Protecting Privacy of Semantic Trajectory

by

Roya Shourouni

M.Sc, Khaje Nasir Toosi University, 2015

B.Sc., Khaje Nasir Toosi University, 2012

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Geography
Faculty of Environment

SIMON FRASER UNIVERSITY
Summer 2021

Declaration of Committee

Name:	**Roya Shourouni**
Degree:	**Master of Science (Geography)**
Title:	**Protecting Privacy of Semantic Trajectory**
Committee:	**Chair:** **Suzana Dragicevic** Professor, Geography **Nadine Schuurman** Supervisor Professor, Geography **Luke Bergmann** Committee Member Associate Professor Department of Geography University of British Columbia **Martin Andresen** Examiner Professor, Criminology

Abstract

The growing ubiquity of GPS-enabled devices in everyday life has made large-scale collection of trajectories feasible, providing ever-growing opportunities for human movement analysis. However, publishing this vulnerable data is accompanied by increasing concerns about individuals' geoprivacy. This thesis has two objectives: (1) propose a privacy protection framework for semantic trajectories and (2) develop a Python toolbox in ArcGIS Pro environment for non-expert users to enable them to anonymize trajectory data. The former aims to prevent users' re-identification when knowing the important locations or any random spatiotemporal points of users by swapping their important locations to new locations with the same semantics and unlinking the users from their trajectories. This is accomplished by converting GPS points into sequences of visited meaningful locations and moves and integrating several anonymization techniques. The second component of this thesis implements privacy protection in a way that even users without deep knowledge of anonymization and coding skills can anonymize their data by offering an all-in-one toolbox. By proposing and implementing this framework and toolbox, we hope that trajectory privacy is better protected in research.

Keywords: Individuals' geoprivacy; semantic trajectory; location swapping; unlinkability; ArcGIS python toolbox

Acknowledgments and Dedication

This thesis is dedicated to my beloved parents. Although far from me, they always supported me through every academic undertaking and believed in me throughout. Thank you for not just your love and support, but your constant inspiration

I would like to acknowledge my committee member, Dr. Luke Bergmann. I am especially grateful for the support of my supervisor, Dr. Nadine Schuurman. This thesis would not be possible without the help and support of my supervisor, Dr. Nadine Schuurman. She has shaped not only my thesis but has also taught me the foundations of being a strong and dedicated academic. Thank you for walking me through the various lessons and experiences of this degree.

Table of Contents

List of Tables

List of Figures

List of Acronyms

GPS	Geographic Positioning System
LBS	Location-Based Service
MFL	Most Frequented Location
POI	Point of Interest
PPDM	Privacy Preserving Data Mining
PPDP	Privacy Preserving Data Publishing
QI	Quasi Identifier
SA	Sensitive Attribute
SKA	Spatial K-anonymity

Chapter 1. Introduction

The vast majority of adults carry mobile devices today (The World Bank, 2016) and their locations are continuously recorded by network providers based on their nearest cellular towers (Ahas et al., 2015). One-third of the world's population is using social media (Ali et al., 2018) mainly through a mobile device. With vast numbers of people in possession of GPS-enabled mobile devices, and interested in location collection and sharing, location has become one of the main foci of many services, applications, and technologies, allowing the location to become a context item strongly affiliated with individuals' identities. Collecting such valuable data has provided new opportunities for mobility-related decision-making (Cáceres et al., 2019; Hwang et al., 2013; Rudenko et al., 2019; Zheng & Zhou, 2011). This ability, however, has also ignited concern about privacy violations due to the inherent sensitivity of the spatial data being gathered (Keßler & McKenzie, 2018; Leszczynski, 2017). An example of this vulnerability is the publication of a Strava global heatmap of location data which accidentally revealed sensitive information about the United States military soldiers on active services resulted in disclosure of United States secret military bases that were abroad (Hern, 2018). This is just one of the many instances where geoprivacy of individuals has been breached, even from aggregated location data.

Another more common form of individuals' geoprivacy breach occurs when publishing person-specific movement data in the form of trajectory micro-data since these trajectories can reveal private information of individuals such as where they live, where they work, and what Points of Interest (PoIs) they frequently visit which can show their habits and sensitive information such as health conditions (e.g., frequent visits to a medical center), daily activities and routines. For example, Hasan et al. (2017) highlighted that publishing trajectories by bike-sharing companies even after removing users' identities can disclose their privacy with their spatial behaviors and movement patterns. Keßler and McKenzie (2018) demonstrated how people often share their location coupled with other personal private information, such as credit card or license plate numbers, with a large number of services in their daily life. Publishing such vulnerable personal data can cause significant risks if it is used for malicious purposes such as unsolicited advertisements, scams, or physical violence (Krumm, 2009).

As a result of these privacy concerns, some regulation efforts and frameworks have been made which details key considerations and obligations with regard to privacy protection—for example, the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans in Canada, the Belmont Report, and Common Rule in the United States (CIHR, NSERC, & SSHRC, 2014; Department of Health and Human Services, 2018), the General Data Protection Regulation (GDPR) for the European Union (Radley-Gardner et al., 2016). There has been a variety of similar policies and regulations published in Europe, the United States, Japan, and Australia, and by the Asia-Pacific Economic Cooperation (APEC). Nevertheless, the focus of these policies is not spatial data privacy or geoprivacy, and therefore they require interpretation when being applied to spatial data. Extending the concept of privacy to spatial data recognizes that this constitutes "a special type of information privacy which expresses the individual rights to determine when, how, and to what extent their spatial information is shared with others" (Duckham & Kulik, 2006). In the context of research, developing policies that explicitly address spatial data would emphasize the sensitivity of participants' spatial information and the need to protect it in ways that will prevent the inference of their identity, true locations, and other personal information.

So far, numerous privacy protection techniques, also known as anonymization, have been proposed and developed with the goal of offering advances over previous efforts to preserve geoprivacy while maintaining data utility. The paradigm of creating a balance between individuals' privacy and the utility of a dataset at data publishing time is called Privacy-Preserving Data publishing (PPDP). Ghinita (2009) identified two relevant scenarios where geoprivacy can be protected at data publishing time. The first scenario is in the context of online location-based service (LBS), where privacy must be protected in real-time when providing on-demand service based on the user's query. In LBS, both location privacy and query privacy must be preserved to prevent the service provider from inferring the locations of the users (Shin et al., 2012). The second scenario, namely that of offline and data-centric privacy protection, occurs when the entire trajectory dataset needs to be published at once in an offline format without imperiling the privacy of the individuals. Privacy protection approaches at offline data publishing are not, in general, applicable for the context of LBS and vice versa (Abul et al., 2010). While both scenarios, and the challenges they pose, are equally important, there have been fewer efforts in the context of offline data publishing, especially in the context of trajectory data.

In this thesis, the focus is on understanding the current state of offline trajectory publication in academic research, the related concepts and privacy concerns, and the gaps in current privacy protection methods. With that background established, this thesis proposes a new framework for the enhanced privacy protection of trajectories and their utility and details the development of a set of tools meant to enable that protection. What follows in this introductory section is a brief overview of related conceptual definitions and privacy criteria, and lastly, an overview of the thesis structure.

1.1. Conceptual definitions

What is a trajectory and how it could be represented from a human point of view? What are the potential privacy threats when publishing trajectories and how their privacy can be protected? To answer these questions, we begin by exploring the concepts of the terms and these terms are ordered by their relationships.

Trajectory dataset: A trajectory dataset contains the mobility history of moving objects which can be collected either by each moving object for a certain period of time, or centrally by a server that can track the location of query issuers in real-time (Gambs et al., 2010).

Trajectory: A trajectory is a temporally ordered sequence of timestamped spatial points which can be a numerical location (geographic coordinate) or a meaningful location such as an address or a semantic label (e.g., “home” or “work”). The timestamp can be the exact date and time, a time interval (e.g., from 8:00-9:00 am), or a period of day (e.g., morning). Each trajectory is characterized by a unique userID, which can be the real identifier of the moving object (e.g., “Tom”), a pseudonym (anonymous username), or even the value “unknown”. Different trajectories recorded by the same moving object have the same pseudonym so that a pseudonym is still able to link different actions performed by the same moving object (Gambs et al., 2010).

Semantic trajectory: A semantic trajectory is a temporally ordered sequence of visited meaningful locations, referred to as semantic stops, which represent activities at stop time and are connected by moves that represent speed distribution, and acceleration patterns (Alvares et al., 2007; Baglioni et al., 2009; Palma et al., 2008). Visited meaningful locations correspond to the set of points of a trajectory within a specific spatiotemporal window that are important depending on the specific application. A spatiotemporal window can be defined for every type of location category or function.

