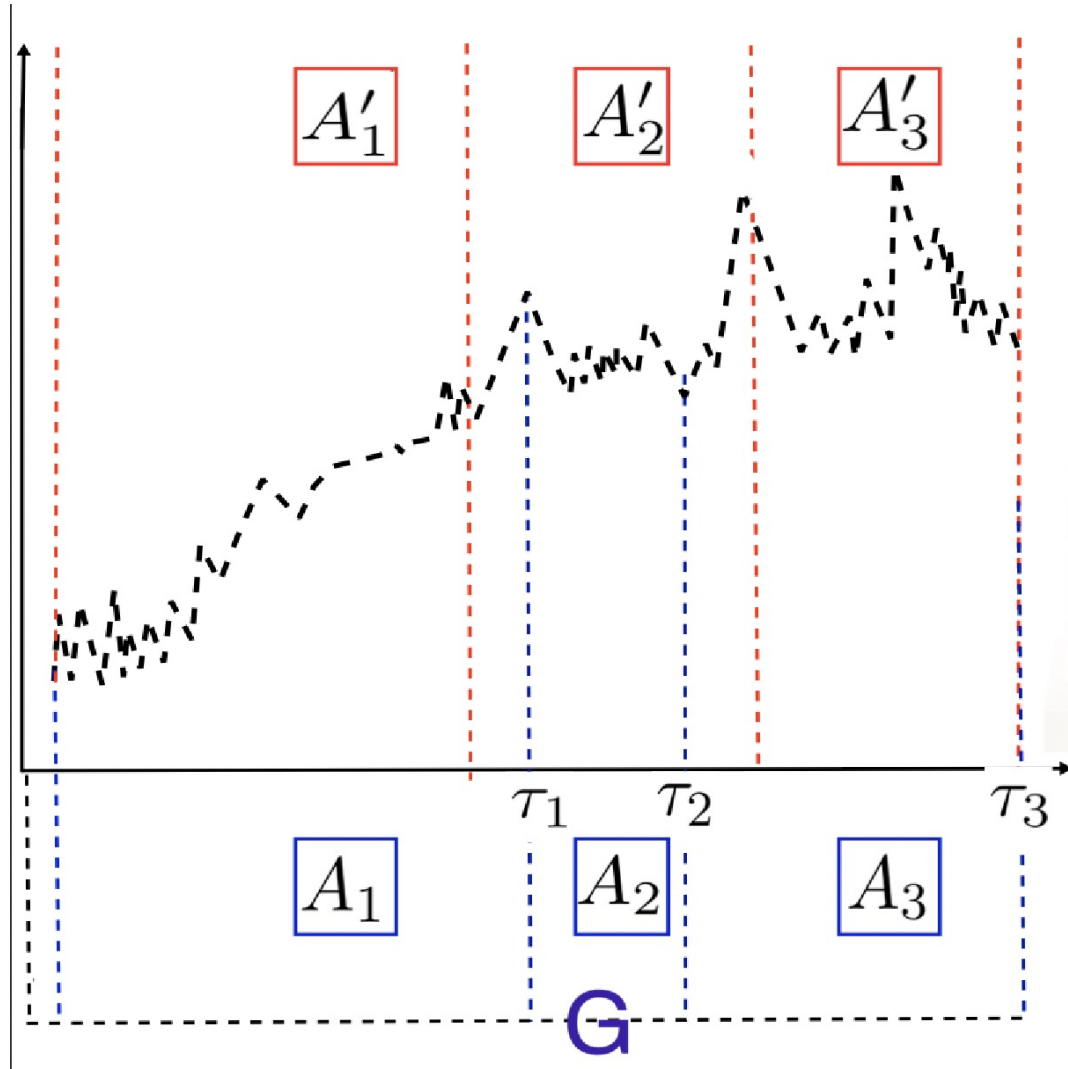
$$C_d = 1 - \frac{\sum (y - Y_r)^2}{\sum (y - Y_a)^2}$$

Burada Y_r , önceki regresyon denklemlerine göre beklenen (tahmin edilen) y değeridir ve Y_a , tüm y değerlerinin ortalamasıdır.

C_d katsayısı, 0 ile 1 arasında değer alır. Saf, parçalanmamış doğrusal regresyonda, C_d ve R_a^2 değerleri eşittir. Parçalı regresyonda, C_d 'nin R_a^2 'den anlamlı derecede büyük olması, parçalanmanın haklı olduğunu gösterir.

Değişim noktasının en uygun değeri, C_d katsayısının maksimum olduğu noktada bulunabilir.

1.3 Kapsama Ölçütü



Şekil 1.1: Kapsama Ölçütü

Değişim noktası tespiti için metrikler, kümeleme metrikleri veya sınıflandırma metrikleri uyarlanarak kullanılabilir. Bu bölümde kümeleme analizinden uyarlanarak hesaplanan kapsama ölçütü (cover metric) tanıtılmıştır.

Kapsama ölçütü, gerçekteki değişim noktası ile bulunan değişim noktasının ne kadar iyi örtüştüğünü ölçer. Jaccard benzerliğini kullanarak örtüşmeyi hesaplar (Van den Burg ve Williams (2020)).

K işaretleyici sayısı ($k = 1, \dots, K$) olmak üzere, k . işaretleyici tarafından belirlenen değişim noktalarının konumları $T_k = \{\tau_1, \dots, \tau_{n_k}\}$ ile gösterilir. $\tau_i < \tau_j$ koşulunu sağlamak için $i < j$ olmalıdır. T_k , $[1, T]$ aralığını A_j adı verilen ayrık kümeler halinde bölen bir bölüm G_k 'yi belirtir. Burada A_j , $j = 1, \dots, n_k + 1$ için $\tau_j - 1$ 'den τ_j - 1'e olan parçadır. Notasyon kolaylığı için $\tau_0 = 1$ ve $\tau_{n_k+1} = T + 1$ kullanılır.

İki küme için $A, A' \subseteq [1, T]$ Jaccard indeksi, ayrıca iki kümenin kesişimi bölü birleşimi olarak da bilinir.

$$J(A, A') = \frac{|A \cap A'|}{|A \cup A'|}$$

Kapsama metriğini G ve G' olarak tanımlarız.

$$C(G, G') = \frac{1}{T} \sum_{A \in G} |A| \cdot \max_{A' \in G'} J(A, A')$$

İşaretleyiciler tarafından sağlanan temel doğruluk bölümlerinin $\{G_k\}_{k=1}^K$, bir algoritma tarafından verilen bir S bölümü için, $C(G_k, S)$ ortalamasını hesaplayarak işaretleyicilerin tek bir performansını değerlendirilir.

