POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Yörünge verisi yayınlamada mahremiyet duyarlı yeni bir model önerisi ve uygulaması

A new privacy-aware model proposal and application on trajectory data publishing

***Yazar(lar) (Author(s))**: Murat AKIN[1], Yavuz CANBAY[2], Şeref SAĞIROĞLU[3]*

ORCID[1]: 0000-0003-0001-1036

ORCID[2]: 0000-0003-2316-7893

ORCID[3]: 0000-0003-0805-5818

Bu makaleye şu şekilde atıfta bulunabilirsiniz(To cite to this article): Akın M., Canbay Y. ve Sağıroğlu S., "Yörünge verisi yayınlamada mahremiyet duyarlı yeni bir model önerisi ve uygulaması", ***Politeknik Dergisi***, 24(3): 1275-1286, (2021).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: **10.2339/politeknik.916234**

Yörünge Verisi Yayınlamada Mahremiyet Duyarlı Yeni Bir Model Önerisi ve Uygulaması

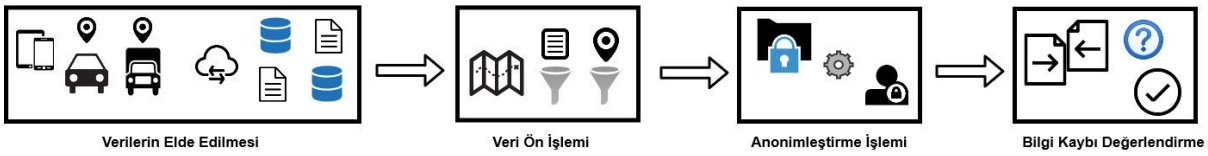
A New Privacy-aware Model Proposal and Application on Trajectory Data Publishing

Önemli noktalar (Highlights)

- *Yörünge Verisi Yayınlama / Trajectory Data Publishing*
- *Mahremiyet Duyarlı Model / Privacy-aware Model*
- *Diferansiyel Mahremiyet / Differential Privacy*

Grafik Özet (Graphical Abstract)

Bu çalışmada, yörünge verilerinin yayınlanmasında diferansiyel mahremiyeti kullanan yeni bir model önerilmiş, başarıyla geliştirilmiş ve gerçek bir veri kümesi üzerinde test edilmiştir. / In this paper, a new trajectory data publishing model using differential privacy was proposed, developed and tested on a real dataset.



Şekil. Önerilen modelin akış şeması /**Figure.** General flowchart of the proposed model

Amaç (Aim)

Mahremiyet korumalı yörünge verisi yayınlamak için yeni bir model geliştirilmesi amaçlanmıştır. / It was aimed to develop a new model for privacy preserving trajectory data publishing.

Tasarım ve Yöntem (Design & Methodology)

Önerilen modelde, veri kümesinde bulunan her bir yörünge verisi için pencere temelli bir yaklaşım kullanılarak gruplama, ortalama alma ve diferansiyel mahremiyet temelli gürültü ekleme işlemi gerçekleştirilir. / In the proposed model, for each trajectory data in the data set, grouping, averaging and differential privacy based noise adding steps were performed by using a window-based approach.

Özgünlük (Originality)

Pencere tabanlı bir yaklaşımla yörünge verilerini gruplayıp lokal diferansiyel mahremiyeti uygulayan ilk mahremiyet modelidir. / This is the first privacy model of applying local differential privacy with grouping trajectory data by a window-based approach.

Bulgular (Findings)

PB=2 ve ε=2 için yapılan testlerde hem örnek araç hem de tüm araçların X ve Y koordinatları için 0,001'den küçük hata değerleri alınmıştır. Bu değerler veri faydası açısından anonimleştirilmiş yörünge bilgilerinin kullanılabilir olduğunu göstermektedir. / In the experiments, less than 0,001 error rates were obtained for X and Y coordinates of both all vehicles and sample vehicle when PB=2 and ε=2. The findings have shown the anonymized trajectory data is usable in terms of data utility.

Sonuç (Conclusion)

Yapılan deneyler sonucu elde edilen bulgular incelendiğinde, önerilen modelin mahremiyet korumalı yörünge verisi yayınlamada başarıyla kullanılabileceği görülmüştür. / When the results of experiments were evaluated, it was clearly seen that the proposed model can be used in privacy preserving trajectory data publishing successfully.

Etik Standartların Beyanı (*Declaration of Ethical Standards*)

Bu makalenin yazar(lar)ı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler. / The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Yörünge Verisi Yayınlamada Mahremiyet Duyarlı Yeni Bir Model Önerisi ve Uygulaması

Araştırma Makalesi / Research Article

Murat AKIN[1,3*], Yavuz CANBAY[2], Şeref SAĞIROĞLU[1]
[1]Gazi Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü, Ankara, Türkiye
[2]Kahramanmaraş Sütçü İmam Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Türkiye
[3]Başarsoft Bilgi Sistemleri A.Ş., Ankara, Türkiye

(Geliş/Received : 15.04.2021 ; Kabul/Accepted : 20.05.2021 ; Erken Görünüm/Early View : 01.06.2021)

ÖZ

Konum tabanlı servisler (KTS), sağladıkları bilgi ve yönlendirmeler ile gündelik hayatı kolaylaştırmaktadır. Kullanıcıların KTS'leri kullanarak gezinmesi sonucu elde edilen konum bilgileri zamana göre sıralandığında, yörünge verileri oluşmaktadır. Bu veriler, KTS sağlayıcıları tarafından toplanmakta, depolanmakta, işlenmekte ve çeşitli gerekçelerle yayınlanmaktadır. Yörünge verileri kişisel veri olarak değerlendirildiği için, bu tür veriler orijinal hali ile yayınlanırsa, saldırganlar kurbanları hakkında hassas bilgilere ulaşabilir ve ifşa saldırıları düzenleyebilir. Bu problemi gidermek için mahremiyet koruyucu güncel yaklaşımlara her zaman ihtiyaç vardır. Bu çalışmada, yörünge verilerinin mahremiyetini koruyarak yayınlanmasını sağlamak için diferansiyel mahremiyet tabanlı yeni bir anonimleştirme modeli önerilmiş, geliştirilmiş ve başarıyla test edilmiştir. Elde edilen sonuçlar, önerilen modelin mahremiyet korumalı yörünge verisi yayınlamada sadece araştırmalar için değil aynı zamanda gerçek uygulamalar için de başarıyla kullanılabileceğini göstermektedir.

Anahtar Kelimeler: Konum verisi, yörünge verisi, anonimleştirme, diferansiyel mahremiyet, veri yayınlama.

A New Privacy-Aware Model Proposal and Application on Trajectory Data Publishing

ABSTRACT

Location-based services (LBS) make daily life easier with the information and directions they provide. Trajectory data is generated when the location information acquired from users utilizing LBS is sorted according to time. Such kind of data are collected, stored, processed and published by LBS providers for various reasons. Since trajectory data is considered as personal data, attackers may obtain sensitive information about their victims and perform disclosure attacks if the trajectory data is published in original form. There is always a need for up-to-date privacy preserving approaches to address this problem. In this study, in order to publish privacy preserved trajectory data, a new anonymization model based on differential privacy was proposed, developed and successfully tested. The obtained results have shown that the proposed model might be successfully used for privacy preserving trajectory data publishing, not only research purposes but also real time applications.

Keywords: Location data, trajectory data, anonymization, differential privacy, data publishing.

1. GİRİŞ (INTRODUCTION)

İnsanlar arası iletişimin artması, kullanılan internetin, cihazların ve uygulamaların yaygınlaşması ile gündelik hayatı kolaylaştıran birçok uygulama ve servis geliştirilmektedir. Konum tabanlı uygulamalar ve servisler ise bunlardan en popüler olanıdır. Yol tarifi, araç takibi, rota belirleme gibi pek çok alanda insanlara büyük kolaylıklar sunan bu uygulamalar, temel olarak Küresel Konumlama Sistemi (GPS) bilgisine dayalı olarak çalışmaktadır.

Bu tür uygulamaları geliştiren kurum ya da şirketler, kişilerin kullandıkları cihazlardaki servislerden elde ettikleri GPS verilerini kendi sistemlerinde depolamakta ve işlemektedir. Çoğu zaman KTS sağlayıcıları tarafından işlenen bu veriler daha fazla değer üretilmesi ve analizler yapılması amacıyla kamuya veya araştırmacılara açılabilmekte ve yayınlanabilmektedir. Ancak böylesi bir durumda, veri orijinal hali ile yayınlanırsa saldırganların doğrudan hedefi haline gelebilmekte, saldırganlar bu veriler üzerinde kurbanları hakkında pek çok hassas bilgiye erişebilmekte ve kimlik ifşaları gerçekleştirebilmektedir [1]. Bunu önlemek için veri yayınlamada mahremiyete her zaman önem verilmesi gerektiği aşikârdır.