Trajectory data publishing: Data publishing is the process of publishing a trajectory dataset by a data owner that could be an organization to a data recipient that could be public or researchers to allow them to do significant research on the published dataset. It is important not to publish trajectories directly since they contain the identity and personal information of individuals.

Quasi-identifier: A quasi-identifier (QI) is an attribute that cannot explicitly reveal a user’s identity, but coupled together or combined with external information can potentially re-identify the user (Samarati & Sweeney, 1998) and expose their sensitive attributes (SAs) (e.g., income, religion, and health status) (N. Li et al., 2007; Machanavajjhala et al., 2006). To adopt the notion of QI and SA in the context of semantic trajectory, one can classify point samples of a trajectory into QIs and SAs, where QIs can be any random point and SAs can relate to personally important places (e.g., home/workplaces) or semantically sensitive places (e.g., hospital, church) (Sila-Nowicka & Thakuriah, 2016).

Attacker: In the data publishing process, the data recipient could be an attacker. An attacker is an unauthorized entity who gains access to the published dataset and links it with their background knowledge to further expand their knowledge about a target user or all users whose trajectories are contained within the dataset. Background knowledge is the extra information about the user(s) that is achieved from external resources apart from the published dataset (Narayanan & Shmatikov, 2006).

Attack: an attack is a process of inferring SAs of a user or all users carried out by a potential attacker who has background knowledge about a set of QIs of the user(s) (Krumm, 2007). The purpose of an attack can be inferring the identity of the user(s), their SAs (such as sensitive locations visited by the user(s)), and/or any new personal information about the user(s) which is implicitly present in the dataset. There are three main types of attack when publishing a trajectory dataset: record linkage attack, attribute linkage attack, and probabilistic attack (Fung et al., 2010). In a record linkage attack, also known as a re-identification attack, given a published trajectory dataset where each trajectory is uniquely associated with an anonymous user, attackers try to match their background knowledge to the corresponding trajectory of the target user(s) to infer the users' identities (Cecaj et al., 2016; de Montjoye et al., 2013; Ma et al., 2013; Rossi et al., 2015). The goal of an attribute linkage attack, also known as homogeneity attack, is inferring SAs of the user(s) without uniquely re-identifying the user(s) in a dataset. In this case, known QIs can be linked to multiple users, but all users have the same SAs (such as visiting from mosques, hospitals, strip clubs,…) (Sui et al., 2016; Tu et al., 2019). The objective of a probabilistic attack is further expanding the knowledge of an attacker about a user (e.g., any part of the user's trajectory) after having access to a published dataset, rather than linking background knowledge to the user's identity or SAs (Dwork, 2006; Gramaglia & Fiore, 2015).

Privacy-Preserving data publishing: PPDP is a process of transforming data before data publishing to preserve the identities, SAs, and/or any further information about users while providing measurable benefits to data recipients for data analysis.

1.2. Privacy criteria

A key purpose of privacy protection is limiting the disclosure of users' identities and SAs from an anonymized dataset (Skinner & Elliot, 2002). However, there is no universally acceptable level of privacy or identity disclosure risk for a given dataset yet. The removal of explicit identifiers (e.g., user name, SIN) is not a solution since QIs can be combined to explicitly disclose users' identities (Samarati & Sweeney, 1998). To limit the identity disclosure risk of objects, a privacy criterion named "k-anonymity" was introduced by Samarati & Sweeney, (1998). It was initially designed for relational datasets to make each record showing user information in a table indistinguishable from at least k-1 other records/users on one or a set of QIs. Thus, an attacker without background knowledge from external sources only has the probability of 1/*k* to identify a target user in a dataset, by which the user's real identity could be preserved. A group of k users that have the same QIs after anonymization is called an anonymity set. This concept was soon adopted for a spatial dataset (Kalnis & Ghinita, 2009), referred to as spatial k-anonymity (SKA) as a metric to assess the re-identification risk of users' locations in a dataset. The value of k shows the number of potential locations that could be corresponded to the true location of a user in an anonymity set.

k-anonymity can prevent the re-identification attack, although it is not resistant against attribute linkage attack or simply attribute attack. Attribute attack happens when the identity of users or their true locations are not precisely revealed, but their SAs (such as religion or health status) can be inferred from anonymity sets (N. Li et al., 2007; Machanavajjhala et al., 2006). For example, in the context of spatial data, anonymization techniques adhere to SKA may create a spatial region that includes k locations, but all have one sensitive semantic function (e.g., hospital) and as a result, the sensitive information of users can be identified without uniquely re-identifying the user. Further, Wong et al., (2009) improved the concept of k-anonymity by proposing (α,k)-anonymity to upper bound the confidence of inferring SAs from any k-anonymity set to a certain threshold α. l-diversity was proposed as a stronger version of k-anonymity to ensure each anonymity set has at least l diverse values for each SA (Machanavajjhala et al., 2006). In the context of spatial data, location-diversity (Xue et al., 2009) was introduced to ensure each anonymity set must include at least l semantically diverse locations. Li et al., (2007a) showed that l-diversity is neither required nor sufficient to prevent attribute attack. They proposed the t-closeness as a stricter criterion that stipulates that the distribution of SAs in any anonymity set is close to the overall dataset's, where closeness is limited by the threshold *t*. Although the increase of the value of location diversity or t-closeness can effectively prevent the disclosure of SAs within an anonymity region, they cannot maintain the semantic consistency of trajectory to its original value.

All aforementioned privacy criteria try to protect users' identities and SAs in a released dataset and prevent re-identification and attribute attacks. However, none of them investigates how the probability of privacy breach differs based on the presence and absence of one or a set of the user's data, which leads to a probability attack. Dwork (2006) defined a more robust privacy model, named "differential privacy", to determine how the privacy risk of an individual and the result of an analysis on a dataset differ with and without the existence of a single element of a user in a dataset. Differential privacy ensures that any change of a single element in a dataset does not significantly affect the result of statistical analysis and does not allow an attacker to achieve more knowledge from the dataset, regardless of their background knowledge. Therefore, it means that any privacy attack will not be a result of a user's participation in the dataset. The basic idea is to add random or bounded noise and merge trajectories and return a sampled result from the set of all possible ones. Noise can be added either to each coordinate of a GPS point or to the whole trajectory (you may refer to Dwork’s work for more details). Differential-privacy publishes a set/cluster of merged trajectories with their counts (noisy counts of trajectories) as a result.

Although finding an optimal k-anonymization is NP-hard (Meyerson & Williams, 2004) and k-anonymity is not resistant against certain types of attacks (e.g., attribute and probability attacks) (N. Li et al., 2007; Machanavajjhala et al., 2006), every newly proposed model has shortcoming as well (Clifton & Tassa, 2013; Dankar & Emam, 2013; Pinto, 2012). These newer models may work well in relational micro-data, but their concept is complicated to implement at the trajectory level. It has been shown that k-anonymity, due to the simplicity of its definition, can be reasonably achieved in most cases, and based on it more secure privacy protection models can be developed (Kenig & Tassa, 2012). Moreover, efforts to improve k-anonymity are still in progress (H. Dai et al., 2020; J. Wang & Kwan, 2020).

1.3. Data utility

The task of privacy protection methods is to publish safe private data while providing measurable benefits to data recipients (Sankar et al., 2010). Data utility is described as the value of a given dataset in terms of providing information to different data recipients to support an analysis to be valid and meaningful. The anonymization process can potentially downgrade the data utility of a dataset through data transformation to satisfy a privacy criterion. The balance between the level of privacy to be protected and the degree to which data utility is downgraded or information (spatial, temporal, and/or semantic information) is lost, is considered a primary and yet under-assessed issue that all data anonymizers deal with. However, researchers still struggle with deciding upon the minimally acceptable level of privacy. In fact, the main goal of developing an anonymization method is to maximize this trade-off.