1.4 F1 Puanı

Değişim noktası tespiti algoritmalarını değerlendirmeye ilişkin alternatif bir görüş, değişim noktası tespitini bir sınıflandırma problemi olarak kabul eder. Burada her bir gözlem “değişim noktası” veya “tipik gözlem” sınıflarına ait olarak değerlendirilir (Aminikhanghahi ve Cook, 2017; Killick ve diğerleri, 2012). X , bir tespit algoritması ile bulunan değişim noktası konumları olsun. $T^* = \cup T_k$, tüm işaretleyicilerin birleşik kümesidir. Değişim noktalarının gerçek konumları T için, algoritmalar ile elde edilen gerçek pozitifler $TP(T, X)$ olarak tanımlanır ($\tau \in T$). Bu küme içindeki her bir τ , $|\tau - x| \leq M$ koşulunu sağlayan en az bir $x \in X$ için bulunur. Ancak bir $\tau \in T$ için yalnızca bir $x \in X$ kullanılır. Bu çalışmada $M = 5$ alınmıştır.

Doğruluk değeri, 0 ile 1 arasındadır. Doğru tahmin edilen değerlerin, tüm değerlere bölünmesi ile elde edilmektedir. Doğruluk değeri 1'e yaklaştıkça başarı artmaktadır (Kelle ve Hüseyin, 2022).

$$Dogruluk = \frac{(GP + GN)}{(GP + GN + FP + FN)}$$

Değişim noktalarının sayısı genellikle serideki gözlem sayısından çok küçüktür. Bu nedenle doğruluk oranı gibi ölçütler oldukça çarpık olacaktır. Bu durumda F puanı kullanılabilir. F puanında, bir algoritmanın performansı kesinlik (P) (doğru tespit edilen değişim noktalarının tespit edilen değişim noktalarının sayısına oranı) ve duyarlılık (doğru tespit edilen değişim noktalarının gerçek değişim noktalarının sayısına oranı) kullanılarak ölçülür ($\beta = 1$ olarak varsayılır).

$$F_\beta = \frac{(1 + \beta^2) \cdot P \cdot R}{(\beta^2 \cdot P) + R}$$

GERÇEK	TAHMİN	
	Gerçek Pozitif (GP)	Yanlış Negatif (YN)
	Yanlış Pozitif (YP)	Gerçek Negatif (GN)

Şekil 1.2: Yanılgı matrisi

- Gerçek Pozitif (GP): Doğru olarak tahmin edilen, gerçekte de doğru olan değerler,
- Gerçek Negatif (GN): Yanlış olarak tahmin edilen, gerçekte de yanlış olan değerler,
- Yanlış Pozitif (YP): Doğru olarak tahmin edilen, gerçekte yanlış olan değerler,
- Yanlış Negatif (YN): Yanlış olarak tahmin edilen, gerçekte doğru olan değerleri ifade etmektedir.

Duyarlılık (recall), doğru olarak tahmin etmemiz gereken değerlerin, ne kadarını doğru tahmin ettiğimizi belirtmektedir. Duyarlılık değeri, gerçekte doğru olan ve doğru olarak tahmin edilen değerlerin, tüm doğru değerlere bölünmesi ile elde edilmektedir(Kelle ve Hüseyin, 2022).

$$Duyarlılık = \frac{GP}{(GP + YN)}$$

Kesinlik (precision), doğru olarak tahmin ettiğimiz değerlerin, ne kadarının gerçekte doğru olduğunu göstermektedir. Kesinlik değeri, gerçekte doğru olan ve doğru olarak tahmin edilen değerlerin, doğru olarak tahmin edilen tüm değerlere bölünmesi ile elde edilmektedir(Kelle ve Hüseyin, 2022).

$$Kesinlik = \frac{GP}{(GP + YP)}$$

F1 puanı, duyarlılık ve kesinlik değerlerinin harmonik ortalamasının hesaplanması ile elde edilmektedir. Her iki değerinde hesaplamaya katılarak dengeli bir değer elde edilmesi amaçlanmaktadır. Dengesiz verilerde başarılı sonuçlar elde etmek için kullanılmaktadır(Kelle ve Hüseyin, 2022).

$$F1-Puanı = \frac{2 * Kesinlik * Duyarlılık}{(Kesinlik + Duyarlılık)}$$

Bölüm 2

Uygulama

Uygulamada yirmi yapay veri, on bir gerçek veri ve on beş kademeli eşik egzersiz test verisi kullanılmıştır. Yapay verilerin her biri farklı sayıda değişim noktasına sahip olacak şekilde üretilmiştir. Gerçek veriler ise farklı alanlardan alınmıştır. Verilerin performanslarını belirlemek amacıyla F1 puanı ve kapsama ölçütü kullanılmıştır. Çalışmada R ve Python programlama dilleri kullanılmıştır. Uygulamada kullanılan veriler ve kodlar github sayfasında (<https://github.com/eyildiztepe/ChangePointDetection>) paylaşılmıştır.

2.1 Veri

Bu bölümde üretilen yapay veriler ve uygulamada kullanılan gerçek veriler hakkında bilgi verilmiştir.

2.1.1 Yapay Veri

Çalışmada kullanılan yapay veriler farklı sayıda değişim noktasına (DN) sahip olacak şekilde R programlama dili kullanılarak üretilmiştir. Yapay verilerin örneklem genişlikleri, ortalaması 2000, standart sapması 500 olan Normal dağılımından, verideki değişim noktası sayısı ortalaması 2.8 olan Poisson dağılımından ve DN'lerin konumları ise Uniform dağılımından üretilmiştir. Kullanılan verilerin özellikleri Tablo 2.1'de verilmiştir.

Tablo 2.1: Yapay Veriler

Veri	Gözlem Sayısı	DN Sayısı	DN Konumu
1	1687	4	516, 578, 779, 1499
2	2092	3	564, 1003, 1347
3	1582	4	175, 553, 1186, 1347
4	2798	3	951, 985, 2315
5	2165	3	1034, 1835, 1892
6	1590	4	631, 698, 1208, 1481

Veri	Gözlem Sayısı	DN Sayısı	DN Konumu
7	2244	1	1578
8	2369	3	788,958,1768
9	2288	4	316,493, 587,1606
10	1847	4	153,300, 469,1172
11	2756	3	2119,2168,2377
12	2195	5	909,1004,1317,1422,1749
13	1689	2	479,611
14	893	2	552,837
15	2562	1	575
16	1978	1	293
17	1992	2	955,1798
18	2472	3	1470,1786,2365
19	2411	3	393,874,1047
20	2297	2	79,1622

2.1.2 Gerçek Veri

Çalışmada farklı alanlardan gerçek veriler kullanılmıştır. Veriler (<https://github.com/alan-turing-institute/TCPD/tree/master/datasets>) adresinden temin edilmiştir. Gerçek verilerin özellikleri Tablo 2.2’de verilmiştir.

Çalışmada kullanılan gerçek veriler 5 ayrı işaretleyici tarafından işaretlenmiştir (Van den Burg ve Williams, 2020). Bu nedenle değişim noktası konumları değişiklik gösterebilmektedir.