KTS tabanlı uygulamalarda konum verileri ile birlikte zaman bilgileri de tutulmaktadır. Mekân-zamansal bilgi olarak ifade edilen konum-zaman verileri, zamana göre sıralandığında kişiye ait rota ve yörünge bilgileri oluşturulur. Yörünge verilerinden elde edilen davranış örüntülerinin analizi ve madenciliği, trafik tıkanıklık yönetimi, şehir sakinlerinin günlük yolculuklarını planlama gibi birçok konuda şehir planlama ve akıllı ulaşım sistemleri için karar mekanizmalarına destek olmaktadır [2].

*Sorumlu Yazar (Corresponding Author)
e-posta : muratakin@gazi.edu.tr

Mahremiyet, literatürde yalnız bırakılma hakkı olarak tanımlanır [3]. Gerçek hayatta kendisini varlığı ile temsil eden insanoğlu, dijital ortamda bu temsili kişisel verisi ile gerçekleştirir. Böylesi bir temsilde ön plana çıkan veri mahremiyeti genel olarak, "muhatabın bilgilerinin doğru kullanımı ve hangi bilgisinin kiminle ve ne derecede paylaşılmasına karar verme mekanizması" [4] olarak ifade edilmektedir.

Günümüzde hayatı kolaylaştıran birçok yazılım ve servis kişisel verileri kullandığı için, veri sahiplerinin mahremiyetini koruma hususu da büyük önem kazanmıştır. Özellikle son yıllarda popülerliğini koruyan Kişisel Verileri Koruma Kanunu (KVKK) ve GDPR (General Data Protection Rule) gibi mevzuatlar kapsamında konum verileri de kişisel veri olarak [5, 6] ele alındığı için, kişisel konum verilerinin anonimleştirilmesi ihtiyacı doğmuştur. Teknolojik olarak gelinen durum ve kişisel verilerin farklı ortamlardan çok rahatlıkla elde edilmesiyle birlikte, mahremiyet kavramı bu tür verilerin toplandığı ve işlendiği her alanda dikkate alınması gereken önemli bir konu olmuştur.

Literatürde yörünge mahremiyeti sağlamada, k-anonimlik [7] ve l-çeşitlilik [8] gibi temel modellerinin yanı sıra son zamanlarda diferansiyel mahremiyet modelinden de faydalanıldığı görülmektedir. İstatistiksel sorgulara gürültü ekleme yaklaşımına dayanan diferansiyel mahremiyette, ilgili veri kümesinde bir kaydın çıkarılması veya eklenmesi o veri kümesinden elde edilecek istatistiksel cevapları etkilememektedir [9]. Bunun temel gerekçesi ise sorgu sonuçlarına çeşitli mekanizmalar kullanılarak gürültüler eklenmesidir [10, 11].

Bu çalışmada; mahremiyet korumalı yörünge verisi yayınlama üzerine odaklanılmış ve bu verilerin anonimleştirmesi için yeni bir model geliştirilmiş, uygulanmış ve farklı parametrelerle test edilerek modelin başarısı ortaya konulmuştur.

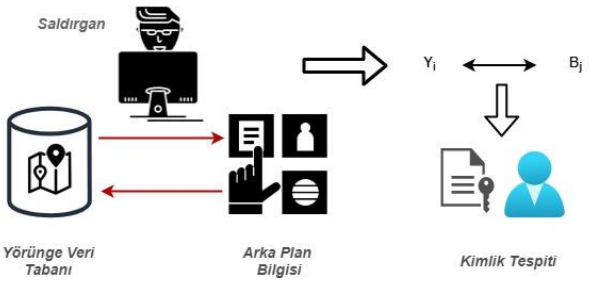
Bu makale 8 bölümden oluşmaktadır. Makalenin ikinci bölümünde problem tanımı verilmiş, üçüncü bölümünde literatür özeti sunulmuş, dördüncü bölümünde yörünge mahremiyeti hakkında bilgi verilmiş, beşinci bölümünde önerilen model hakkında temel bilgiler sunulmuş, altıncı bölümünde önerilen model detaylı olarak açıklanmış, yedinci bölümünde deneysel çalışmalara yer verilmiş ve son bölümde ise sonuçlar verilerek çeşitli değerlendirmelerde bulunulmuştur.

2. PROBLEMİN TANIMI (PROBLEM DEFINITION)

Yörünge verileri, içerisinde birden fazla konum verisi ve bunlarla eşleştirilmiş zaman verisi barındıran, çeşitli rotalarda hareket eden mekân-zamansal verileri temsil eder. Günümüzde özellikle çok sayıda cihazdan farklı KTS'lerle kolay bir şekilde toplanabilen yörünge verileri ile pek çok alanda değerler üretilebilmektedir.

Yörünge verisi mahremiyeti, bu verilerden veri sahibine ulaşmayı mümkün olduğu kadar engelleyen kurallar bütünü olarak tanımlanabilir [12]. Kamuyla ya da araştırmacılarla paylaşılan yörünge verileri kimliksizleştirilse bile farklı veri kaynaklarından elde edilen bilgilerle eşleştirilmesi sonrası kimlik belirleme işlemi yapılabilir ve böylelikle kimlik ifşası gerçekleşebilmektedir. Örneğin, bir filo şirketine ait kimliksiz olarak yayınlanan yörünge veri kümesi, arka plan bilgileri ile eşleştirilerek kurban hakkında sağlık, dini görüş, siyasi görüş, ev ve iş yeri bilgisi, sürüş davranışı ve sağlık durumu gibi pek çok hassas bilgiyi elde etmeyi mümkün kılar [13].

Bu çalışmada ele alınan problem Şekil 1'de gösterilmektedir. Yörünge veri tabanı *Y*, kimlikli arka plan bilgi kümesi *B*, $Y=\{Y1,\ldots Yn\}$ ve $B=\{B1,\ldots Bm\}$ olmak üzere; eğer çeşitli özellikler kullanılarak Y_i ile B_j eşleştirilirse bu durumda *Y* veri tabanındaki kurbanın kimliği açığa çıkar ve ifşa saldırısı gerçekleşir.



Şekil 1. Yörünge verilerinde ifşa saldırısı problemi (Disclosure attack problem on trajectory data)

3. LİTERATÜR TARAMASI (LITERATURE REVIEW)

Bu bölümde, mahremiyet korumalı yörünge verisi yayınlama üzerine bir literatür araştırması yapılmış, kullanılan veri kümeleri ve yöntemler sunularak karşılaştırılmıştır.

Li ve arkadaşları [14], yörünge verilerinin yayınlanmasında mahremiyet koruyucu önlem olarak diferansiyel mahremiyet yöntemini uygulamış, mevcut yöntemlerin zayıf yanlarını iyileştiren yeni bir yaklaşım önermişlerdir. Microsoft tarafından yayınlanan T-Drive verisi üzerinde yapılan testlerde, Mutual Information metriği veri faydasını ölçmek için kullanılmış ve önerilen modelin diğer modellere göre daha yüksek veri faydası sunduğu belirtilmiştir.

Chen ve arkadaşları [15], farklı türdeki yörünge verilerine uygulanabilecek veri bağımlı ve diferansiyel mahremiyet temelli yeni bir veri yayınlama algoritması geliştirmiştir. Literatürde k-anonimliği uygulayan çeşitli çalışmaların yeterli düzeyde mahremiyeti sağlayamadığı ilgili çalışmada belirtilmiş ve bu olumsuzluk giderilmeye çalışılmıştır. STM tarafından yayınlanan yörünge veri kümesi üzerinde yapılan testlerde, ortalama bağıl hata metriği kullanılarak iki farklı senaryoya göre deneyler yapılmış ve yüksek veri faydası elde edilmiştir.

Han ve arkadaşlarının yaptığı çalışmada [16], literatürdeki çalışmaların eksikliklerinden yola çıkarak,

mekânsal parçalama tabanlı diferansiyel mahremiyet duyarlı yeni bir veri yayınlama metodu geliştirilmiştir. Testlerde Microsoft tarafından yayınlanan T-Drive verisi kullanılmış ve Haussdorf uzaklığı ile de veri faydası ölçülmüştür. Önerilen modelin, literatürdeki diğer modeller ile kıyaslandığında daha yüksek veri faydası sunduğu belirtilmiştir.

He ve arkadaşları [17], diferansiyel mahremiyet temelli yörünge verisi yayınlama modeli geliştirmiştir. 2009 yılı Mayıs ayında Pekin'de bulunan taksilerden toplanan veri kümesi ve Oldenburg sentetik veri kümesi üzerinde yapılan deneylerde, geliştirilen 3 farklı metrik kullanılarak veri faydası ölçülmüş ve başarılı sonuçlar elde edilmiştir.