In general, two approaches exist when assessing the utility of anonymized data. The first is data-centric which is generic and regardless of the end-uses, measures the dissimilarity between an original and anonymized dataset. The second one is application-centric which considers how a specific application is affected by the precision of the anonymized dataset and could benefit from data. Examples of data utility include average displacement distance (Richter, 2018), descriptive statistics (e.g., mean center)

(Richter, 2018), the amount of data suppression (Eom et al., 2020), Ripley's K Function (Zhang et al., 2017), Kernel Density, preservation of the spread of disease, hotspots or any particular pattern (T. Li et al., 2020), and more. To enhance both privacy and the amount of meaningful information that the data recipients can receive, contextual information has been used such as census-based population data (Kwan et al., 2004), land-use or Point of Interest database (Tu et al., 2017; Zhang et al., 2017), and road networks (Swanlund, Schuurman, Zandbergen, et al., 2020). Association with contextual information that is not always readily available and can lead to the need for extra work to find, clean, and process these additional data. Nonetheless, some believe that this added burden is worthwhile, given the importance of privacy protection (Kounadi & Leitner, 2014).

1.4. Thesis structure

This introductory chapter is followed by three more chapters. The second chapter details the design of a privacy protection framework for trajectories. The framework identifies visited meaningful locations related to users' habits and makes them indistinguishable from k-1 near locations with the same function in a land-use map by swapping them with one of the candidate locations to prevent their re-identification. Next, it unlinks the different locations of users and cuts off the relationship between actual users and their trajectories using the pseudonym swapping approach, consequently, to prevent users' re-identification when having background knowledge about any random spatiotemporal points or mobility model of users. Finally, the framework produces anonymized trajectories with maximum semantic consistency and minimum deviation from original trajectories, while preserving a healthy balance between privacy and data utility. With these future developments for trajectory privacy in mind, the third chapter describes the implementation of an all-in-one python toolbox in the ArcGIS Pro environment to enable non-expert users to anonymize trajectory data before publishing. The toolbox provides tools to process raw GPS points through the construction of semantic trajectories depending on user-defined parameter values, to identify personal frequent locations based on the frequency of locations' occurrence, and lastly, it offers a variety of anonymization techniques. These two chapters are standalone articles that were originally written for publication in peer-reviewed journals. Finally, the fourth chapter concludes the thesis by summarizing and synthesizing the themes presented throughout the two main chapters, explicating the contributions made by the framework and the toolbox, and suggesting directions for future work.

Chapter 2. A framework for protecting the privacy of semantic trajectories and locations related to users' habits

2.1. Abstract

The advancement of GPS-enabled devices has enabled people to record their locations and offered the promise of anonymization. However, as surveillance has continued and intensified, this promise has not been met. While point data can reveal a user's identity, it is easier to mask. Trajectories are more challenging as the combination of locations, in particular locations related to users' habits, is often unique. In this paper, we offer a combined privacy protection framework that a) prevents true detection of locations related to users' habits by swapping them while preserving their functions; and b) prevents entire trajectories from being identified by unlinking different parts of trajectories from each other and their actual owners using the pseudonym swapping method. This framework attempts to prevent users' re-identification when knowing important locations or any spatiotemporal points of a user while maintaining the semantics of locations and the minimum shape deviation from the original trajectory. To conduct an experimental evaluation, we used pseudo-anonymized Geolife and T-Drive datasets. Finally, we demonstrate the trade-off between privacy level and the data utility of datasets after anonymization.

2.2. Keywords

Privacy protection, semantic trajectory, indistinguishability, unlinkability, locations related to users' habits.

2.3. Introduction

Digital technology is still relatively young but has already affected every facet of modern life. The advancement of GPS-equipped devices has enabled individuals to record their spatial data (Kiukkonen et al., 2010; Zheng et al., 2010). A sequence of time-stamped spatial data recorded by an individual forms trajectory that contains a rich source of spatial, temporal, and semantic information for mobility-related decision making and movement analysis (Giannotti, Nanni, Pedreschi, Renso, et al., 2009; Yan et al., 2013). However, there is a high risk of privacy violation when publishing such person-specific data (Duckham & Kulik, 2006) because important locations of individuals, such as home/workplaces and Places of Interest (POIs) as well as their movement patterns, might be explicitly revealed due to the regularity of spatial behaviour. As a consequence and drawing on the concept of individuals' geoprivacy—the individual right to determine how, if, and when their spatial data are shared with others (Duckham & Kulik, 2006; Elwood & Leszczynski, 2011; Kwan et al., 2004) — trajectory anonymization methods were devised to better safeguard the identity, exact locations and sensitive information of individuals.

Given that people have periodic and regular movement patterns, some locations are frequently visited within that pattern and reveal users' identities with a higher probability. Given the importance of locations for a user, they must be treated differently (Sila-Nowicka & Thakuriah, 2016). For example, locations such as home (Krumm, 2007), home-workplaces (Freudiger et al., 2012; Golle & Partridge, 2009), most frequented locations (Zang & Bolot, 2011; Rossi & Musolesi, 2014), and/or long-stayed locations (Beresford & Stajano, 2003; Y. Dai et al., 2018; Unnikrishnan & Naini, 2013) can lead to the re-identification of users and using their personal information in an unauthorized way, even after aggregating spatial and temporal units. Zang & Bolot (2011) showed that it is necessary to aggregate frequent locations to the city level to sufficiently anonymize even a week's worth of location data. Important locations of users are the most common type of background knowledge in attacks against users' re-identification (Fiore et al., 2020). To protect users' re-identification when knowing their important locations, in particular mostly frequented locations, the framework presented here employs the location swapping method (Zhang et al., 2017) which displaces locations to new locations with similar functions. The purpose of our framework is to preserve semantic information of a dataset for meaningful and efficient analyses. For example, if an office, visited by a person from 8:00 am to 5:00 pm, is relocated to a bar or a cinema can result in non-sense and inaccurate analysis.

An important issue in anonymizing trajectories is that a person being at a particular place at a particular time is a highly unique combination of attributes (Bettini et al., 2005), and only having as few as three to four of these spatiotemporal pairings makes the odds of someone else matching that profile astronomically small, even in an aggregated dataset (de Montjoye et al., 2013). The more unique a trajectory, the higher risk of a privacy breach and user re-identification. Another goal of the proposed framework, therefore, is to provide “unlinkability” between users and their trajectories by removing linkage between visited locations of a user and reduce the uniqueness of trajectories in which the combination of locations is unique. “Unlinkability” of locations means that within a dataset, an attacker cannot sufficiently distinguish whether these locations are related to the same user or not and is unable to infer the movement pattern of the user (Pfitzmann & Köhntopp, 2001). To this end, our framework tries to break the continuity of location exposure, drawing on the concept of a mix-zone (Guo et al., 2018; Salas et al., 2018). Mix-zones are spatiotemporal areas where co-occurred trajectories pass close enough at a specific distance within a specific time interval, and their pseudonyms are swapped or changed to new random ones to confuse the attacker about the association of incoming and outgoing trajectories. In this study, co-occurred co-located trajectories are selected as candidates for pseudonym swapping that are similar in terms of direction to prevent non-valuable and inefficient pseudonyms. In the end, move points within mix-zones are suppressed to avoid the connection between trajectories after pseudonym swapping.

The contribution of the paper is to describe the development of a privacy protection framework to prevent re-identification attacks when knowing important locations or any random spatiotemporal points of users. While simple masking of point data has higher degrees of certainty associated with the process (Swanlund, Schuurman, & Brussoni, 2020), trajectory masking is more complex and cannot offer statistical measures of certainty. The endeavour is nonetheless critical to ensuring greater privacy for individuals. Anonymizing only important locations, which have more importance for both users and attackers, rather than all sampling points, reduces the computation cost and information loss and produce a dataset beneficial for location-based analysis (e.g., hotspot detection, understanding the spatial distribution of facilities) since the majority of sampling points are kept intact. Pseudonym swapping of trajectories keeps movements and the shape of trajectories intact and produces a dataset beneficial for making mobility maps (e.g., for urban planning, proposing transportation services), outlier detection to find trajectories that do not comply with the general behavior of the dataset, trajectory classification and inferring activities and transportation modes, clustering similar sub-trajectories, making landmark graphs where landmarks are frequently traversed routes (used for route recommendation, etc.), detecting social ties, and more. However, this framework does not preserve the relationship among visited locations due to frequently changing pseudonyms of trajectories.

The output of the framework is an anonymized dataset that contains de-personalized micro-trajectories with the least deviation of trajectories. Personalized Micro-trajectories—trajectories of a particular person—are information-rich, but runs a high risk of enabling re-identification, while aggregated trajectories—trajectories of multiple users with average information about them—are safer but have low utility. Between those extremes, de-personalized micro-trajectories are trajectories of single users which is not linked to a particular person. We felt this output achieved a good balance between privacy and information.