Tablo 2.2: Gerçek Veriler

Veri	N	DN Sayısı(1)	DN Sayısı(2)	DN Sayısı(3)	DN Sayısı(4)	DN Sayısı(5)
Bitcoin	774	4	1	1	7	7
Brent-spot.	500	3	2	5	9	11
Children-per women	301	2	1	2	4	2
Co2- canada	215	2	6	2	5	7
Debt -Ireland	21	2	2	2	4	2
Rail-lines	37	2	2	2	2	1
Rather-stock	600	1	2	1	2	2
Scanline-42049	481	10	7	2	7	7
Shangai-license	205	1	1	1	1	2
Usd-isk	247	2	4	1	3	2
Well-log	675	11	9	9	2	17

2.2 Değişim Noktası Analizi

Bu bölümde kullanılan yöntemler ile elde edilen sonuçlara yer verilmiştir. Kullanılan fonksiyonların varsayılan ayarları ile elde edilen sonuçlar F1 puanı varsayılan ve kapsama ölçütü varsayılan olarak adlandırılmıştır. Ayrıca, argümanların değerleri değiştirilerek elde edilen en iyi sonuçlar F1 puanı ayarlanmış ve kapsama ölçütü ayarlanmış olarak sunulmuştur.

2.2.1 Yapay Veriler için Sonuçlar

Bu bölümde yapay veriler için elde edilen sonuçlara yer verilmiştir.

Tablo 2.3: F1 puanı (Varsayılan)

Veri	AMOC	BINSEG	SEGMENTED	PROPHET	PELT
1	0,5714	0,9091	0,2857	0,2857	0,4000
2	0,6667	0,4000	0,3333	0,1818	0,4444
3	0,5714	0,9091	0,1290	0,2000	0,4000
4	0,6667	0,8000	0,3333	0,2222	0,7500
5	0,6667	0,6000	0,4444	0,2500	0,6666
6	0,5714	0,5455	0,4000	0,1538	0,4000
7	0,5000	0,2500	0,5000	0,3333	0,5000
8	0,6667	0,8000	0,5000	0,2222	0,6000
9	0,2857	0,5455	0,4000	0,1818	0,3636
10	0,5714	0,7273	0,6000	0,1666	0,4615
11	0,6667	0,6000	0,5000	0,2222	0,4444
12	0,5000	0,8333	0,1666	0,1666	0,5000
13	0,8000	0,6667	0,6666	0,2857	0,5714
14	0,8000	0,6667	0,6666	0,2857	0,5714
15	0,9998	0,5000	0,5000	0,4000	0,4000
16	0,9999	0,5000	0,5000	0,4000	0,4000
17	0,8000	0,6667	0,6666	0,2857	0,5714
18	0,6667	0,8000	0,5000	0,2222	0,4000
19	0,6667	0,8000	0,5000	0,2222	0,4444
20	0,8000	0,4444	0,3333	0,2857	0,2857

Algoritmalar varsayılan parametre ayarlarıyla çalıştırıldığında (2.3), yirmi yapay verinin sekizinde BinSeg, altısında AMOC ve birinde PELT 0,7 ve üzerinde F1 puanına sahipken, parçalı regresyon ve Prophet algoritmasında F1 puanı 0,7 ve üzerinde olan veri bulunmamaktadır. Varsayılan parametrelerle çalıştırılan algoritmalarından Prophet on yedi, parçalı regresyon üç, AMOC, BinSeg ve Pelt algoritmalarında birer tane 0,3'ün altında F1 puanına sahip veri bulunmaktadır.

Tablo 2.4: Kapsama Ölçütü (Varsayılan)

Veri	AMOC	BINSEG	SEGMENTED	PROPHET	PELT
1	0,5975	0,9895	0,3740	0,3607	0,6843
2	0,5388	0,8881	0,5185	0,5600	0,7686
3	0,4937	0,9943	0,2397	0,5151	0,7953
4	0,5850	0,8989	0,4876	0,5495	0,8897
5	0,7707	0,9687	0,8126	0,6164	0,7706
6	0,6079	0,8822	0,7007	0,3625	0,5339

Veri	AMOC	BINSEG	SEGMENTED	PROPHET	PELT
7	0,9648	0,7995	0,7807	0,4670	0,8523
8	0,6084	0,8848	0,6725	0,6173	0,7047
9	0,6629	0,9467	0,4937	0,5129	0,6060
10	0,6269	0,9288	0,7368	0,4726	0,4510
11	0,8764	0,9464	0,8185	0,9991	0,6856
12	0,5639	0,8220	0,6032	0,6158	0,5461
13	0,8771	0,7478	0,8771	0,5563	0,4660
14	0,8932	0,9567	0,8911	0,4304	0,7085
15	0,9992	0,9863	0,8634	0,5091	0,4972
16	0,9950	0,7715	0,8828	0,4541	0,4058
17	0,8408	0,8620	0,8403	0,5809	0,5871
18	0,7744	0,9956	0,8502	0,4184	0,4871
19	0,6932	0,9925	0,7623	0,4422	0,4285
20	0,5905	0,7208	0,4578	0,4711	0,5644

Tablo 2.4'daki sonuçlara göre, BinSeg varsayılan parametre ayarlarıyla çalıştırıldığında tüm verilerde 0,7'nin üzerinde kapsama ölçütü değerine sahiptir. Parçalı regresyon on ikisinde, AMOC dokuzunda, Pelt yedisinde ve Prophet birinde 0,7 ve üzerinde kapsama ölçütü değerine sahip olduğu görülmektedir.

Varsayılan ayarlar çalıştırılan algoritmalarından sadece parçalı regresyonda bir veri 0,3'ün altında kapsama ölçütü değerine sahiptir.

Tablo 2.5: F1 Puanı (Ayarlanmış)

Veri	AMOC	BINSEG	SEGMENTED	PROPHET	PELT
1	0,5714	0,9997	0,4444	0,2500	0,6000
2	0,6667	0,5000	0,5000	0,4000	0,5000
3	0,5714	0,9994	0,4000	0,2500	0,4000
4	0,6667	0,9996	0,4000	0,2500	0,7500
5	0,6667	0,7500	0,4000	0,2857	0,6666
6	0,5714	0,6000	0,4000	0,3333	0,2857
7	0,5000	0,5000	0,5000	0,4000	0,5000
8	0,6667	0,7500	0,5000	0,2500	0,5454
9	0,2857	0,6000	0,4000	0,2222	0,3636
10	0,5714	0,8000	0,6000	0,2000	0,6000
11	0,6667	0,7500	0,5000	0,2500	0,5714
12	0,5000	0,8333	0,1666	0,3636	0,5000
13	0,8000	0,9987	0,6666	0,3333	0,6666
14	0,8000	0,6667	0,6666	0,3333	0,6666
15	0,9998	0,9989	0,5000	0,5000	0,5000
16	0,9999	0,9967	0,5000	0,5000	0,5000
17	0,8000	0,9985	0,6666	0,3333	0,6666
18	0,6667	0,9978	0,5000	0,2500	0,5000

Veri	AMOC	BINSEG	SEGMENTED	PROPHET	PELT
19	0,6667	0,9949	0,5000	0,2500	0,5000
20	0,8000	0,6667	0,3333	0,3333	0,3333

Ayarlanmış parametreler sonucunda en iyi F1 puanı değerleri BinSeg ile elde edilmiştir. Prophet ve parçalı regresyon algoritmalarında hiç bir veri 0,7 ve üzerinde F1 puanına sahip değildir. Pelt algoritmasında bir tane verinin 0,7'nin üzerinde olduğu görülmektedir.