Gürsoy ve arkadaşları çalışmalarında [18], diferansiyel mahremiyeti kullanarak yüksek veri faydası sunan yeni bir mahremiyet korumalı veri yayınlama modeli önermiştir. Microsoft'un yayınladığı Geolife veri kümesi, Portekiz-Porto'daki taksilerden toplanan GPS konum veri kümesi ve Oldenburg sentetik veri kümesinin kullanıldığı çalışmada ortalama bağıl hata metriğinden faydalanılarak veri faydası ölçülmüş ve farklı epsilon değerleri için yapılan testlerde önerilen yöntemin yüksek veri faydası sunduğu görülmüştür.

Cao ve Yoshikawa [19], geliştirdikleri modelde yörünge verilerinin mahremiyetini korumak için diferansiyel mahremiyet tekniğinden faydalanmıştır. Veri faydasının ortalama mutlak hata metriği ile ölçüldüğü çalışmada, PeopleFlow, Geolife ve T-Drive veri kümeleri üzerinde testler gerçekleştirilmiştir. Yapılan testler sonucunda, önerilen modelin literatürdeki dört farklı modelden daha yüksek veri faydası sunduğu bildirilmiştir.

Tian ve arkadaşları [20], mahremiyet duyarlı yörünge verisi yayınlamada kişiselleştirilmiş diferansiyel mahremiyet yaklaşımını uygulayan yeni bir model önermiştir. Yörünge verisinin karakteristiğini çıkarmak amacıyla Hilbert eğrisinin kullanıldığı çalışmada, önerilen modeli test etmek amacıyla Microsoft'un yayınladığı T-Drive veri kümesinden faydalanılmıştır. Hausdorff uzaklığı metriği kullanarak veri faydası ölçülmüş, farklı küme sayıları için bu metrik değerlendirilmiş ve başarılı sonuçlar elde edilmiştir.

Zhao ve arkadaşları [21], yörünge verilerinin mekân-zamansal bütünlüğünü koruyarak yeni bir mahremiyet korumalı veri yayınlama modeli geliştirmiştir. Diferansiyel mahremiyetin kullanıldığı çalışmada, Gutenberg sentetik verisi üzerinde önerilen model test edilmiştir. Bağıl hata metriği veri faydası ölçmede kullanılmış, farklı epsilon değerlerine göre testler yapılmış ve literatürdeki yöntemlerle yapılan kıyaslamalara göre önerilen modelin daha yüksek veri faydası sunduğu bildirilmiştir.

Zhao ve arkadaşları [22], diferansiyel mahremiyeti kullanarak risk duyarlı yeni bir yörünge verisi yayınlama modeli geliştirmiştir. Literatürdeki diğer modellerin veri kümesindeki her bir yörünge verisinin tekil kişilere ait olmasını kabul etmesini bir dezavantaj olarak gösteren bu çalışma, belirtilen problemin çözümüne odaklanmıştır. Diferansiyel mahremiyetin ele alındığı çalışmada, D4D yörünge veri kümesi üzerinde testler gerçekleştirilmiştir. Veri faydasını ölçmede Top-k metriği kullanılmış ve önerilen model literatürdeki farklı modellerle karşılaştırılmıştır. Farklı parametrelere göre yapılan testler sonucunda, önerilen modelin daha yüksek veri faydası sunduğu ifade edilmiştir.

Jiang ve arkadaşları [23], diferansiyel mahremiyeti uygulayan yörünge verisi yayınlama modeli önermiştir. Tüm yörünge verisine, örnekleme tabanlı olarak her bir pozisyona ve her bir koordinata olmak üzere üç farklı yaklaşımla gürültü ekleme işleminin yapıldığı modelde, üçüncü yaklaşımın en yüksek mahremiyeti sağladığı vurgulanmıştır. Uygulamaya özel olarak toplanan yörünge verileri üzerinde yapılan testlerde, Öklit uzaklığı veri faydası ölçmede kullanılmış ve başarılı sonuçlar elde edildiği belirtilmiştir.

Han ve arkadaşları ise [24], lokal diferansiyel mahremiyeti uygulayarak yeni bir yörünge verisi yayınlama modeli geliştirdi. Korelasyon tabanlı hassas verilerin hassas olmayanlarla değiştirilmesi planına dayanan modeli test etmede Gowalla veri kümesi kullanılmıştır. Veri faydası metriği olarak kl-divergence tekniğinden faydalanılmış ve literatürdeki farklı modellerle yapılan kıyaslamalarda yüksek veri faydası elde edildiği gözlemlenmiştir.

Nakamura ve Nishi [25], k-anonimlik yönteminden türetilen yeni bir yaklaşım önermiştir. Önerilen yaklaşımda konum verilerine rastgele ve belirli ölçek aralığında gürültü eklenerek mahremiyet sağlanmıştır. Xie ve arkadaşları [26] ise kullanıcılara yöneltilen anket soruları ve kayıt altına alınan konum bilgilerinin mahremiyetini koruyarak veri toplanmasını amaçlamıştır. Verileri toplarken Gauss dağılım kullanan araştırmacılar, önerilen yöntemin normal dağılımla yapılan çalışmalara oranla daha iyi sonuç verdiğini ifade etmiştirler. Her iki çalışmada da veri faydasının ölçümü için RMSE (Root Mean Square Error) yöntemi kullanılmıştır.

Çizelge 1'de sunulan bilgiler ele alındığında, genellikle açık veri kümelerinin kullanıldığı, RMSE hata metriğinin veri faydasını ölçmede tercih edildiği, çalışmalara özgü farklı yöntemlerin kullanılmasıyla beraber diferansiyel mahremiyetin sağlanması için Laplace gürültüsünden faydalanıldığı, global ve lokal seviyede diferansiyel mahremiyetin sağlandığı görülmüştür. Bu gerekçeler dikkate alınarak bu çalışmada RMSE metriğinden ve diferansiyel mahremiyet için Laplace mekanizmasından faydalanılmış, önerilen modelin bir gereksinimi olarak lokal seviyede veri mahremiyeti uygulanmıştır.

4. YÖRÜNGE VERİSİNE YÖNELİK MAHREMİYET İFŞA SALDIRILARI VE MAHREMİYET KORUMA YÖNTEMLERİ (PRIVACY DISCLOSURE ATTACKS AND PRIVACY PRESERVING METHODS ON TRAJECTORY DATA)

Çizelge 1. Literatürdeki çalışmaların karşılaştırılması (Comparison of studies in the literature)

No	Veri Kümesi	Yöntem	Diferansiyel Mahremiyet Yaklaşımı	Metrik	Kod Erişilebilirliği
[14]	T-Drive	*k*-means kümeleme	Global, Laplace Gürültüsü	Mutual Information	Mevcut değil
[15]	STM	Prefix Tree	Global, Laplace Gürültüsü	Ortalama bağıl hata	Mevcut değil
[16]	T-Drive	Hilbert eğrisi veri uzayı parçalama	Global, Laplace Gürültüsü	Haussdorf uzaklığı	Mevcut değil
[17]	Pekin taksi verisi ve Oldenburg sentetik verisi	Hiyerarşik referans sistemi, Prefix Tree	Global, Laplace Gürültüsü	Çalışmaya özgü üç farklı metrik	Mevcut değil
[18]	Geolife, Porto taksi verisi, Oldenburg sentetik verisi	İstatistiksel tabanlı sentetik veri üretme	Global, Laplace Gürültüsü	Ortalama bağıl hata	Mevcut değil
[19]	PeopleFlow, Geolife, T-Drive	İstatistik tabanlı yakınsama stratejisi	Global, Laplace Gürültüsü	Ortalama mutlak hata	Mevcut değil
[20]	T-Drive	Hilbert lineer indeks kümeleme	Global, Laplace Gürültüsü	Haussdorf uzaklığı	Mevcut değil
[21]	Gutenberg sentetik verisi	R-tree veri uzayı parçalama	Global, Laplace Gürültüsü	Bağıl hata	Mevcut değil
[22]	D4D	Risk temelli yaklaşım	Global, Laplace Gürültüsü	Top-k	Mevcut değil
[23]	Özel veri kümesi	Örneklem tabanlı yaklaşım	Global, Laplace Gürültüsü	Öklid	Mevcut değil
[24]	Gowalla	Korelasyon tabanlı hassas veri değişimi	Lokal, Laplace Gürültüsü	KL-Divergence	Mevcut değil
[25]	Cleveland ve Hungarian veri kümesi	TMk-anonymity yaklaşımı	Global, Laplace Gürültüsü	RMSE	Mevcut değil
[26]	General Practices Surgeries veri kümesi	Gaussian dağılımından geliştirilen yeni bir yaklaşım	Global, Laplace Gürültüsü	RMSE	Mevcut değil
Mevcut Çalışma	Özel veri kümesi (filo verisi)	Ortalama alma	Lokal, Laplace Gürültüsü	RMSE	Mevcut[1]

Bu bölümde, yörünge verilerinin mahremiyetini ifşa etmede kullanılan saldırılar ve mahremiyet koruma yöntemleri sunulmuştur.