To empirically assess our proposed framework, we used two pseudonymized trajectory datasets, Geolife and T-drive in which each user has a unique user ID. In the pre-processing step, we structured raw trajectories in the form of visited meaningful locations, referred to as semantic stops, and map-matched moves. Then we determined frequented locations of each user and accordingly employed a location swapping method to prevent their re-identification; finally, we protected the privacy of the whole trajectory by disassociating trajectory owners from their entire trajectories using a pseudonym swapping approach. To assess our proposed framework, we evaluated the balance between data utility and privacy of the output.

This paper begins by providing a review of the range of available anonymization techniques for protecting the privacy of trajectories at publishing time. Next, the paper turns to the design of a series of methods used to construct semantic trajectories and protect their components. Finally, we present the experimental results of the application of the proposed protocol on two real-world trajectory datasets, before offering a discussion of the merits and shortcomings, as well as future research directions.

2.4. Related works

So far, many scholars and experts have carried out studies about the anonymization of trajectories and several ways of classifying trajectory anonymization methods have been introduced. In this section, the classifications that matter for our framework are described. One way of classification was offered by Bonchi et al. (2011), which arranged methods based on how they deal with trajectory components into two main classes. The first class anonymizes the entire trajectory at once to satisfy full-length anonymity (Abul et al., 2010, 2008; Gramaglia & Fiore, 2015; Nergiz et al., 2008), leading to huge computation costs while ignoring the semantics of points. The second class, which is the focus of this paper, breaks up each trajectory into sets of QIs, to be dealt with individually (Huo et al., 2012; Monreale et al., 2010; Poulis, Loukides, et al., 2013; Terrovitis & Mamoulis, 2008; Y. Wang & McArthur, 2018; Yarovoy et al., 2009) rather than treating all trajectory points equally (Abul et al., 2008; Andrienko et al., 2009; Monreale et al., 2010). The reason for this approach is that attackers mostly have background knowledge about individuals' visited locations or parts of trajectories rather than the entire spatiotemporal points (Huo et al., 2012).

The second way of classification that is relevant to our framework divides methods into three main classes, based on their objectives to protect the geoprivacy of the individual (Fiore et al., 2020): indistinguishability adhere to k-anonymity and its extensions, undetectability adhere to differential privacy, and unlinkability adhere to mitigation principle. The third way of classification we drew upon is more technical, assigning techniques that will protect geoprivacy to each of the objectives (Fiore et al., 2020): generalization/cloaking/gird-masking, suppression, clustering, perturbation, adding random noise, and pseudonym changing/swapping. Following the classification of the objectives and techniques of anonymization, we explain each of the objectives and briefly present some examples of different techniques utilized depending on each objective.

“Indistinguishability”, means granting that each object in a dataset must not be distinguishable from a set of other objects in the same dataset, called anonymity set. Based on this concept, k-anonymity was introduced for a table to ensure each record in a table, containing information of a user, is indistinguishable from at least k-1 other records in anonymity set in which all records have the same QI (Sweeney, 2002). The notion of k-anonymity was adapted for a spatial database (Gruteser & Grunwald, 2003) by making a location indistinguishable from k-1 other locations, before being adapted specifically to trajectories by Nergiz et al., (2008) after which it has been referred to as “trajectory k-anonymity.” The anonymization techniques assigned to indistinguishability can be categorized as the combination of clustering and perturbation and the combination of generalization and suppression. For example, Abul et al., (2010) satisfied trajectory k-anonymity by clustering k-identical trajectories within a distance threshold and perturbating trajectories to the center of clusters to make them indistinguishable (Abul et al., 2010, 2008). However, perturbation trajectories without considering the semantic of points destroy the semantic consistency of trajectories and association with road links.

Generalization involves coarsening at least one characteristic of trajectory, time, location, and semantics to create anonymity regions with k users to make each user indistinguishable from k-1 others. Suppression eliminates particular points that do not satisfy privacy criteria (Chen et al., 2013; Terrovitis & Mamoulis, 2008), and is often combined with other techniques for better protection (Andrienko et al., 2009; Gramaglia & Fiore, 2015; Sila-Nowicka & Thakuriah, 2016).

Despite these technical advances, a growing body of literature has shown that even when spatiotemporal information has been greatly generalized, there is still a high risk of re-identification when locations related to users' habits are known—e.g., home (Krumm, 2007), home-workplaces (Golle & Partridge, 2009), N most frequented locations (Zang & Bolot, 2011), making it necessary to treat locations differently based on their personal importance. Long-stayed locations and frequent locations, referred to as personally important locations, can be used as a mobility signature of a user and disclose habits of individuals and sensitive information with higher probability. Many studies have shown that protecting the privacy of only these locations can be sufficient instead of all sampling points in a trajectory. A simple approach to preserve the privacy of personally important locations is to suppress them (Sila-Nowicka & Thakuriah, 2016), although it can cause high information loss since they contain valuable information. Wang & McArthur (2018) improved information loss by cloaking a subset of important locations in anonymity regions with k locations adhere to spatial k-anonymity (Kalnis & Ghinita, 2009). Dai et al. (2018) protected personally important locations as well as user-defined semantically sensitive locations, which required users' participation in defining sensitivity, by replacing them with POIs with a similar semantic category. However, trajectories are sufficiently unique that even protecting all important locations cannot prevent re-identification when few random spatiotemporal points are known. Ma et al. (Ma et al., 2013) showed that 50% of trajectories can be uniquely revealed with having only eight spatiotemporal points. Rossi et al. (2015), showed that 5% to 60% of users can be uniquely re-identified even in a dataset with a spatial granularity of 10 km^2. Cecaj et al., (2016) showed that about four spatiotemporal; points are enough to uniquely locate 90% of the users by matching their geotagged messages released by a social network to their corresponding trajectories in a real-world trajectory dataset by applying a simple statistical learning approach. They also showed that about 20% of users in the trajectory dataset are associated with a unique social network user through sharing a number of common points.

“Undetectability” of a user means that the attacker cannot sufficiently detect whether the user exists in a released aggregate dataset or not (Pfitzmann & Köhntopp, 2001). These approaches add noise to each coordinate of a GPS point or the whole trajectory either randomly or boundedly (Dankar & Emam, 2013; Deldar & Abadi, 2018; Duckham & Kulik, 2005; Srivatsa & Hicks, 2012) adhere to differential privacy (Dwork et al., 2006), and merge trajectories and return a sample as a result, so that an anonymized dataset cannot reveal whether a specific individual is present or not. These methods publish aggregate statistics such as trajectories’ density (Acs & Castelluccia, 2014; Alaggan et al., 2015) or people counts (Hay et al., 2016) rather than atomic trajectories, and are useful for examining the statistical behaviour of the user population. However, Srivatsa & Hicks (2012) used different models of random noise and finally observed that the risk of user re-identification can be reduced by adding only a high amount of noise. Moreover, due to the high dimensionality of trajectory data and strict mathematical complexity of differential privacy methods, no method has been devised yet to practically achieve this goal by adding noise to trajectories. Existing methods either consider relaxed notions of differential privacy or generate synthetic data based on characteristics of actual trajectories derived from historical trajectories.

“Unlinkability” of locations of a user means cutting off the linkage between locations of a user such that an attacker cannot track a user and re-identify the entire trajectory of the user in a released dataset. To address the aforementioned issues, more heuristic techniques with an objective of unlinkability adhere to mitigation principle have been devised to reduce privacy risks without satisfying specific privacy criteria (Fiore et al., 2020). These anonymization techniques are based on the concept of pseudonym changing/swapping combined with the concept of mix-zone. The idea of path perturbation was introduced by Baik Hoh & Gruteser (2005) by which the positions of two co-located co-occurred trajectories are slightly perturbed to get so close to each other and they are swapped and at the end, each trajectory is paths of different users. However, this method was not applied and tested to a real-world dataset.