Prophet algoritmasının en kötü F1 puanı değerlerini verdiği görülmektedir. AMOC, parçalı regresyon ve Pelt algoritmalarında 0,3 'ün altında birer tane veri bulunmaktadır.

Tablo 2.6: Kapsama Ölçütü (Ayarlanmış)

Veri	AMOC	BINSEG	SEGMENTED	PROPHET	PELT
1	0,5975	0,9930	0,9930	0,6104	0,7081
2	0,5388	0,8914	0,8914	0,6793	0,8194
3	0,4937	0,9950	0,9950	0,6152	0,9246
4	0,5850	0,9957	0,9957	0,6830	0,8840
5	0,7707	0,9688	0,9688	0,7494	0,7706
6	0,6079	0,8853	0,8853	0,3967	0,5599
7	0,9648	0,9648	0,9648	0,6002	0,9670
8	0,6084	0,8831	0,8831	0,6700	0,7364
9	0,6629	0,9467	0,9467	0,5665	0,6067
10	0,6269	0,9092	0,9092	0,4866	0,7201
11	0,8764	0,9689	0,9689	0,9998	0,9387
12	0,5639	0,8220	0,8220	0,6338	0,5604
13	0,8771	0,9976	0,9976	0,6158	0,6509
14	0,8932	0,9567	0,9567	0,5356	0,8589
15	0,9992	0,9992	0,9992	0,5255	0,8189
16	0,9950	0,9950	0,9950	0,5825	0,7296
17	0,8408	0,9980	0,9980	0,7171	0,7905
18	0,7744	0,9976	0,9976	0,4701	0,6171
19	0,6932	0,9983	0,9983	0,4816	0,6680
20	0,5905	0,9854	0,9854	0,5511	0,7184

Ayarlanan parametreler ile hesaplanan kapsama ölçütü değerinin BinSeg ve parçalı regresyon algoritmalarının tamamında 0,8'in üzerinde olduğu görülmektedir. Pelt için on dört, AMOC için dokuz ve Prophet algoritması için üç verinin 0,7'nin üzerinde kapsama ölçütü değerine sahip olduğu görülmektedir. Ayarlanan parametrelerle çalıştırılan beş farklı algoritmanın kullanıldığı yirmi verinin hiçbirinde 0,3'ün altından kapsama ölçütü değeri bulunmamaktadır.

Elde edilen sonuçlara göre yapay veriler için algoritmaların performansları arasında anlamlı bir fark olup olmadığı Friedman testi ile araştırılmıştır. Takip eden bölümlerde

PMCMRplus paketindeki `friedman.test()` ve `frdAllPairsNemenyiTest()` fonksiyonları ile elde edilen sonuçlara yer verilmiştir. Hipotez testlerinde ($\alpha = 0.05$) alınmıştır.

Loading required package: PMCMRplus

Friedman rank sum test

```
data: sim
Friedman chi-squared = 54.674, df = 4, p-value = 3.803e-11
```

H_0 : Algoritmaların performansları arasında istatistiksel olarak anlamlı bir fark yoktur.
 H_1 : En az bir algoritmanın performansı farklıdır.

p -değeri 0.05 (α)'den küçüktür. Bu nedenle H_0 hipotezi reddedilmiştir, en az bir algoritmanın performansının diğerlerinden farklı olduğu söylenebilir.

İkili karşılaştırmalar için Nemenyi testi (Nemenyi (1963); Demšar (2006)) kullanılmıştır.

Pairwise comparisons using Nemenyi-Wilcoxon-Wilcox all-pairs test for a two-way balance

```
data: y
```

	AMOC	BinSeg	Parçalı Reg	Prophet
BinSeg	0.00074	-	-	-
Parçalı Reg	0.00074	1.00000	-	-
Prophet	0.49732	2.8e-07	2.8e-07	-
PELT	0.99964	0.00166	0.00166	0.37348

P value adjustment method: single-step

Buna göre; AMOC ve BinSeg algoritmaları arasında; parçalı regresyon ve AMOC algoritmaları arasında, Prophet ile BinSeg ve parçalı regresyon algoritmaları arasında, PELT ile BinSeg ve parçalı regresyon algoritmaları arasında istatistiksel olarak anlamlı fark olduğu, sonucuna varılmıştır.

2.2.2 Gerçek Veriler için Sonuçlar

Bu bölümde gerçek veriler için elde edilen sonuçlara yer verilmiştir.

Tablo 2.7: F1 puanı (Varsayılan)

Veri	AMOC	BINSEG	PELT	SEGMENTED	PROPHET
Bitcoin	0,3670	0,4897	0,2663	0,3670	0,2341
Brent-spot	0,2718	0,6431	0,4244	0,6590	0,3904
Children-per women	0,6175	0,5902	0,3366	0,8440	0,5168
Co2- canada	0,5441	0,8194	0,8194	0,3938	0,6315
Debt -Ireland	0,7603	1,0000	0,7603	0,7603	0,6086
Rail-lines	0,8462	0,8000	0,4690	1,0000	0,2666
Rather-stock	0,2718	0,3392	0,4710	0,4886	0,5292
Scanline-42049	0,4926	0,7400	0,5151	0,2463	0,3902
Shangai-license	0,8679	0,6511	0,6666	0,6495	0,5316
Usd-isk	0,7854	0,6093	NA	0,7881	0,5956
Well-log	NA	0,7289	0,4235	0,6912	0,3589

Varsayılan parametre ayarlarıyla algoritmalar çalıştırıldığında, on bir verinin beşinde BinSeg, dördünde AMOC ve parçalı regresyon, ikisinde PELT 0,7 ve üzerinde F1 puanına sahipken Prophet algoritmasında F1 puanı 0,7 ve üzerinde olan veri bulunmamaktadır.

Varsayılan ayarlarda, on bir verinin ikisinde AMOC ve Prophet ile, birinde PELT ve parçalı regresyon ile 0,3'ün altında F1 puanı elde edilmiştir.

BinSeg algoritmasında F1 puanı 0,3 ve altında olan veri bulunmamaktadır. Değişim noktası tespit edilemeyen durumlar (NA) ile gösterilmiştir.