4.1 Yörünge Verisine Yönelik Mahremiyet İfşa Saldırıları (Privacy Disclosure Attacks On Trajectory Data)

Saldırganlar, yörünge verilerinden veri sahiplerine ulaşmak için çeşitli ifşa saldırıları düzenleyebilir. Kamuya açık verilerden elde edilebilen arka plan bilgileri, ifşa saldırılarının yaşanmasında önemli rol oynar ve bu bilgiler yayınlanan veri kümeleriyle eşleştirildiğinde çeşitli mahremiyet ihlalleri ortaya çıkar [27].

Kurbanı hakkında arka plan bilgisine sahip bir saldırgan;

- Kimlik bilgileri içeren veri kümeleri ile yayınlanan kimliksiz verileri çeşitli öznitelikler düzeyinde eşleştirerek yayınlanan kimliksiz veri içerisindeki kurbana ait kimlik bilgilerini ifşa edebilir [28],
- Yayınlanan kimliksiz verideki hassas özniteliklerin homojen dağılımına bağlı olarak kurbanın hassas bilgilerini ifşa edebilir [29] veya
- Yayınlanan kimliksiz veride çeşitli üyelik çıkarımları yapabilir [30].

Yukarıdaki bilgilerden görüldüğü üzere, saldırgan kurbanı hakkında arka plan bilgisine sahip olduğunda pek çok saldırı düzenleyerek onun hassas bilgilerine ulaşabilme imkânına sahip olabilir. Bu tür saldırıları bertaraf etmek için literatürde çeşitli yaklaşımlar mevcut olup sonraki bölümde bunlar hakkında detaylı bilgiler verilmiştir.

4.2. Yörünge Verisinde Mahremiyet Koruma Yaklaşımları (Privacy Preserving Approaches on Trajectory Data)

Yörünge verisi mahremiyetinin ihlali ile ortaya çıkabilecek sorunları önlemek için kişilerin konum ve yörünge bilgisini koruyan çeşitli yaklaşımlar literatürde mevcuttur. k-anonimlik [7] ve l-çeşitlilik [31] gibi temel mahremiyet koruma yöntemlerinin yanı sıra, diferansiyel mahremiyet yöntemi de bu alanda sıklıkla kullanılmaktadır.

k-Anonimlik: en temel mahremiyet koruma yaklaşımı olup, veri kümesindeki bir verinin en az k-1 tane veriden ayırt edilememesini sağlar [32]. Bir saldırgan önceden elde ettiği kimlikli bir veri kümesini k-anonim olarak yayınlanan veri kümesi ile eşleştirse bile kurbanını ifşa edebilme ihtimali 1/k oranındadır. Optimum k-

[1] https://github.com/ycanbay/anonym_traj

anonimliği sağlamanın NP-Zor bir problem olduğu çeşitli çalışmalarda [33-36] ispatlanmıştır.

l-Çeşitlilik: Machaanavajjhala ve arkadaşları [29] tarafından önerilen bu yaklaşım, k-anonimliğin hassas öznitelikleri koruyamaması sorununu çözmek için geliştirilmiştir. Hassas verilerin ifşa edilmesini mümkün olabildiği kadar engellemek amacıyla bu verilerin çeşitliliğinin en az l sayıda olmasını sağlar.

Diferansiyel mahremiyet: Dwork ve arkadaşları [9] tarafından önerilen bu yaklaşım, veri tabanları üzerinde yapılan istatistiksel analizlere gürültü ekleme metoduna dayanır. k-anonimlik ve l-çeşitlilik gibi arka plan bilgisi saldırısına karşı zafiyeti olan yöntemlere kıyasla, bu saldırıyı bertaraf edecek bir yapıya sahip olmasından dolayı günümüzde sıklıkla tercih edilmektedir.

Diferansiyel mahremiyetin temel çalışma prensibi, veri tabanına gönderilen sorgunun döndüreceği gerçek cevaba belirli ölçülerde gürültü ekleme işlemine dayanır. Şekil 2'de diferansiyel mahremiyetin uygulanması süreci gösterilmiştir. Bu süreçte araştırmacı, veri tabanına istatistiksel bir sorgu ("Sum", "Count", "Mean" vb.) gönderir. Bu sorgu veri tabanında çalıştırılır ve dönecek gerçek cevap diferansiyel mahremiyet mekanizmasına iletilir. Bu mekanizma kullanılarak gerçek cevaba gürültü eklenerek araştırmacıya gönderilir. Bu şekilde araştırmacının gerçek cevaba ulaşması yerine yaklaşık cevaplara ulaşması sağlanır.

Şekil 2. Diferansiyel mahremiyet süreci (Process of differential privacy)

Diferansiyel mahremiyet genellikle şu şekilde ifade edilir; d boyutlu r kayıtlarının oluşturduğu bir veri uzayında; D_1 ve D_2 komşuluk veri kümelerini ve f ise bir sorgu fonksiyonunu temsil etsin. Diferansiyel mahremiyetin temel amacı D_1 ve D_2 veri kümeleri üzerinde çalıştırılacak f sorgusunun sonuçları arasındaki farkı minimize etmektir. Bu veri kümelerine uygulanacak herhangi bir f sorgusunun sonuçları arasındaki maksimum farkın Δf hassasiyeti olarak adlandırıldığı bu modelde mahremiyet koruma işlemi, D veri kümesine rasgeleleştirilmiş bir M mekanizması uygulanması ile sağlanır [37].

ε-Diferansiyel Mahremiyet; bir M mekanizması; her S çıktı kümesi için, herhangi D_1 ve D_2 komşu veri kümeleri üzerinde Eş. 1'deki şartı sağlaması halinde ε-diferansiyel mahremiyeti sağlar;

$$Olasılık[M(D_1) \in S] \leq \exp(\varepsilon) x\, Olasılık[M(D_2 \in S] \quad (1)$$

Bir sorgunun hassasiyeti, seçilen mekanizmanın yapacağı verideki bozma miktarını belirler. Eğer bir f sorgusu D veri kümesine uygulanırsa, f sorgusunun sonucuna eklenecek olan gürültünün miktarı Δf tarafından belirlenir.

Literatürde diferansiyel mahremiyeti sağlamak amacıyla kullanılan Laplace, Eksponansiyel ve Gauss gibi çeşitli mekanizmalar mevcuttur [38]. Bu çalışmada ise literatürde sıklıkla tercih edildiği için Laplace mekanizmasından faydalanılmıştır.

5. TEMEL BİLGİLER (PRELIMINARY)

Bu bölümde, çalışmada kullanılan veri kümesi modellenmiş ve veri faydası metriği açıklanmıştır.

5.1. Veri Modeli (Data Model)

Yörünge verileri genel olarak içerisinde zaman damgası ve konum bilgisi (koordinat) barındıran veriler olarak temsil edilir. Eş 2 ile görülen l yörünge bilgisi x enlem ve y boylam bilgilerinden oluşurken, t ise zaman olmak üzere, T_i yörünge verisi Eş 3'deki gibi modellenebilir.

$$l_i = \{x_i, y_i\} \quad (2)$$

$$T_i = \{(t_1, l_1), (t_2, l_2), \ldots\ldots, (t_n, l_n)\} \quad (3)$$

Yukarıda sunulan veri modelinde tek bir yörünge verisi modellenmiştir. Genel olarak veri kümeleri çok sayıda yörünge verisi içerebileceğinden dolayı böylesi bir T yörünge veri kümesi Eş 4'deki gibi modellenebilir.

$$T = \{T_1, T_2, \ldots\ldots, T_m\} \quad (4)$$

5.2. Veri Faydası Metriği (Data Utility Metric)

Veri faydası, anonim verinin orijinal veriye olan yakınlığı olarak ifade edilir [1]. Bu yakınlık ne kadar fazla olursa, anonim veri üzerinde geliştirilecek olan modelin doğruluğu, orijinal veri üzerinde geliştirilecek olan modelin doğruluğuna o kadar yakın olur. Bir veri kümesinin orijinal halinin maksimum seviyede veri faydası sunduğu kabul edilirken, anonimleştirme sonrası veri kaybı oluşacağı için anonim veri orijinal veriye göre daha düşük seviyede veri faydası sunacaktır.

Literatürde veri faydasını ölçmek için çeşitli metrikler mevcut olup [39-42], bu metrikler genellikle probleme ve algoritmaya uygun olarak seçilir. Bu çalışmada RMSE metriğinden [25] faydalanılmıştır. Örnek sayısının n, hesaplanan değerin $\overline{y}$, gerçek değerin y ile ifade edildiği Eş 5'de RMSE formülü verilmiştir.