Based on the concept of pseudonym changing, trajectory segmentation was introduced by Song et al. (2014) to break each trajectory into shorter sub-trajectories using specific time windows and assign a different pseudonym to each segment. Yet still, 80% of sub-trajectories, especially long ones, were unique even with a time window of 6 hours. The concept of pseudonym changing/swapping within mix-zones is one of the fundamental approaches implemented for location privacy (Arain et al., 2017). Mix-zones are spatial areas where at least two co-occurred trajectories pass close enough at a specific distance within a specific time window. Users' spatiotemporal points within mix-zones are not available, such that the pseudonyms of trajectories after entering mix-zones are swapped or changed to new ones and the linkage between the incoming and outgoing trajectory segments of a user is obfuscated. The concept of the road network-based mix-zone (Buttyán et al., 2007) was introduced and then improved by Palanisamy & Liu (2011) as an alternative and complementary approach to spatial generalization/cloaking to protect the privacy of vehicles traveling on road networks. Drawn on the concept of mix-zone, a trajectory swapping method, called SwapMob, was recently introduced by Salas et al. (2018). SwapMob tries to remove the linkage between an actual user and their trajectories by swapping pseudonyms of trajectories within the mix-zones. After swapping, each trajectory is linked with multiple users. However, generating mix-zones at inappropriate occasions without considering time information and direction constraints can be invalid and ineffectual. Swapped trajectories can be mapped to the actual trajectories if there is a huge difference between the direction or time information of two trajectories before and after swapping. Another shortcoming of the SwapMob method is that POIs of users are disclosed at their exact positions and there is still a risk of users' re-identification when knowing their POIs if not involved in the trajectory swapping process. The conceptualization of our framework is very close to the SwapMob method with some improvements to generate mix-zones at proper occasions while considering time and direction constraints and prevent detection of the exact position of POIs and solve the problem of swapping in datasets with a low density of trajectories.

2.5. Proposed framework

The purpose of the proposed combined privacy protection framework is to address current gaps in the technology and create an accessible approach to prevent the exposure of users' frequent locations and trajectories with relatively little modification of points and information loss. To meet this purpose, we have sought to prevent the inference of locations related to users' habits by relocating these locations to new locations with a similar semantic, while keeping the rest of the locations intact and to reduce the uniqueness of trajectories by breaking their continuity and obfuscating the linkage between actual users and their trajectories. The end product of the framework is micro-trajectories where each trajectory is a combination of sub-trajectories of different users. This output has the potential to be used by both location-based analyses where visited locations matter and movement-based analyses where detailed, not-aggregated but still anonymized information about movements matters. Our framework shifts the focus from aggregated trajectories with average information to de-personalized micro-trajectories with information about single individuals. The framework consists of three separate, but interconnected steps as shown in Figure 2.1. In the first step, we collect individuals' GPS points and convert them to trajectories with meaningful inner structures by using stop detection and map-matching methods. The second step involves the detection of personal frequent locations. The last step employs location swapping to prevent true detection of individuals' frequent locations and pseudonym swapping to prevent trajectory tracking and expanding knowledge about unknown parts of a user's trajectory. When publishing the anonymized dataset, the methods used to anonymize the dataset are shared.

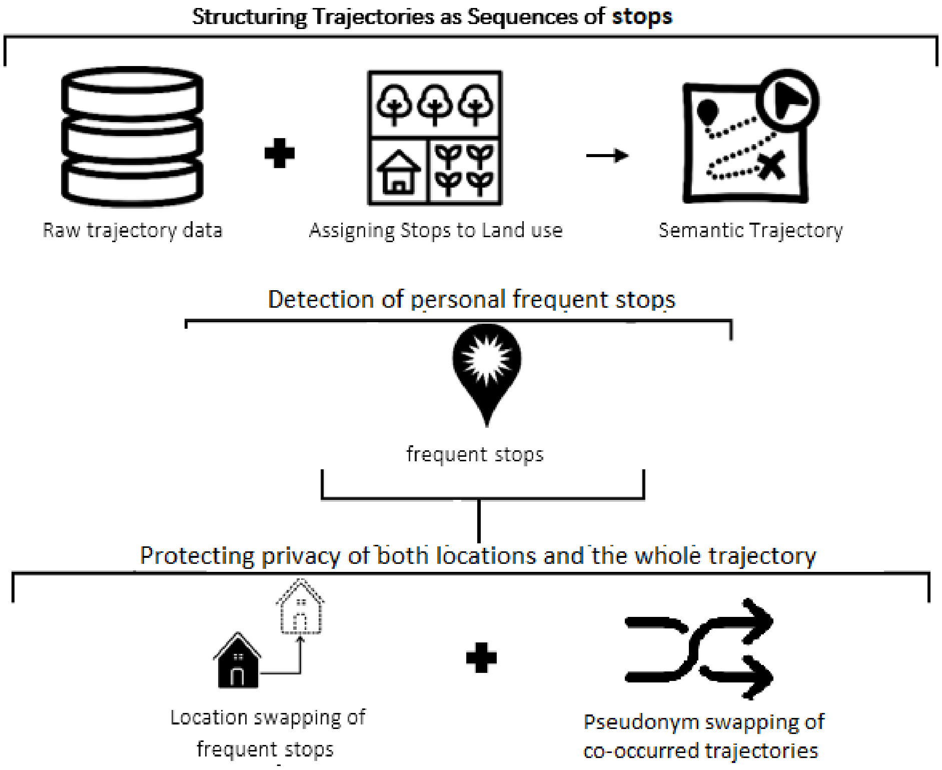

Structuring Trajectories as Sequences of stops
Raw trajectory data
Assigning Stops to Land use
Semantic Trajectory
Detection of personal frequent stops
frequent stops
Protecting privacy of both locations and the whole trajectory
Location swapping of frequent stops
Pseudonym swapping of co-occurred trajectories

Figure 2.1. A visual representation of the included steps in the development of the proposed anonymization framework. In the first step, raw trajectories are deconstructed into sequences of semantic stops and moves. In the second step personal frequent locations are detected. In the third step, several anonymization techniques are employed to protect locations and the entire trajectory.

2.5.1. Deconstructing trajectories to the sequences of semantic stops and moves

This subsection describes how we constructed the stop and move structure of raw trajectories. First, users’ trajectories were divided into daily sub-trajectories. Then their inner structure was formed using stop detection and map-matching methods. A stop, also known as an episode or stay point, is a part of a trajectory where a user has stayed over a period of time for a specific activity. To detect stops, we computed low-speed clusters with a speed lower than maxSpeed over a minimal possible staying time, minTime, and a maximal possible staying distance, maxDis, and assigned all points within a cluster to the center of the cluster indicating a stop point.

In developing our framework, we are looking for semantic stops that visit a meaningful place for minimal time duration and are associated with a certain activity/semantics. Not all stops meaningfully represent an activity, such as stops in traffic jams, traffic lights, and so on. To ignore such noisy stops, we used a road network dataset and employed a near analysis function to determine whether a stop appears on a road or not. Once noisy stops were detected and considered as move points, we linked candidate stops to the intersected or the nearest parcels/units in the land-use map to annotate semantic information such as land-use function/category. However, there were redundant stops at neighboring time windows that appeared in the same location and were part of the same stop as highlighted in Figure 2.2. To solve this problem, we merged redundant stops at neighboring time windows (in our case 10 minutes) and represented them as a single semantic stop. The value of 10 minutes is arbitrarily used for testing. Finally, a stop was described as follows: stopID is the identifier of the stop, userID is the identifier of an individual user, locID is the identifier of the corresponding location, (t_enter, t_leav) is the enter and leave time as the timestamps of the first and the last GPS point within the stop, and avgSpeed is the average speed of the stop points

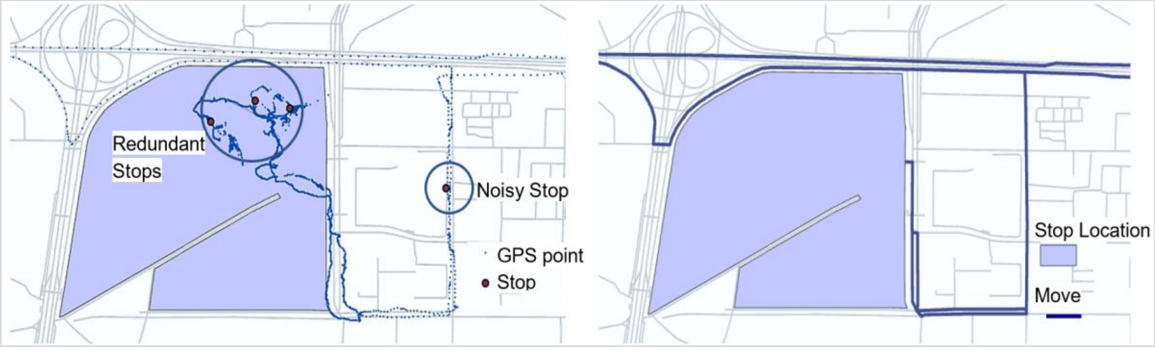


Figure 2.2. The need to ignore problematic stops (noisy and redundant). Noisy stops such as stops in traffic jams are on the roads. Redundant stops are separate stops but part of the same stop and appearing in the same parcel in the land-use at proximate time windows (e.g., one stop at 9 AM and one stop at 9:30 am).