Tablo 2.8: Kapsama Ölçütü (Varsayılan)

Veri	AMOC	BINSEG	PELT	SEGMENTED	PROPHET
Bitcoin	0,7640	0,7354	0,3022	0,5322	0,1941
Brent-spot	0,4251	0,5921	0,4535	0,4945	0,2992
Children-per women	0,7838	0,7663	0,7721	0,6282	0,2753
Co2- canada	0,5264	0,7291	0,7393	0,5135	0,3409
Debt -Ireland	0,5844	0,6607	0,5446	0,5861	0,4000
Rail-lines	0,7682	0,7732	0,4408	0,7890	0,3081
Rather-stock	0,3870	0,3923	0,3970	0,5164	0,3470
Scanline-42049	0,4305	0,7502	0,4157	0,3859	0,3249
Shangai-license	0,9105	0,7691	0,3132	0,8272	0,2458
Usd-isk	0,8577	NA	NA	0,5248	0,2752
Well-log	0,4527	0,7696	0,4285	0,4146	0,3743

Varsayılan ayarlar ile, on bir verinin yedisinde BinSeg, beşinde AMOC, ikisinde PELT ve parçalı regresyon ile 0,7 ve üzerinde kapsama ölçütü değerleri elde edilmiştir.

Prophet algoritmasında hiçbir veri için kapsama ölçütü değeri 0,7 nin üzerinde değildir. Varsayılan ayarlar ile Prophet algoritmasında başarılı kapsama ölçütü değerleri elde edilemediği görülmektedir. Prophet on bir verinin beşinde 0,3 ve altında

kapsama ölçütü değerine sahiptir. Değişim noktası tespit edilemeyen durumlar (NA) ile gösterilmiştir.

Tablo 2.9: F1 Puanı (Ayarlanmış)

Veri	AMOC	BINSEG	PELT	SEGMENTED	PROPHET
Bitcoin	0,3670	0,6124	0,3940	0,4523	0,2743
Brent-spot	0,2718	0,6341	0,4481	0,6590	0,3704
Children-per women	0,6175	0,5902	0,6388	0,8440	0,5168
Co2- canada	0,5441	0,8776	0,3595	0,7142	0,6045
Debt -Ireland	0,7603	0,9583	0,9583	0,9795	0,9795
Rail-lines	0,8461	0,8316	0,7234	0,9655	0,9655
Rather-stock	0,2718	0,3728	0,5316	0,4243	0,5292
Scanline-42049	0,4926	0,8331	0,5704	0,8648	0,4581
Shangai-license	0,8679	0,6511	0,2025	0,6495	0,5316
Usd-isk	0,7854	0,6093	0,6007	0,7881	0,5956
Well-log	0,2791	0,7289	0,5760	0,6912	0,3589

Ayarlanmış parametreler ile en iyi F1 puanı değerleri parçalı regresyon ile alınmıştır. On bir veri setinin altısında parçalı regresyon, beşinde BinSeg, dördünde AMOC, ikisinde PELT ve Prophet'in 0,7 ve üzerinde F1 puanına sahip olduğu görülmektedir.

Tablo 2.9'e göre parçalı regresyon ve BinSeg algoritmaları denenen tüm gerçek veriler için 0,3 ve üzerinde F1 puanı elde etmişlerdir. Ancak AMOC algoritması için F1 puanı onbir verinin üçünde 0,3'ün altındadır.

Tablo 2.10: Kapsama Ölçütü (Ayarlanmış)

Veri	AMOC	BINSEG	PELT	SEGMENTED	PROPHET
Bitcoin	0,7640	0,4352	0,2182	0,3833	0,2045
Brent-spot	0,4251	0,4858	0,3257	0,5273	0,3206
Children-per women	0,7938	0,7625	0,7635	0,7062	0,3485
Co2- canada	0,5264	0,7001	0,7494	0,6622	0,4366
Debt -Ireland	0,5844	0,8136	0,7042	0,8217	0,6919
Rail-lines	0,7665	0,7731	0,5113	0,6964	0,6222
Rather-stock	0,3870	0,3941	0,4840	0,4488	0,3470
Scanline-42049	0,4305	0,8570	0,4157	0,7944	0,4097
Shangai-license	0,9105	0,7961	0,3955	0,8058	0,3625
Usd-isk	0,8577	0,7549	0,5609	0,7819	0,4211
Well-log	0,4527	0,6371	0,4873	0,5511	0,4615

Ayarlanmış parametreler ile algoritmalar çalıştırıldığında, kapsama ölçütüne göre en iyi sonuçları BinSeg algoritması vermiştir. Denenen on bir verinin yedisinde BinSeg, beşinde AMOC ve parçalı regresyon, üçünde PELT 0,7 ve üzerinde kapsama metriğine sahipken Prophet algoritmasında kapsama metriği 0,7 ve üzerinde olan veri bulunmamaktadır.

Çalışmada incelenen algoritmaların ihtiyaç duyduğu parametre değerleri verilere göre en uygun hale getirildiğinde BinSeg ve parçalı regresyon algoritmalarının diğerlerine göre değişim noktalarının konumlarını belirlemede daha başarılı olduğu görülmüştür.

Elde edilen sonuçlara göre gerçek veriler için algoritmaların performansları arasında anlamlı bir fark olup olmadığı Friedman testi ile araştırılmıştır. Takip eden bölümlerde PMCMRplus paketindeki `friedman.test()` ve `frdAllPairsNemenyiTest()` fonksiyonları ile elde edilen sonuçlara yer verilmiştir. Hipotez testlerinde ($\alpha = 0.05$) alınmıştır.

Friedman rank sum test

```
data: real
Friedman chi-squared = 19.564, df = 4, p-value = 0.0006088
```

H_0 : Algoritmaların performansları arasında istatistiksel olarak anlamlı bir fark yoktur.

H_1 : En az bir algoritmanın performansı farklıdır.

p -değeri 0.05 (α)’den küçüktür. Bu nedenle H_0 hipotezi reddedilmiştir, algoritmaların performanslarının en az birinin farklı olduğu söylenebilir.

Pairwise comparisons using Nemenyi-Wilcoxon-Wilcox all-pairs test for a two-way

```
data: y
```

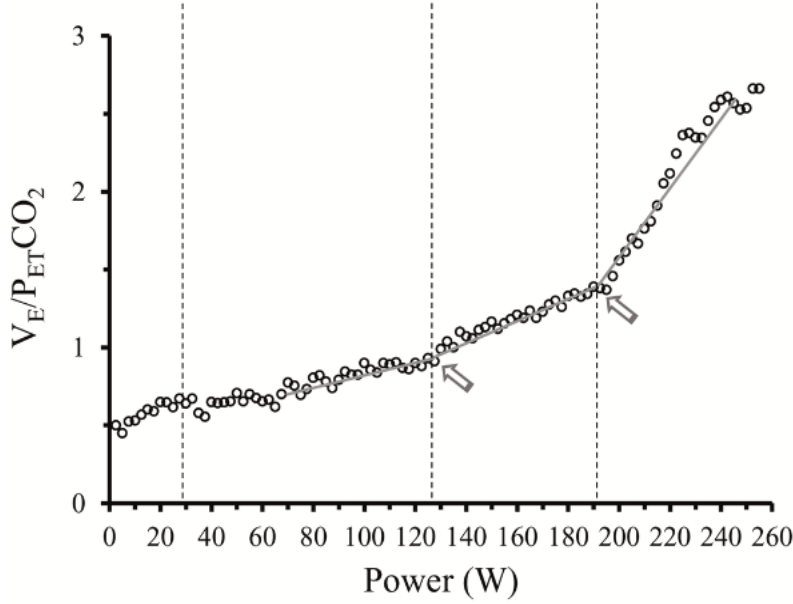
	AMOC	BinSeg	PELT	Parçalı Reg
BinSeg	0.87972	-	-	-
PELT	0.96200	0.48569	-	-
Parçalı Reg	0.96200	0.99884	0.66078	-
Prophet	0.02510	0.00088	0.14727	0.00252

```
P value adjustment method: single-step
```

Nemenyi testi sonuçlarına göre Prophet ile AMOC, BinSeg ve parçalı regresyon algoritmaları arasında istatistiksel olarak anlamlı bir fark vardır.