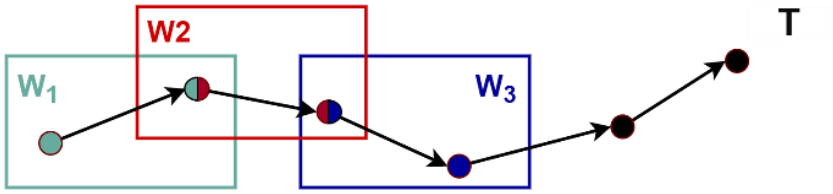
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\overline{y_i} - y_i)^2} \quad (5)$$

6. ÖNERİLEN ANONİMLEŞTİRME MODELİ (PROPOSED ANONYMIZATION MODEL)

Önerilen model, lokal diferansiyel mahremiyeti sağladığı için ilgili model her bir yörünge verisine tek tek uygulanarak anonimleştirme yapılmıştır. Model, yörünge verilerinin pencere tabanlı gruplandırılması, grupların ortalamalarının alınması ve ortalamalara Laplace gürültüsü eklenmesi olmak üzere üç aşamadan oluşmakta olup bunlar aşağıda açıklanmaktadır.

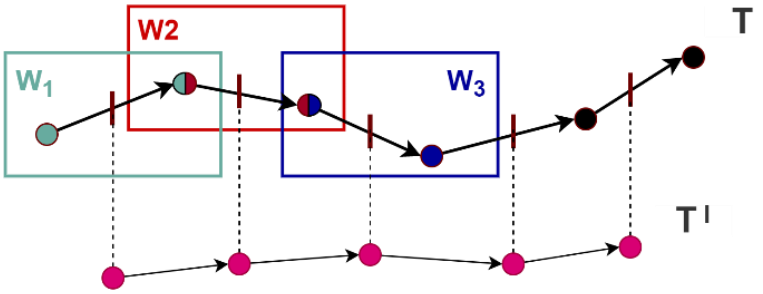
a) Pencere tabanlı gruplama: Önerilen model yörünge verilerini pencere tabanlı olarak ele almaktadır. Pencereden kasıt yörünge verilerinin belirli alanlarda gruplanmasıdır. Pencere Boyu (PB) bir parametre olarak

sisteme verilmektedir. İlgili pencere belirlenen bu boya göre her adımda bir sonraki alana kaydırılarak yeni gruplar oluşturulur. Şekil 3'de örnek bir yörünge verisine pencere yaklaşımı uygulanarak yapılan gruplama işlemi gösterilmektedir. Verilen şekilde, W_i pencere yapısını ve T ise ilgili yörünge verisini temsil etmek üzere; W_1, W_2 ve W_3 pencereleri ve bu pencerelere düşen veri grupları gösterilmektedir.



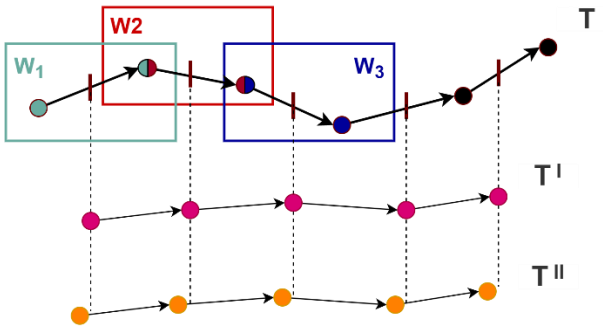
Şekil 3. Yörünge verisinin pencere yapısı ile gruplandırılması (Grouping trajectory data with a window structure)

b) Grupların ortalamasını alma: Pencere tabanlı gruplama ile alt gruplara ayrılan veri kümesine diferansiyel mahremiyet yaklaşımını uygulayabilmek amacıyla, ilgili pencerelere denk gelen verilerin ortalamasının alınarak yeni bir yörünge verisinin (T') oluşturulduğu aşamadır. Şekil 3'de sunulan örnekteki pencerelerin içerisinde bulunan verilerin ortalamaları alınarak T' elde edilmiş ve bu durum Şekil 4'de gösterilmiştir.



Şekil 4. Pencere içindeki konum noktalarının ortalamasını alma ve T' yörünge verisi oluşturma (Averaging location points in the window and constructing T' trajectory data)

c) Ortalamalara gürültü ekleme: Bir önceki aşamada elde edilen T' yörüngesine bu aşamada Laplace gürültüsü eklenerek anonimleştirilmiş yörünge verisi (T'') oluşturulmakta ve bu süreç Şekil 5'de gösterilmektedir.



Şekil 5. T' yörünge verisine gürültü ekleyerek T'' anonim yörünge verisini oluşturma (Constructing anonymous T'' trajectory data by adding noise to T')

Algoritma 1'de önerilen modelin sözde kodu sunulmuştur. İlgili kodda ilk olarak, algoritmaya girdi olarak verilen yörünge veri kümesi pencere tabanlı olarak gruplanır. İkinci olarak, yörünge bazlı elde edilen grupların her birinin kendi içerisinde ortalaması alınarak yeni yörünge verileri oluşturulur. Son olarak ise yeni yörünge verilerine gürültü eklenerek anonim yörünge verileri elde edilir.

Çalışma kapsamında yörünge verilerinin anonim hale getirilmesi için uygulanan genel iş akış süreci Şekil 6'da gösterilmektedir. Başarsoft firmasının çalışma için sağladığı araç verileri bir saatlik Ankara iline ait veriler olup araç kimliği, zaman damgası, enlem ve boylam bilgisi, hız değeri ve açısal yön bilgisi olarak altı adet öznitelikten oluşmaktadır.

Verilerin elde edilme süreci olan ilk aşamada, araçlarda bulunan mobil cihaz verileri veri tabanına metin dosyası şeklinde kaydedilmektedir. Verilerin ön işlemi sürecinde araç verileri öncelikle 32.815056-32.603345 doğu boylam, 39.912720-39.851896 kuzey enlem noktaları arasında ve 10 dakikalık zaman dilimi için filtrelenmiştir. Filtreleme işlemi sonrasında araç verileri harita eşleme işlemine tabi tutulmuş ve sadece Eskişehir yolu üzerinde hareket eden araç verileri elde edilmiştir. Eskişehir yolu üzerinde hareket eden araçlar iki yönlüdür ve aynı yönde ilerleyen araçlar için açısal yön bilgileri 180-270 olan araçlar seçilmiştir. Açısal yön için 0 derece kuzey yönünü göstermekte olup açısal yön bilgileri 180-270 arasındaki araçlar seçildiğinde Eskişehir yolu için doğu-batı yönünde ilerleyen araçlar elde edilecektir ve çalışmada belirtilen yönde ilerleyen araçlar ele alınmıştır. Veri ön işlemi uygulanan araç verilerine üçüncü aşamada önerilen anonimleştirme işlemi uygulanmış ve son aşamada ise anonim verinin sunduğu veri faydası ölçülmüştür.

Algoritma 1. Önerilen anonimleştirme modelinin algoritması (Algorithm of proposed anonymization model)

Girdi: veri kümesi $T = \{T_1, T_2, \dots\dots, T_m\}$, pencere boyu s

Çıktı: anonim veri kümesi $T'' = \{T_1'', T_2'', \dots\dots, T_m''\}$;

Her bir $T_i \in T$ için yinele:

1) Pencere tabanlı grupla:

$$gruplar = \{\{g_1, g_2, \dots, g_m\} = w(T_i, s)\}, g_i = \{(t_x, l_x), (t_y, l_y), \dots\dots, (t_s, l_s)\}$$

2) Ortalama hesapla:

Her bir $g_i \in gruplar$ için yinele:

- Ortalama hesapla; $g_j^{ortalama} = ortalama\_hesapla(g_j)$
- Yeni gruplar kümesine ekle; $gruplar_j^{ortalama} = g_j^{ortalama}$

Dönüştürülen yeni yörünge verisini oluştur; $T_i' = gruplar^{ortalama}$

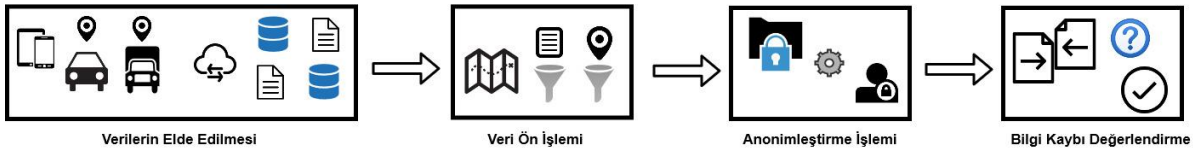
3) Gürültü ekle:

Her bir $g_i \in gruplar$ için yinele:

- Gürültü ekle; $g_j^{ortalama\_laplace} = ortalama\_laplace(g_j)$
- Yeni gruplar kümesine ekle; $gruplar_j^{ortalama\_laplace} = g_j^{ortalama\_laplace}$

Dönüştürülen anonim yörünge verisini oluştur; $T_i'' = gruplar^{ortalama\_laplace}$

Döndür: T'' anonim veri kümesi



Şekil 6. Çalışmada uygulanan genel iş akış adımları (General workflow steps applied in the study)

7. DENEYSEL ÇALIŞMALAR (EXPERIMENTAL STUDIES)

Bu bölümde çalışmanın gerçekleştirildiği ortam, kullanılan veri kümesi ve gerçekleştirilen deneyler ile elde edilen sonuçlar sunulmuştur. Çizelge 1'de görüldüğü üzere, literatürdeki diğer çalışmaların kodları açık olarak yayınlanmadığı, farklı parametrelere göre deneyleri gerçekleştirdikleri ve bu çalışmada önerilen modele yakın bir model bulunmadığı için önerilen modelin test sonuçları diğer çalışmalarla doğrudan karşılaştırılamamıştır. Ancak deneylerde elde edilen test sonuçları, önerilen modelin mahremiyet korumalı yörünge verisi yayınlamada başarıyla kullanılabileceği göstermektedir.