The moves are GPS points recorded between stops and/or gaps. The gaps are parts of trajectories that occur between two consecutive points which are spatially and temporally distant, depending on the average sampling rate of trajectories (by default five times larger than sampling interval), and do not match any road segment. We assume that users travel by road or subway and map move points between stops and/or gaps to traveled road or subway segments by using GIS-based map-matching

(Dalumpines & Scott, 2011). It is worth noting that map-matching, which is time-consuming and has high computation costs especially in large-scale datasets, is an optional process in this study.

2.5.2. Personal frequent locations detection

Most people repeat the same spatial routes almost every day of the week, resulting in the re-occurrence of some locations that reveal the identity of the users, their habits, and associated private information with a higher probability. For example, if a user usually travels from location A to B at a specific time window in the morning, and travels from B to A at a specific time window in the afternoon, attackers can easily infer that A is home and B is the workplace and can discover a user’s home/work addresses by means of geocoding tools. The number of times that a location occurs in the trajectories of a user is called the frequency, showing the significance of that location for the user. With a set of meaningful locations, we detected personal frequent locations relying on the fact that the knowledge of the frequency of a location leads to a better assessment of the privacy disclosure risk of that location. For instance, the top two frequent locations likely correspond to users’ homes/workplaces, and the third frequent location is likely to be a personal POI that a user often visits. If an attacker knows frequent locations even without time information, there is a high risk of user re-identification. To find frequent locations, we defined *frequency* threshold as (frequency of a stop)∕(total number of user's daily trajectories) which lower bounded the occurrence of a location in a trajectory set of a user. Thereby, a location is frequent if its frequency is greater than the *frequency* threshold. For example, when the value of frequency threshold was 0.5 and a location (e.g., home location) occurred in more than half of the user’s daily trajectories, it was considered a user’s frequent location.

2.5.3. Location and trajectory privacy protection

After detecting personal frequent locations, a privacy protection framework was developed to protect those locations while maintaining a satisfactory level of data utility and paralleling their semantics.

Location swapping of frequent locations

To anonymize frequent locations, the location swapping method was employed to randomly relocate a location to one of the potential locations with a similar function (e.g., households) in the land-use map within a buffer with a defined radius around the original location. The radius of buffers varied depending on the local density of locations to contain k similar locations in terms of semantics with the original location. One limitation is that in areas with low location density (e.g., suburbs), there would be huge information loss to achieve a satisfying level of privacy. After relocating the original location to one of k random locations, SKA, also known as nth nearest neighbor number close to the relocated location, was measured. SKA is the number of potential locations that fall within a buffer centered at each masked location with a radius r, where r equals the distance from the masked location to the original location. The basic idea is preventing true detection of home/workplaces and POIs when publishing an anonymized dataset with the probability of 1/SKA. The spatial distance between an original location and a new location is anonymous and just the value of SKA is published.

Location swapping was the most reasonable method to anonymize frequent locations since it provides a healthy balance between privacy and data utility and maintains the semantics of locations. As frequent locations mostly correspond to home/workplaces (as evident from section 2.6.3), after anonymization they still have a similar category but at a different position. For example, if an attacker knows the approximate home/work address of a target user, even by reverse geocoding of the published locations he/she cannot easily identify the user in a dataset, since the user's true location has been displaced to a new home/work address. However, we assume that false identification, meaning incorrect association of a sensitive attribute to a household (Seidl et al., 2018), is not a concern. One of the most challenging aspects of anonymizing frequent locations is that if a person frequently visits a location, e.g., several times a week, they need to be relocated to the same location. To deal with this issue, we perturbated a frequent location to a single random but a fixed alternative; otherwise, the median of a cluster of new different locations revealed the approximate position of the original location.

To maintain the integrity and shape of an entire trajectory and increase data utility while reducing computation cost, only personal frequent locations were displaced, rather than all visited locations. Non-frequent locations can be just as, if not more, sensitive than frequent ones, and still pose re-identification risk. To protect the privacy of non-frequent locations, we applied the pseudonym swapping method for trajectories that meet each other around non-frequent locations as described in the next section to disconnect a sequence of visited locations from the actual user.

Pseudonym swapping

To protect the privacy of an entire trajectory where a combination of spatiotemporal points is unique, motivated by the concept of mix-zones, we broke the continuity of trajectory disclosure by swapping pseudonyms of trajectories within the mix-zones. Swapping pseudonyms confuses the incoming and outgoing trajectories and prevents attackers to expand their knowledge to unknown parts of a trajectory and visited locations of a user by following a wrong user. This approach also preserves the traveling information of users, such as speed, direction, and time of travel (in general). To carry out the pseudonym swapping process, dynamic circular mix-zones were created where at least two trajectories pass close enough at a specific distance (equal to the radius of mix-zone) within a specific time window at a road intersection or around visited locations (non-frequent) and their pseudonyms were swapped. The time window and/or the radius of mix-zone was considered by default six times larger than the sampling interval of the dataset for distance or time.

We generalized the time resolution of the dataset, and consequently the speed of trajectories according to the time window of mix-zones to prevent timing and speed attacks, where an attacker compares the time and speed of incoming and outgoing trajectories from mix-zones and tries to associate between the incoming and outgoing trajectories based on the similarity of time and speed information. However, the deviation in the direction of trajectories after pseudonym swapping and the non-uniformity in trajectories could still provide valuable information to the attacker to map the new pseudonyms to true trajectories. Given that, we only chose trajectories as candidates for pseudonyms swapping, if for each incoming trajectory, there was at least one outgoing trajectory from mix-zone (except the true outgoing trajectory) with an angle less than 270 degrees (as shown in Figure 2.3). To reduce the probability of mapping trajectories based on the deviation in direction and to avoid the sharp change of positions on the swapped trajectories, we suppressed corresponding move points within mix-zones.

We restricted the distance between mix-zones such that every two pseudonyms of a trajectory must be at least 10 times larger than the sampling rate (or any arbitrary value) apart. However, two users might travel together along the same roads and their trajectories get frequently close to each other, resulting in generating inordinate mix-zones. To avoid ineffective pseudonym swapping, consecutive mix-zones for the same trajectories with frequent intersections were disregarded and only the last mix-zone was considered adequate for pseudonym swapping.