2.2.3 Kademeli Egzersiz Test Verileri

Bu bölümde bir başka gerçek veri için uygulama yapılmıştır. On beş sporcunun kademeli bir rampa testi sonucunda elde edilen verileri kullanılarak eşik egzersiz şiddetlerini değişim noktası belirleme yöntemleri ile tespit etmek denenmiştir.



Şekil 2.1: $V_E/P_{ET}CO_2$ - Power (W) (Ozkaya, Balci, As ve Yıldıztepe, 2021)

Uygulamada $V_E/P_{ET}CO_2$ (Dakika ventilasyonu bölü ekspirasyon sonu CO_2 kısmi basıncı) değerleri kullanılarak değişim noktası egzersiz şiddeti (Power(W)) olarak bulunmaya çalışılmıştır.

Tablo 2.11: Kapsama Ölçütü (Ayarlanmış)

Veri	BINSEG	SEGMENTED	PELT
1	0,7023	0,6688	0,7340
2	NA	0,8232	NA
3	NA	0,5097	NA
4	0,7485	0,9547	0,7308
5	0,9326	0,8972	0,7314
6	NA	0,7793	NA
7	NA	0,9493	NA
8	1,0000	0,8930	0,8425
9	0,8181	0,9545	0,8723
10	0,8039	0,6843	0,7189
11	1,0000	0,6111	0,5965
12	1,0000	0,7270	0,9353
13	0,6520	0,8486	0,9805
14	0,5434	0,7423	1,0000
15	0,7141	0,4571	0,8194

Ayarlanan parametreler ile hesaplanan kapsama ölçütü değerinin parçalı regresyon algoritmasında yedi, PELT ve BinSeg algoritmalarında ise altı veride 0,8'in üzerinde

olduğu görülmektedir. Ayarlanan parametrelerle çalıştırılan üç farklı algoritmanın kullanıldığı on beş verinin hiçbirinde 0,3'ün altından kapsama ölçütü değeri bulunmamaktadır.

İki değişim noktası tespit edilemeyen durumlar (NA) ile gösterilmiştir.

Tablo 2.12: F1 Puanı (Ayarlanmış)

Veri	BINSEG	SEGMENTED	PELT
1	0,6600	0,6600	0,6600
2	NA	1,0000	NA
3	NA	0,6600	NA
4	1,0000	1,0000	1,0000
5	1,0000	1,0000	1,0000
6	NA	1,0000	NA
7	NA	1,0000	NA
8	1,0000	1,0000	1,0000
9	0,8571	0,8571	1,0000
10	1,0000	1,0000	1,0000
11	0,7499	0,8571	0,8571
12	1,0000	1,0000	1,0000
13	0,5714	0,8571	1,0000
14	0,2857	0,8571	1,0000
15	0,2857	0,4615	0,6600

Ayarlanmış parametreler sonucunda en iyi F1 puanı değerleri parçalı regresyon ile elde edilmiştir. Parçalı regresyon algoritmasında on iki veri 0,8 ve üzerinde F1 puanına sahiptir. Pelt algoritmasında dokuz, BinSeg algoritmasında ise altı verinin 0,8'in üzerinde F1 puanına sahip olduğu görülmektedir.

Pelt ve parçalı regresyon algoritmalarının 0,3'ün altında F1 puanı yokken, BinSeg algoritmasında ise iki veride 0,3'ün altında F1 puanı elde edilmiştir.

İki değişim noktası tespit edilemeyen durumlar (NA) ile gösterilmiştir.

Aşağıda PMCMRplus paketindeki `friedman.test()` ve `frdAllPairsNemenyiTest()` fonksiyonları ile elde edilen sonuçlara yer verilmiştir. Hipotez testlerinde ($\alpha = 0.05$) alınmıştır.

Friedman rank sum test

```
data: ve_petco2
Friedman chi-squared = 0.72727, df = 2, p-value = 0.6951
```

H_0 : Algoritmaların performansları arasında istatistiksel olarak anlamlı bir fark yoktur.

H_1 : En az bir algoritmanın performansı farklıdır.

p -değeri 0.05 (α)’den büyüktür. Bu nedenle H_0 hipotezi reddedilememiştir, kademeli egzersiz test verileri için algoritmaların performansları arasında istatistiksel olarak anlamlı bir fark yoktur.

2.3 R Shiny Uygulaması

Bu uygulama genel olarak değişim noktası içeren yapay veriler üretmeyi ve change-point paketindeki fonksiyonlarla değişim noktalarını tespit etmeyi amaçlamaktadır. Kullanıcılar, herhangi bir kod bilgisine gerek kalmadan uygulama aracılığıyla değişim noktaları analizini gerçekleştirebilir ve sonuçları etkileşimli olarak görselleştirebilir. Ayrıca, kullanıcılar kendi veri setlerini yükleyerek bu veriler üzerinde değişim noktalarını analiz etme olanağına sahiptir. Uygulama, R’deki changepoint paketinin fonksiyonlarını kullanarak değişim noktalarını tespit eder ve görselleştirir. Bu uygulama, değişim noktalarının belirlenmesi ve analiz edilmesi sürecini daha erişilebilir hale getirir, böylece araştırmacılar ve veri analistleri bu yöntemi kullanarak veri setlerindeki önemli değişiklikleri kolayca keşfedebilirler.