7.1. Geliştirme Ortamı (Development Environment)

Deneysel çalışmalar, Python dilinde Jupyter Notebook ortamında gerçekleştirilmiştir. Pandas kütüphanesi ile veri seti ortama dahil edilmiş ve Numpy kütüphanesi ile matematiksel işlemler yapılmıştır. Matplotlib kütüphanesi ve OpenStreetMap aracı ile görselleştirme işlemleri gerçekleştirilmiştir. Tüm deneyler 16 GB ram, i7 işlemci ve Windows işletim sistemine sahip bilgisayar üzerinde gerçekleştirilmiş ve sonuçlar alınmıştır.

7.2. Kullanılan Veri Kümesi (Dataset)

Yapılan çalışmada, Başarsoft Bilgi Teknolojileri A.Ş. tarafından sağlanan bir saatlik, Türkiye sınırları içerisinde yer alan filo araçlarına ait GPS verileri kullanılmıştır. Toplamda yaklaşık 3 milyon 800 bin satır veriden oluşan bir saatlik veri seti, hareketli araçlara ait nitelik bilgilerini içermektedir. Araçlara ait nitelikler; araçlara özgü ID, GPS sinyalinin alındığı zaman, araç yörünge bilgisi olan enlem ve boylam, araçların anlık hız ve açısal yön bilgileridir. Veri kümesi bir önceki bölümde anlatıldığı şekilde filtreleme ve harita eşleme ön işlemlerine tutulduğunda tek yönlü araç verileri 10 dakika için yaklaşık 5 bin satıra indirgenmiştir. Böylelikle indirgeme işlemi sonucunda belirlenen yol üzerinde oluşan 477 adet araç yörüngesi üzerinde çalışılmıştır. Filtrelenen ve Eskişehir Yolu'na indirgenen veri kümesi hassas bilgiler barındırdığından dolayı çizelge halinde gösterilmekten ziyade şekilsel olarak gösterilmesi uygun görülmüş ve Şekil 7'de görselleştirilmiştir.

Şekil 7. Çalışmada kullanılan Eskişehir yolundaki araçlara ait yörünge veri kümesi (Used trajectory data set of vehicles on Eskişehir Road)

7.3. Deneyler (Experiments)

Deneysel çalışmalarda, farklı PB ve ε parametrelerine göre testler yapılmış ve sonuçlar RMSE metriği kullanılarak ölçülmüştür. PB değerleri literatürde standart bir değer olmadığı için sezgisel olarak seçilmiş ve sadece 3 farklı değere göre deneylerin yapılması yeterli görülmüştür. Bu kapsamda, veri kümesinden rasgele seçilen örnek bir araç için ilk deney gerçekleştirilmiş ve ilgili test sonuçları alınarak grafiksel olarak gösterilmiştir. İkinci deneyde ise veri kümesindeki tüm araçlar için ilgili testler yapılmış ve elde edilen değerler sunulmuştur.

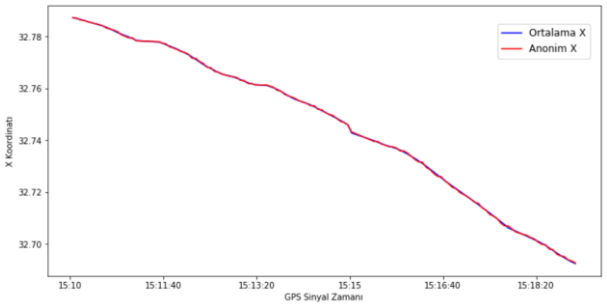
a) Deney 1 - Örnek araç için test sonuçları:

Bu deneyde veri kümesinden rasgele bir araç yörünge verisi seçilerek testler yapılmış ve sonuçları Çizelge 2'de sunulmuştur. Rasgele seçilen bu aracın belirlenen zaman aralığında ürettiği 123 GPS sinyali mevcuttur. Çizelge 2 incelendiğinde, PB arttıkça X ve Y değerlerinde elde edilen RMSE değerlerinin de arttığı görülmüştür. Ayrıca her bir PB içerisinde ε değeri arttıkça RMSE değerinde genel olarak azalma meydana geldiği görülmektedir. Çizelge 2'de sunulan sonuçlara göre, bu deneyde en az hata değerini PB=2 ve ε=2 iken elde edildiği gözlemlenmiştir.

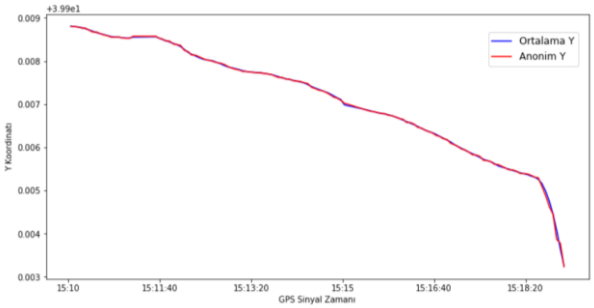
Örnek aracın ortalama ile anonimleştirilmiş X ve Y değerlerinin grafikleri sırasıyla Şekil 8 ve Şekil 9'da çizdirilmiştir. Çizelge 2'de verilen hata değerleri de dikkate alındığında her iki grafikte de araç hareket grafiklerinin birbirine çok yakın olduğu görülmektedir. Şekil 10'da ise aynı örnek aracın X, Y ve GPS sinyal zamanı eksenlerinde hareketinin orijinali ile PB=2 ve ε=2 için anonimleştirilmiş versiyonu gösterilmiştir.

Çizelge 2. Örnek araç için elde edilen deney sonuçları (The obtained experimental results for the sample vehicle)

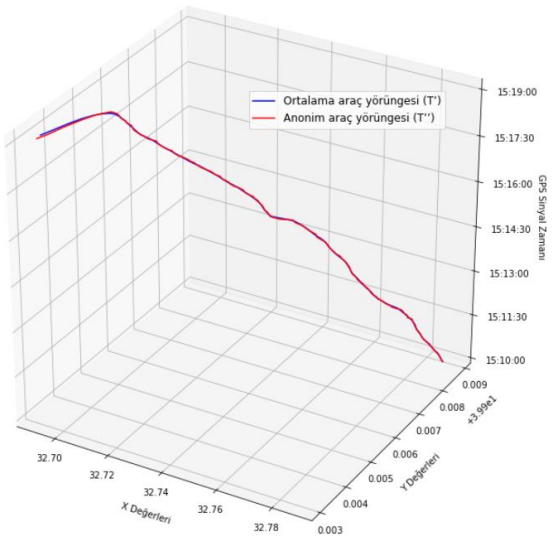
PB	ε	RMSE değeri	
		X	Y
2	0,5	0,00042509	0,00000423
	1,0	0,00041723	0,00000386
	1,5	0,00037905	0,00000256
	2,0	0,00037026	0,00000301
3	0,5	0,00081843	0,00000772
	1,0	0,00071983	0,00000618
	1,5	0,00071359	0,00000579
	2,0	0,00064440	0,00000585
4	0,5	0,00113142	0,00011389
	1,0	0,00102258	0,00011024
	1,5	0,00103406	0,00010254
	2,0	0,00087512	0,00000899



Şekil 8. Örnek aracın ε=2 için ortalama ve anonimleştirilmiş X değerleri (Average and anonymized X values of the sample vehicle for ε=2)