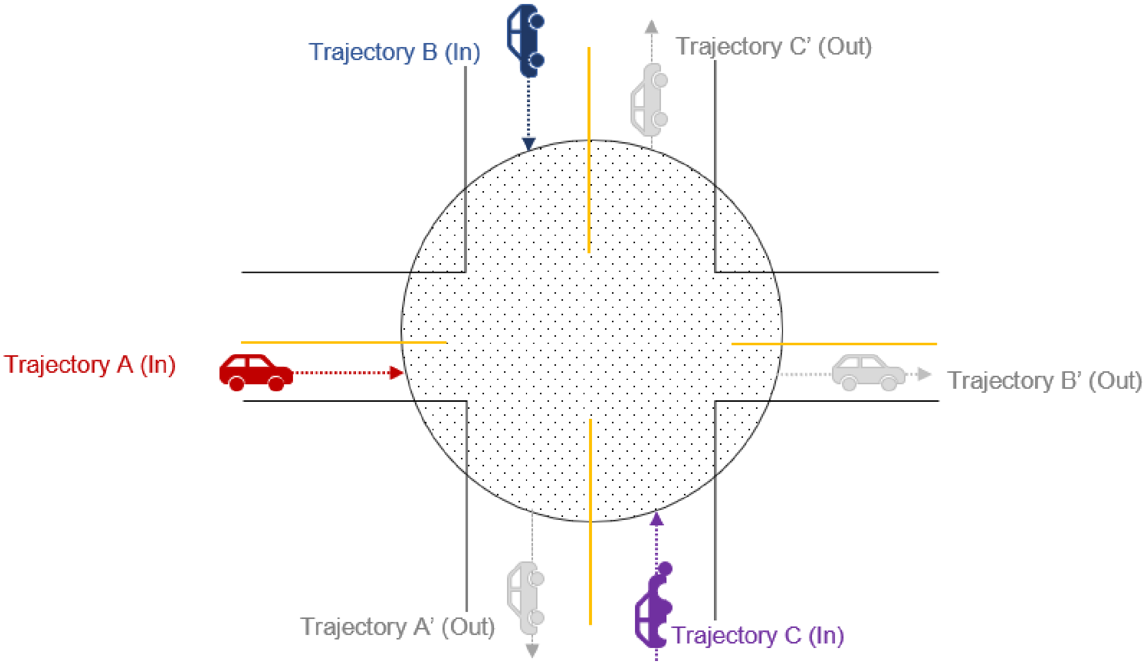


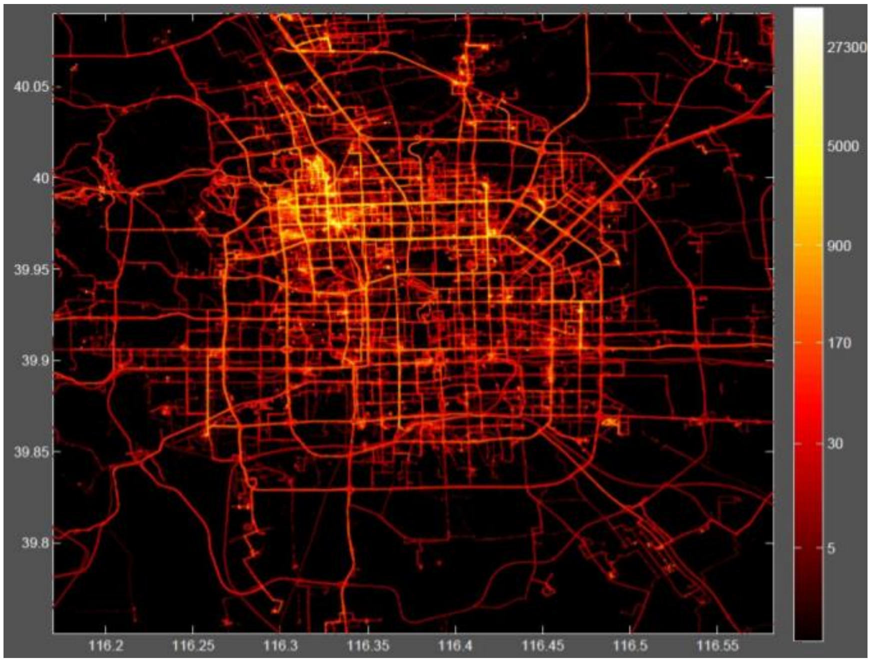
Figure 2.3. Pseudonym swapping of co-occurred candidate trajectories within dynamic mix-zones.

2.6. Evaluation

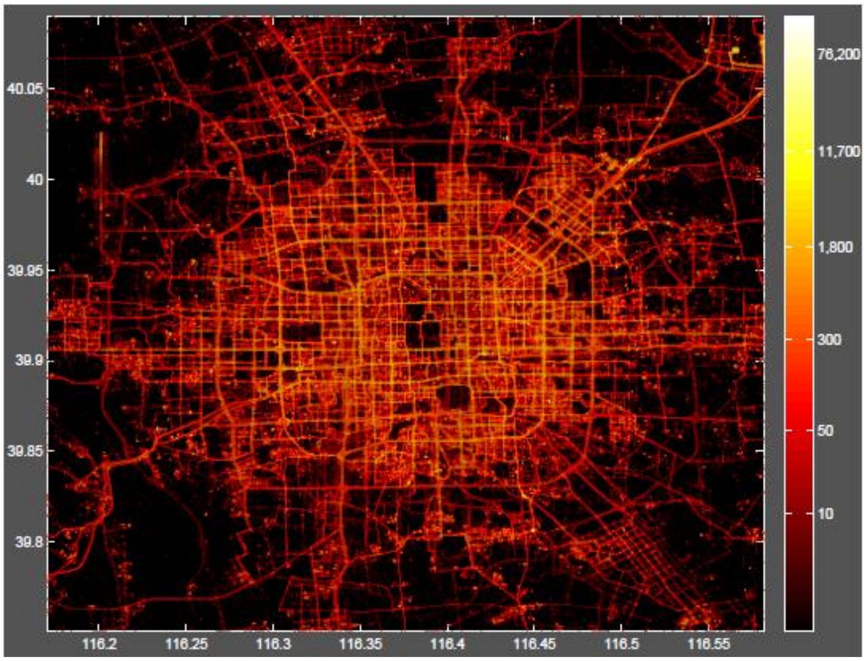
2.6.1. Dataset

This section empirically illustrates the output of the concepts discussed using two real-world datasets. We used pseudo-anonymized GPS records from the Geolife project, a Collaborative Social Networking Service (Zheng et al., 2008, 2010, 2009). In Geolife, 182 users tracked their daily outdoor movements from 2007 to 2012, mainly in Beijing.

The dataset contains the timestamped latitude and longitude of 17,621 daily trajectories with a total distance of about 1.2 million km and a total duration of 48,000+ hours. A high sampling interval (1~5 s) occurred in 91.5 percent of the trajectories. The City of Beijing, China was chosen as the study area since most GPS records are placed in this city. To confirm our method is generalizable, we tested it on a T-Drive dataset which contains GPS trajectories of 10,357 taxis during the period of February 2 to February 8, 2008, within the same study area, but with different sampling intervals. The average sampling interval is about 177 s with a distance of about 623 m. The total number of points in T-Drive is around 15 million and the total distance of the trajectories reaches 9 million km. We filtered out both datasets by removing short-time daily trajectories (less than 10 min) and high-speed movements with a speed of greater than 180 km/h. Moreover, to do an accurate evaluation we removed users who had fewer than three stops. Figure 2.4 visualizes the density distribution of the GPS points in both datasets.



Geolife, Source: (Zheng et al., 2009). See https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/ for re-use terms and conditions.



T-Drive. Source: (Yuan et al., 2011). See https://www.microsoft.com/en-us/research/publication/t-drive-trajectory-data-sample/?from=https%3A%2F%2Fresearch.microsoft.com%2Fapps%2Fpubs%2F%3Fid%3D152883 for re-use terms and conditons.

Figure 2.4. Distribution of GPS points within Beijing, where the color indicates the density of the points.

Moreover, the Beijing land-use shapefile for stop detection and the road network shapefile for the map-matching were collected from Open Street Map (https://download.bbbike.org/osm/bbbike/Beijing/).

In spite of the fact that the framework seems complex, the methods can be summarized in four general steps as described below and six detailed steps as shown in Figure 2.5:

I. Converting individuals' raw trajectories into semantic trajectories through detecting semantic stops/meaningful locations and matching moves to road segments.

II. Identifying personal frequent locations by defining a frequency threshold.

III. Employing location swapping method on personal frequent locations to protect their privacy and maintain their semantics.