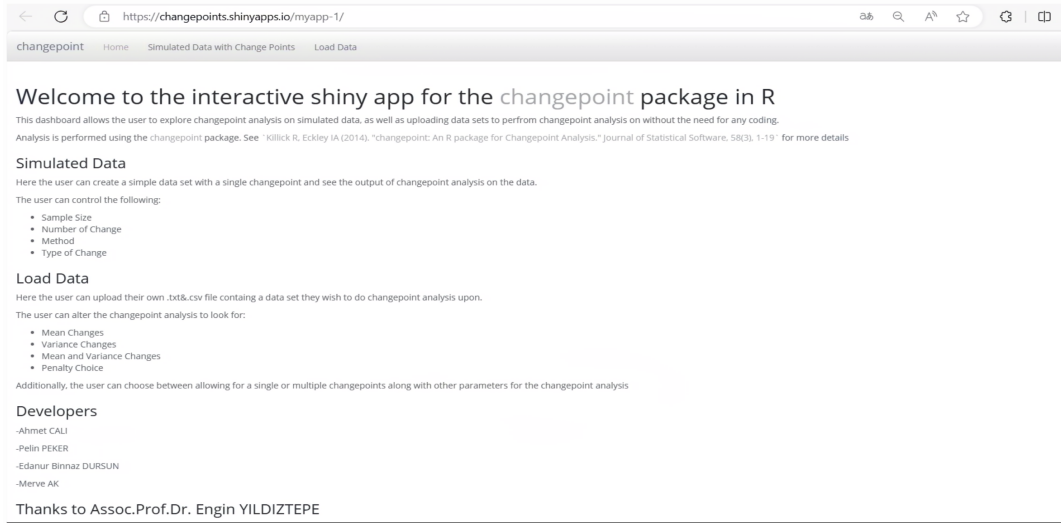
2.3.1 R Shiny Nedir?

R Shiny, R programlama dilinde interaktif web tabanlı uygulamalar oluşturmak için kullanılan bir pakettir. Bu paket, R kullanıcılarının veri analizi ve görselleştirme yeteneklerini web tabanlı arayüzler üzerinden paylaşımlarını sağlar. R Shiny, istatistiksel analizlerin ve veri görselleştirmenin yanı sıra, R dilinin sağladığı zengin paket ve fonksiyonlardan da yararlanabilir. Bu şekilde, kullanıcılar, veri analizlerini ve sonuçlarını paylaşmak için HTML, CSS ve JavaScript gibi web teknolojilerine ihtiyaç duymadan, R kodunu kullanarak web tabanlı arayüzler oluşturabilirler.

2.3.2 Kullanılan Paketler

Bu çalışmada, Shiny uygulamaları oluşturmak için shiny paketi kullanılmıştır. Ayrıca, uygulamanın görünümünü özelleştirmek için shinythemes paketi tercih edilmiştir. Değişim noktaları analizi için ise changepoint paketi kullanılmıştır. Son olarak, veri görselleştirmesi için ggplot2 paketi tercih edilmiştir.

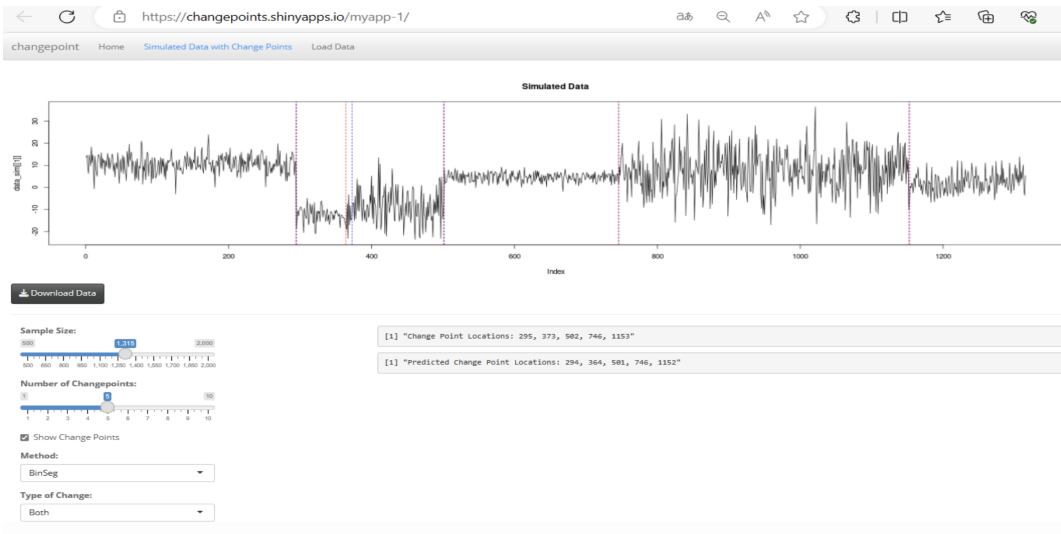
2.3.3 Anasayfa



Şekil 2.2: Anasayfa

Uygulamanın ana sayfası, kullanıcıları karşılayan bir giriş noktasıdır. Burada, uygulamanın genel tanımı ve geliştiricilerin bilgileri yer almaktadır. Ayrıca, kullanıcıları uygulamanın sağladığı imkanlar hakkında bilgilendirir.

2.3.4 Yapay Veri Üretme

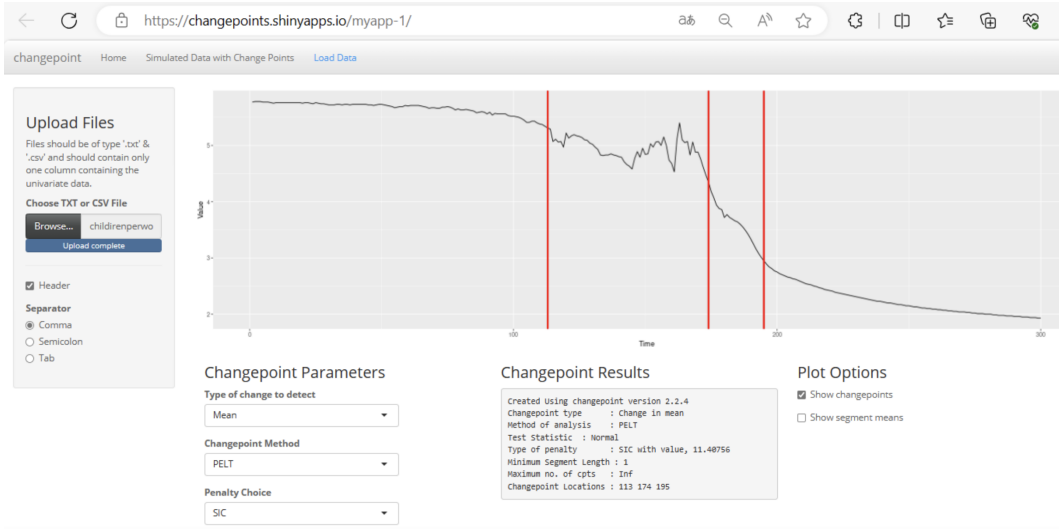


Şekil 2.3: Simüle Edilmiş Veri

Veri Seti Üretme İşlevi: Kullanıcılar, simüle edilmiş veri setini oluşturmak için veri seti üretme panelini kullanabilirler. Bu panel, çeşitli parametreleri içerir:

- Örneklem Sayısı (Sample Size): Kullanıcılar, oluşturulacak veri setinin büyüklüğünü bu parametre ile belirleyebilirler.
- Değişim Noktası Sayısı (Number of Changepoints): Kullanıcılar, veri setindeki değişim noktalarının sayısını belirleyebilirler. Değişim noktaları, veri setinde meydana gelen değişiklikleri temsil eder.
- Kullanıcılar, değişim noktalarını belirlemek için AMOC, BinSeg ve Pelt algoritmaları arasından seçim yapabilirler.
- Değişiklik Türü (Type of Change): Kullanıcılar, değişim noktası tespit etmek için türünü seçebilirler. Bu seçenekler arasında ortalama (Mean), varyans (Variance) veya her ikisi (Both) bulunmaktadır.
- İndirme İşlevi (Download Data): Kullanıcılar, ekranda gözüken veri setini ve analiz sonuçlarını indirebilirler. Dosya .csv formatında kaydedilir. Bu, daha sonra analiz sonuçlarını paylaşmak veya daha detaylı incelemek için kullanılabilir.
- Değişim Noktalarını Göster (Show Change Points): Bu onay kutusu işaretlendiğinde ekranda mavi çizgi ile gerçekte olan değişim noktaları mavi çizgi ile gösterilir.
- Tahmin Edilen Değişim Noktalarını Göster (Show Predicted Show Change Points): Bu onay kutusu işaretlendiğinde ekranda kırmızı çizgi ile algoritmaların tahmin ettiği değişim noktaları mavi çizgi ile gösterilir.

2.3.5 Veri Yükleme



Şekil 2.4: Veri Yükleme

Bu kısımda kullanıcılar tek değişkenli verilerini txt veya csv formatında yükleyerek herhangi bir kod bilgisine gerek kalmadan ‘changepoint’ paketini kullanarak analiz edebilirler. Sunduğu imkanlar şu şekildedir:

- Değişiklik Türü (Type of change to detect): Kullanıcılar ortalama, varyans, ortalama ve varyans seçeneklerinden uygun olanı seçebilirler.
- Değişim Noktası Yöntemi (Changepoint Method):Kullanıcılar, değişim noktalarını belirlemek için AMOC, BinSeg ve Pelt algoritmaları arasından seçim yapabilirler.
- Parametre Seçeneği (Penalty Choice): Kullanıcılar, Changepoint kütüphanesinde bulunan algoritmaların parametre ayarlamalarını, kütüphanenin içerdiği seçeneklerle veya manuel olarak ayarlayabilirler.
- Grafik seçenekleri:Tahmin edilen değişim noktaları ve parça ortalamaları grafik üzerinde opsiyonel olarak gösterilebilir.

Bölüm 3

Sonuç

Bu çalışmada, değişim noktalarını tespit etmek için AMOC, BinSeg, parçalı regresyon, PELT ve Prophet algoritmaları kullanılmıştır ve performansları karşılaştırılmıştır. Yapay veri, gerçek veri ve kademeli egzersiz test verileri kullanılarak gerçekleştirilen analizlerde, algoritmaların performansları F1 puanı ve kapsama ölçütü ile değerlendirilmiştir.

Sonuçlar, yapay verilerde BinSeg algoritmasının F1 puanı açısından en başarılı sonuçları verdiğini, gerçek verilerde ve kademeli egzersiz test verilerinde ise parçalı regresyon algoritmasının en iyi performansı sergilediğini göstermiştir. Kapsama ölçütüne göre, yapay verilerde BinSeg ve parçalı regresyon algoritmalarının, gerçek verilerde ve kademeli egzersiz test verilerinde parçalı regresyon algoritmasının en iyi sonuçları elde ettiği tespit edilmiştir.

Friedman testi sonuçlarına göre, yapay verilerde AMOC ve BinSeg, parçalı regresyon ve AMOC, Prophet ile BinSeg ve parçalı regresyon, PELT ile BinSeg ve parçalı regresyon algoritmaları arasında istatistiksel olarak anlamlı farklar bulunmuştur. Gerçek verilerde ise Prophet algoritması ile AMOC, BinSeg ve parçalı regresyon algoritmaları arasında istatistiksel olarak anlamlı farklar mevcuttur. Ancak, kademeli egzersiz test verilerinde algoritmaların performansları arasında istatistiksel olarak anlamlı bir fark bulunmamıştır.

Bu bulgular, farklı veri türlerinde değişim noktası tespit yöntemlerinin performanslarının değişebileceğini göstermektedir. Bu nedenle, belirli bir veri seti ve analiz amacı için en uygun algoritmanın seçilmesi önem arz etmektedir. Çalışmada elde edilen bulgular, gelecekteki araştırmalara ve pratik uygulamalara rehberlik edecek niteliktedir. Ayrıca, değişim noktası tespiti için geliştirilen Shiny web uygulaması ve ilgili R ve Python kodları, araştırmacıların bu yöntemleri kendi verilerine uygulayabilmeleri için önemli bir kaynak sunmaktadır.

Kaynaklar

- Aminikhanghahi, S. ve Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2), 339-367.
- Birgé, L. ve Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability theory and related fields*, 138, 33-73.
- Chen, J., Gupta, A. K. ve Gupta, A. (2000). *Parametric statistical change point analysis* (C. 192). Springer.
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7, 1-30.
- Guyon, X. ve Yao, J. (1999). On the underfitting and overfitting sets of models chosen by order selection criteria. *Journal of Multivariate Analysis*, 70(2), 221-249.
- Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables.
- Jen, T. ve Gupta, A. K. (1987). On testing homogeneity of variances for Gaussian models. *Journal of Statistical Computation and Simulation*, 27(2), 155-173.
- Kelle, A. C. ve Hüseyin, Y. (2022). MQTT Trafiğinde DoS Saldırılarının Makine Öğrenmesi ile Sınıflandırılması ve Modelin SHAP ile Yorumlanması. *Journal of Materials and Mechatronics: A*, 3(1), 50-62.
- Killick, R., Fearnhead, P. ve Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500), 1590-1598.
- Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal processing*, 85(8), 1501-1510.
- Lykou, R., Tsaklidis, G. ve Papadimitriou, E. (2020). Change point analysis on the Corinth Gulf (Greece) seismicity. *Physica A: Statistical Mechanics and its Applications*, 541, 123630.
- Nemenyi, P. B. (1963). *Distribution-free multiple comparisons*. Princeton University.
- Ozkaya, O., Balci, G. A., As, H. ve Yildiztepe, E. (2021). A new technique to analyse threshold-intensities based on time dependent change-points in the ratio of minute ventilation and end-tidal partial pressure of carbon-dioxide production. *Respiratory Physiology & Neurobiology*, 294, 103735.
- Scott, A. J. ve Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 507-512.
- Sen, A. ve Srivastava, M. S. (1975). On tests for detecting change in mean. *The Annals of statistics*, 98-108.
- Taylor, S. J. ve Letham, B. (2018). Forecasting at scale. *The American Statistician*,

72(1), 37-45.

Van den Burg, G. J. ve Williams, C. K. (2020). An evaluation of change point detection algorithms. *arXiv preprint arXiv:2003.06222*.

Zhao, N., Liu, Y., Vanos, J. K. ve Cao, G. (2018). Day-of-week and seasonal patterns of PM_{2.5} concentrations over the United States: Time-series analyses using the Prophet procedure. *Atmospheric environment*, 192, 116-127.