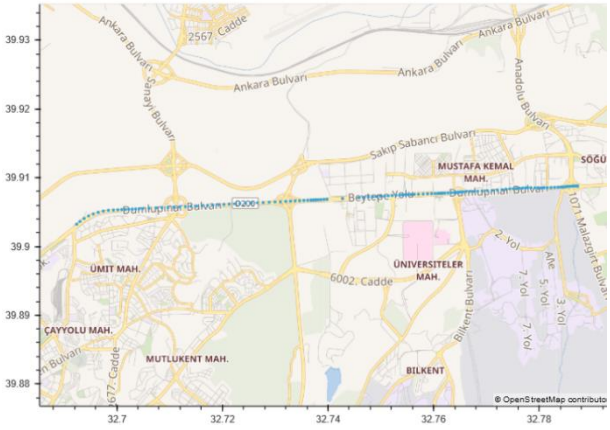
Şekil 9. Örnek aracın ε=2 için ortalama ve anonimleştirilmiş Y değerleri (Average and anonymized Y values of the sample vehicle for ε=2)



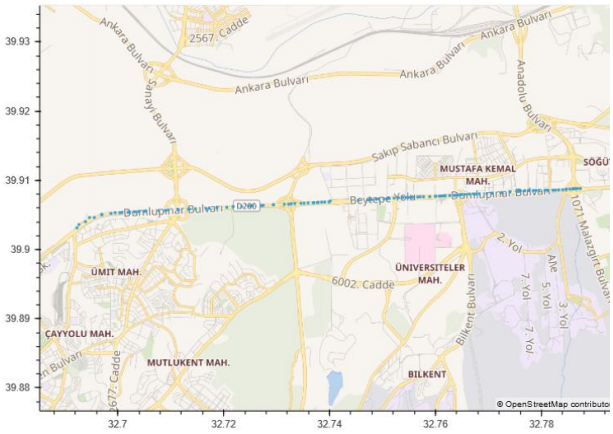
Şekil 10. Örnek aracın X, Y ve zaman boyutunda GPS sinyal grafiği (The GPS signal graph of sample vehicle on X, Y and time dimension)

Şekil 11'de örnek aracın orijinal yörünge verisi, Şekil 12'de ortalama alınmış yörünge verisi ve Şekil 13'de ise diferansiyel mahremiyet uygulanmış yörünge verisi versiyonları harita üzerinde gösterilmiştir.

Şekil 11. Örnek aracın orijinal yörünge verisi ($T_{örnek}$) (Original trajectory data of the sample vehicle (T_{sample}))



Şekil 12. Örnek aracın ortalama alınmış yörünge verisi ($T'_{örnek}$) (Averaged trajectory data of the sample vehicle (T'_{sample}))



Şekil 13. Örnek aracın diferansiyel mahremiyet uygulanmış yörünge verisi ($T''_{örnek}$) (Differentially private trajectory data of the sample vehicle (T''_{sample}))

b) Deney 2 - Veri kümesindeki tüm araçlar için elde edilen sonuçlar:

Bu deneyde veri kümesindeki tüm araçlar için yapılan testler ve elde edilen sonuçlar Çizelge 3'de sunulmuştur. Çizelge 3 incelendiğinde, PB arttıkça X ve Y değerlerinde elde edilen RMSE değerlerinin arttığı, her bir PB özelinde ε değeri arttıkça RMSE değerinde genel olarak azalma meydana geldiği görülmektedir. Elde edilen bu sonuçlara göre, bu deneyde de en az hata değerini PB=2 ve ε=2 iken elde edildiği gözlemlenmiştir. Hem Deney 1'de hem de Deney 2'de, hata değerlerin en küçük PB ve en büyük ε değerinde alındığı görülmektedir. PB=4 ve ε=0,5 değeri en büyük hata oranına sahip olsa da konum verileri kabul edilebilir ve kullanılabilir seviyededir.

Çizelge 3. Tüm araçlar için elde edilen ortalama RMSE değerleri (The obtained average RMSE values for all vehicles)

PB	ε	Ortalama RMSE değeri	
		X	Y
2	0,5	0,00099361	0,00032929
	1,0	0,00092869	0,00031752
	1,5	0,00086412	0,00028746
	2,0	0,00079487	0,00027582
3	0,5	0,00185356	0,00062133
	1,0	0,00170812	0,00058045
	1,5	0,00158801	0,00055717
	2,0	0,00142386	0,00050013
4	0,5	0,00263162	0,00089442
	1,0	0,00246601	0,00083682
	1,5	0,00222765	0,00080369
	2,0	0,00201050	0,00071273

8. SONUÇLAR VE DEĞERLENDİRME (RESULTS AND EVALUATIONS)

Veri mahremiyeti, KTS'lerde ele alınması gereken ve kişisel verilerin korunması için son derece önemli bir konudur. Konum veya yörünge verileri gerek KVKK gerekse de GDPR ile güvence altına alınmış olsa da bu tür verilerin yayınlanırken veya paylaşılırken mahremiyete önem verilmesi ve koruyucu önlemlere tabi tutulması gerekir. Bu çalışmada, konum ve yörünge verilerinin mahremiyetinin korunarak yayınlanması için yeni bir model önerilmiş, önerilen model gerçek bir veri kümesi üzerinde uygulanmış ve başarıyla test edilmiştir. Önerilen modelde araç verileri öncelikle ön işlemden geçirilmiş ve daha sonra da anonimleştirme işlemine tabi tutulmuştur. Anonimleştirme için araç yörünge verileri pencere tabanlı bir yaklaşımla ele alınarak diferansiyel mahremiyet yöntemi uygulanmış ve RMSE metriği kullanılarak veri faydası ölçülmüştür.

Elde edilen bulgular aşağıda verilen maddelerde değerlendirilmiştir.

- PB=2 ve ε=2 için yapılan testlerde hem örnek araç hem de tüm araçların X ve Y koordinatları için 0.001'den küçük hata değerleri elde edilmiştir.
- Örnek araç yörünge verileri haritalar üzerinde görselleştirilmiştir.
- Anonim yörünge verisinin (T'') dönüştürülmüş yörünge verisine (T') göre çok yakın olduğu Şekil 12 ve Şekil 13'de gösterilmiştir.
- Hata oranları ve harita üzerindeki konum noktaları incelendiğinde, anonimleştirilmiş yörünge verisinin anlam bütünlüğünün korunduğu, bu verinin halen kullanılabilir olduğu gerçek zamanlı olarak görülmektedir.
- Özellikle araçlardan elde edilen GPS sinyallerinin bundan sonraki süreçte veri toplama sıklığının artırılması ve saha fazla verinin toplanması durumunda, önerilen modelin veri faydasının daha yüksek olacağı değerlendirilmektedir.
- Çizelge 1'de sunulan literatür çalışmaları incelendiğinde, çalışmaların kodlarının açık olarak yayınlanmadığı, farklı parametrelere göre deneyleri gerçekleştirdikleri, gerçek zamanlı verileri kullanmadıkları ve en önemlisi ise bu çalışmada önerilen modele yakın bir model önerisinin bulunmadığı görülmüştür.
- Önerilen modelde literatür kısmında sunulan modellerden farklı olarak lokal seviyede diferansiyel mahremiyeti uygulaması ve kayıt temelli işlem yapmasıyla, pencere boyutuna göre ortalama alarak orijinal verinin anlam bütünlüğünü koruması açısından diğer modellerden ayrışmaktadır.
- Bu çalışma kapsamında önerilen model, lokal diferansiyel mahremiyet uygulamasından dolayı çizelgede sunulan diğer örneklerle karşılaştırılamamış olsa da test sonuçları başarımın yüksek olduğunu göstermektedir.
- Önerilen model, lokal diferansiyel mahremiyeti uygulamasıyla, özellikle bu tür uygulamalar araçtan doğrudan veri toplama aşamasında kullanılabilir ve böylelikle daha veri toplama aşamasında yüksek seviyede mahremiyet sağlanabilir.
- Yapılan deneysel çalışmalarda gerçek veri kümesi kullanılmıştır. Dolayısıyla önerilen modelin çalışan gerçek sistemlere kolayca adapte edilerek kullanılması rahatlıkla sağlanabilir.
- Önerilen modelin sunduğu veri faydası farklı metriklerle de değerlendirilebilir. Ancak bu çalışmada literatürde sıklıkla kullanılan RMSE metriğinden faydalanılması yeterli görülmüştür.

Sonuç olarak bu çalışma kapsamında sunulan modelin; PB ve ε parametrelerinin değişmesine rağmen elde edilen hata payının verinin kullanılabilirliğini makul seviyelerde değiştirdiğini, böylelikle önerilen modelin yörünge verilerinin mümkün olduğu kadar anlamsal bütünlüğünü koruyarak mahremiyetini sağladığını ortaya koymuştur. Ayrıca, bu çalışmada önerilen modelin bir TÜBİTAK proje kapsamında gerçek bir ürünün geliştirilmesinde kullanılacak olması ise bu çalışmanın diğer bir önemli katkısı olacaktır. Burada geliştirilen modelin, farklı uygulamalarda da kullanılabileceği değerlendirilmektedir.

TEŞEKKÜR (ACKNOWLEDGEMENT)

TÜBİTAK tarafından desteklenen 3191873 numaralı proje kapsamında yapılan bu çalışmada yazarlar; başta sağladığı destekler için TÜBİTAK'a, Başarsoft Bilgi Teknolojileri A.Ş.'ye ve teknolojik altyapı sunan Gazi BIDISEC'e teşekkür ederler.

ETİK STANDARTLARIN BEYANI (DECLARATION OF ETHICAL STANDARDS)

Bu makalenin yazar(lar)ı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

YAZARLARIN KATKILARI (AUTHORS' CONTRIBUTIONS)

Murat AKIN: Makaleyi yazmış, deneyleri yapmış ve sonuçlarını analiz etmiştir.

Yavuz CANBAY: Makaleyi yazmış, deneyleri yapmış ve sonuçlarını analiz etmiştir.

Şeref SAĞIROĞLU: Problemi tanımlamış ve sonuçları analiz etmiştir.

ÇIKAR ÇATIŞMASI (CONFLICT OF INTEREST**)**

Bu çalışmada herhangi bir çıkar çatışması yoktur.

KAYNAKLAR (REFERENCES)

[1] Fung B. C., Wang K., Fu A. W. and Philip S. Y., "Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques". ***CRC Press***, (2010).

[2] Liu X. and Zhu Y., "Privacy and Utility Preserving Trajectory Data Publishing for Intelligent Transportation Systems," ***IEEE Access***, 8, 176454-176466, (2020).

[3] Warren S. D. and Brandeis L. D., "The Right to Privacy," ***Harvard Law Review,*** 193-220, (1890).

[4] Jain P., Gyanchandani M., and Khare N., "Big Data Privacy: A Technological Perspective and Review," ***Journal of Big Data***,3(1): 25, (2016).

[5] De Capitani Di Vimercati S., Foresti S., Livraga G., and Samarati P., "Data Privacy: Definitions and Techniques," ***International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems***, 20(6): 793-817, (2012).

[6] İnternet: "Kişisel Verilerin Korunması Kanunu." Bakanlar Kurulu. http://www.resmigazete.gov.tr/eskiler/2016/04/20160407-8.pdf (11.09.2020).

[7] Abul O., Bonchi F., and Nanni M., "Never walk alone: Uncertainty for anonymity in moving objects databases," in ***International conference on data engineering***, 376-385, (2008).

[8] Wang Y., Xia Y., Hou J., Gao S.-m., Nie X., and Wang Q., "A fast privacy-preserving framework for continuous location-based queries in road networks," ***Journal of Network and Computer Applications***,53, 57-73, (2015).

[9] Dwork C., "Differential Privacy," ***International Colloquium on Automata, Languages and Programming,*** 1-12, (2006).

[10] Ren W. and Tang S., "EGeoIndis: An effective and efficient location privacy protection framework in traffic density detection," ***Vehicular Communications***, 21,100187, (2020).

[11] Zhang G., "A differentially private data aggregation method based on worker partition and location obfuscation for mobile crowdsensing," ***Computers, Materials & Continua***, 63(1): 223-241, (2020).

[12] Liu L., "From data privacy to location privacy: models and algorithms," I***nternational conference on Very large data bases***, Vienna, Austria, (2007).

[13] Hoh B., Gruteser M., Xiong H., and Alrabady A., "Preserving privacy in gps traces via uncertainty-aware path cloaking," ***Conference on Computer and communications security***, 161-171, (2007).

[14] Li M., Zhu L., Zhang Z., and Xu R., "Achieving differential privacy of trajectory data publishing in participatory sensing," ***Information Sciences***, 400, 1-13, (2017).

[15] Chen R., Fung B., and Desai B. C., "Differentially private trajectory data publication," ***arXiv preprint arXiv:1112.2020***, (2011).

[16] Han Q., Xiong Z., and Zhang K., "Research on trajectory data releasing method via differential privacy based on spatial partition," ***Security and Communication Networks,*** 2018, (2018).

[17] He X., Cormode G., Machanavajjhala A., Procopiuc C. M., and Srivastava D., "DPT: differentially private trajectory synthesis using hierarchical reference systems," ***VLDB Endowment***, 8(11):1154-1165, (2015).

[18] Gursoy M. E., Liu L., Truex S., and Yu L., "Differentially private and utility preserving publication of trajectory data," ***IEEE Transactions on Mobile Computing***, 18(10)2315-2329, (2018).

[19] Cao Y. and Yoshikawa M., "Differentially private real-time data release over infinite trajectory streams," in ***IEEE International Conference on Mobile Data Management***, 2, 68-73, (2015).

[20] Tian F., Zhang S., Lu L., Liu H., and Gui X., "A novel personalized differential privacy mechanism for trajectory data publication," in ***International Conference on Networking and Network Applications***, 61-68, (2017).

[21] Zhao X., Dong Y., and Pi D., "Novel trajectory data publishing method under differential privacy," ***Expert Systems with Applications***, 138,112791, (2019).

[22] Zhao J., Mei J., Matwin S., Su Y., and Yang Y., "Risk-Aware Individual Trajectory Data Publishing with Differential Privacy," ***IEEE Access,*** (2020).

[23] Jiang K., Shao D., Bressan S., Kister T., and Tan K.-L., "Publishing trajectories with differential privacy guarantees," in ***International Conference on Scientific and Statistical Database Management,*** 1-12, (2013).

[24] Han Q., Lu D., Zhang K., Du X., and Guizani M., "Lclean: a plausible approach to individual trajectory data sanitization," ***IEEE Access***,6, 30110-30116, (2018).

[25] Singh K., Rong J., and Batten L., "Sharing sensitive medical data sets for research purposes-a case study," in ***International Conference on Data Science and Advanced Analytics***, 555-562, (2014).

[26] Xie H., Kulik L., and Tanin E., "Privacy-aware collection of aggregate spatial data," ***Data & Knowledge Engineering,*** 70(6):576-595, (2011).

[27] Chen B., LeFevre K., and Ramakrishnan R., "Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge," in ***International Conference on Very Large Data Bases***, Vienna, Austria, 770-781, (2007).

[28] Sweeney L., "Computational Disclosure Control: A Primer on Data Privacy Protection," Ph. D. Thesis, ***Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology***, USA, (2001).

[29] Machanavajjhala A., Gehrke J., Kifer D., and Venkitasubramaniam M., "l-Diversity: Privacy Beyond k-Anonymity," ***International Conference on Data Engineering,*** Atlanta, USA, (2006).

[30] Nergiz M. E., Atzori M., and Clifton C., "Hiding the Presence of Individuals from Shared

Databases," in ***International Conference on Management of Data***, Beijing, China, 665-676, (2007).

[31] Wang Y., Xia Y., Hou J., Gao S. M., Nie X., and Wang Q., "A fast privacy-preserving framework for continuous location-based queries in road networks," ***J Netw Comput Appl***, 53,57-73, (2015).

[32] Sweeney L., "k-Anonymity: A Model for Protecting Privacy," ***International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems***,10(5):557-570, (2002).

[33] Kenig B. and Tassa T., "A practical approximation algorithm for optimal k-anonymity," ***Data Mining and Knowledge Discovery***, 25,(1):134-168, (2012).

[34] Meyerson A. and Williams R., "On the Complexity of Optimal k-Anonymity," in ***Symposium on Principles of Database Systems***, Paris, France, 223-228, (2004).

[35] Aggarwal G. *et al.*, "Approximation Algorithms for k-Anonymity," ***Journal of Privacy Technology,*** 1-18, (2005).

[36] Aggarwal G. *et al.*, "Anonymizing Tables," in ***International Conference on Database Theory***, Edinburgh, UK, 246-258, (2005).

[37] Zhu T., Li G., Zhou W., and Philip S. Y., "Differentially private data publishing and analysis: A survey," ***IEEE Transactions on Knowledge and Data Engineering***, 29(8):1619-1638, (2017).

[38] Canbay Y. and Sağıroğlu Ş., "Derin Öğrenmede Diferansiyel Mahremiyet," ***Uluslararası Bilgi Güvenliği Mühendisliği Dergisi***,6(1):1-16, (2020).

[39] Samarati P., "Protecting Respondents Identities in Microdata Release," ***IEEE Transactions on Knowledge and Data Engineering***, 13(6):1010-1027, (2001).

[40] LeFevre K., DeWitt D., and Ramakrishnan R., "Mondrian Multidimensional k-Anonymity," in ***International Conference on Data Engineering,*** Atlanta, USA, 25-25, (2006).

[41] Skowron A. and Rauszer C., "The Discernibility Matrices and Functions in Information Systems," in ***Intelligent Decision Support,*** 331-362, (1992).

[42] Ghinita G., Karras P., Kalnis P., and Mamoulis N., "Fast Data Anonymization with Low Information Loss," in ***International Conference on Very Large Databases***, Vienna, Austria, 758-769, (2007).