IV. Swapping pseudonyms of co-located co-occurred candidate trajectories to protect the entire trajectories.

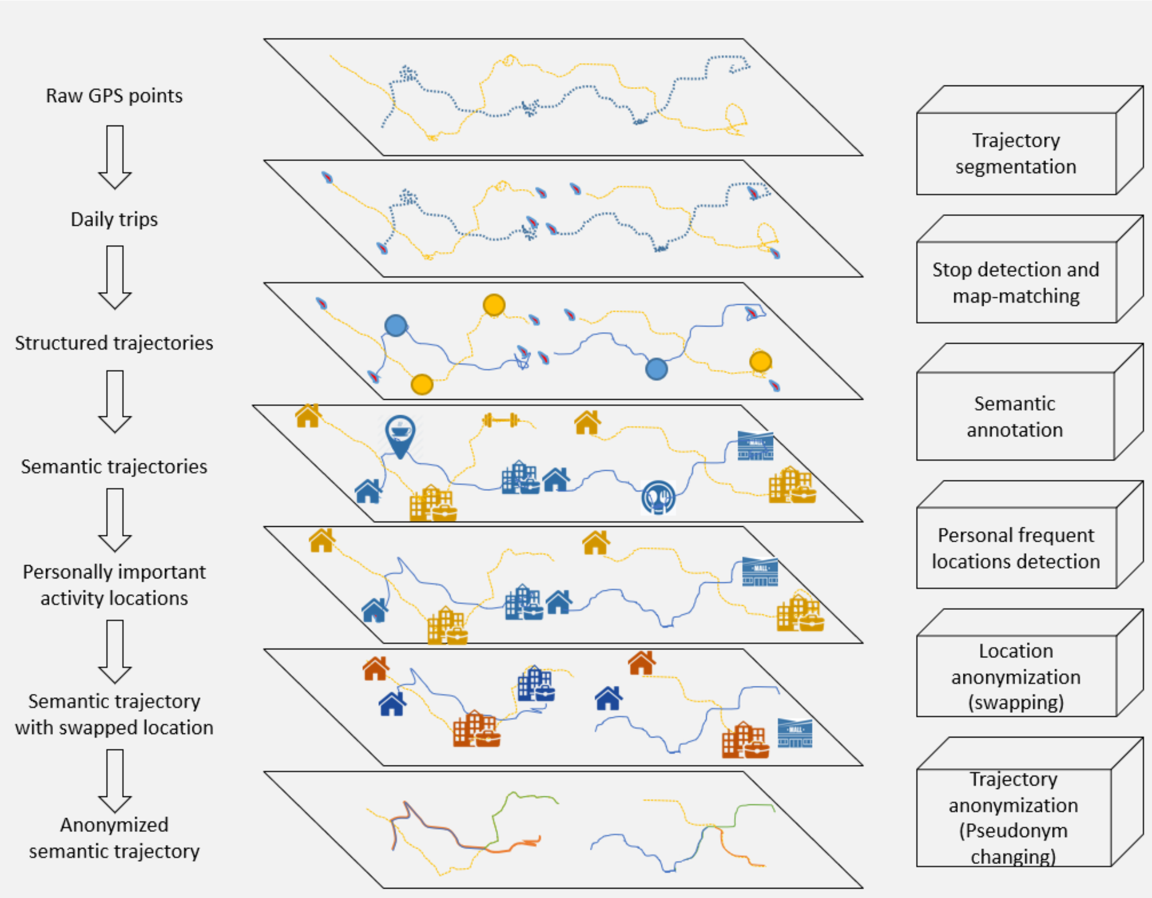

Raw GPS points
Daily trips
Structured trajectories
Semantic trajectories
Personally important activity locations
Semantic trajectory with swapped location
Anonymized semantic trajectory
Trajectory segmentation
Stop detection and map-matching
Semantic annotation
Personal frequent locations detection
Location anonymization (swapping)
Trajectory anonymization (Pseudonym changing)

Figure 2.5. The procedure for trajectory anonymization in a semantic way.

2.6.2. Results of stop detection

The accuracy of stop detection was dependent on the chosen values of its parameters. To this end, we varied the values of *300 s <minTime< 1800 s, 50 m <maxDis< 300 m*, and *0.6 m/s <maxSpeed< 1.8 m/s* (a brisk walking speed for young individuals), respectively and counted low-speed clusters, noisy, redundant, and desire stops. When values were too conservative, for instance, *maxDis=50m* there were more redundant stops, or when *minTime=1800 s*, a greater number of stops were undetected. In contrast, choosing values that were too relaxed led to the detection of a great number of noisy stops. With the values of *MinTime=600 s, MaxDis=300 m, MaxSpeed=1.2 m/s* we observed that 80% of detected stops intersected a polygon in the land-use map, which means the stops had meaningful semantics. Also, there were on average 2.7~3 stops during each daily trajectory in Geolife. Figure 2.6 illustrates the basic statistics of the two datasets by showing the cumulative distribution functions (CDF) of the number of GPS points, daily trajectories, and stop locations per user

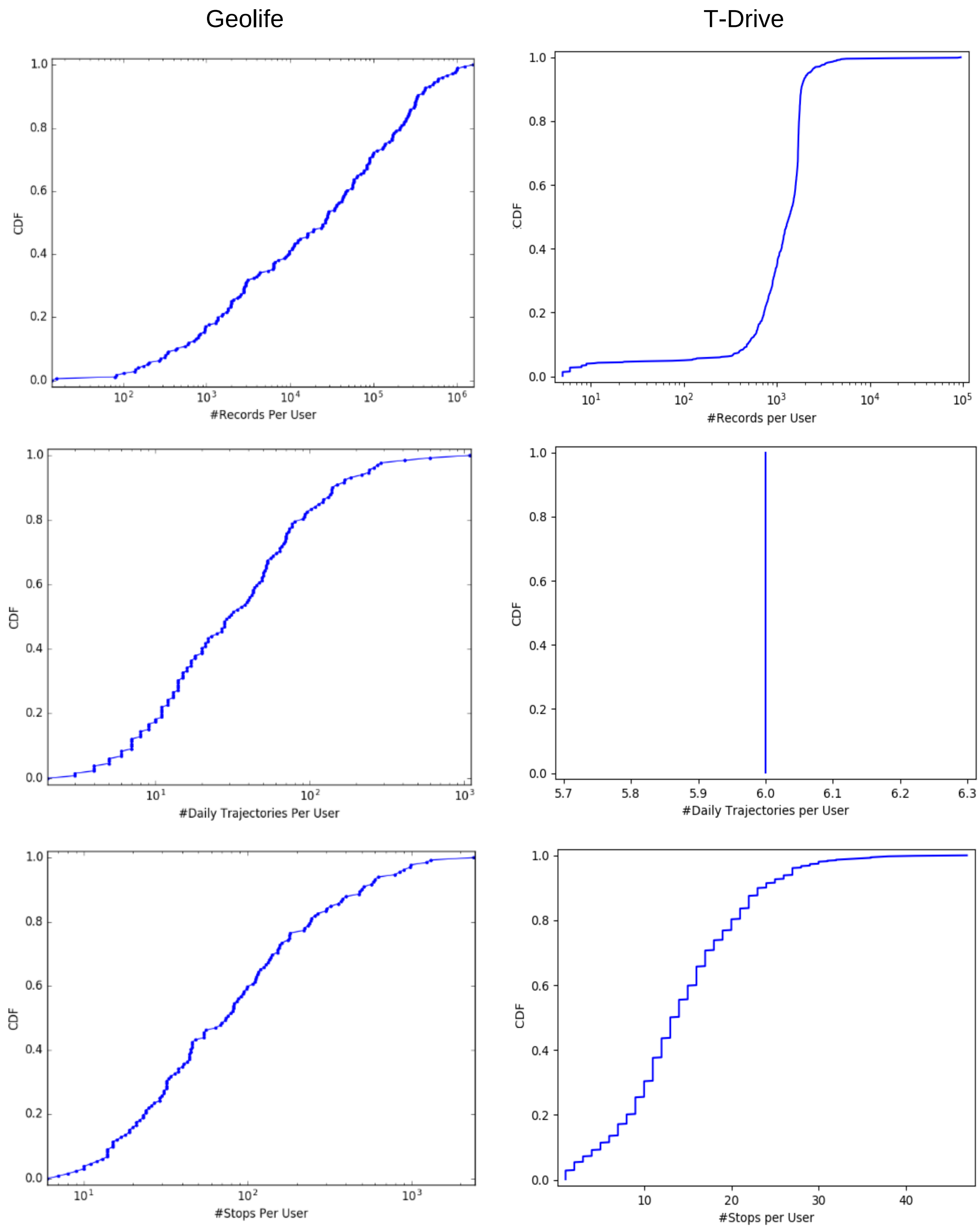


Figure 2.6. CDF of the number of GPS points, daily trajectories, and stops per user after data cleaning for Geolif and T-dive.

As highlighted in Figure 2.6, about 20% of users in both datasets recorded less than 1,000 GPS points. More than half of T-Drive users recorded a large number of GPS points, from 1,000 to 10,000 points, and around 60% of Geolife users have more than 10,000 GPS points. On average, each Geolife user has generated more records compared with T-Drive users because the sampling interval of Geolife is lower than that of T-Drive. By exploring the CDF of the daily trajectories per user, it is shown that about 20% of Geolife users generated less than 10 daily trajectories and around 30% of users have more than 100 daily trajectories, while all T-Drive users have a fixed value of 6 daily trajectories. Finally, around 20% of Geolife users have between 10 to 100 stops and half of the users have more than 100 stops in total, while more than 50% of T-Drive users have less than 20 stops in total and the maximum number of stops for users is 47.

2.6.3. Results of personal frequent stop detection

After detecting semantic stops, frequently visited locations of individuals with different values of *frequency* threshold were identified and the percentage of their population for each land use was measured. As we can notice from Table 2.1, the majority of frequented locations in Geolife belong to residential parcels, followed by educational parcels (university, college, schools, …) which can be workplaces. In T-Drive the majority of personal frequented locations occur in transport facilities (workplaces of taxi drivers), followed by residential parcels. This confirms the fact that frequented locations of individuals are mainly corresponding to homes and workplaces. The increase of frequency threshold caused an increase in the number of residential places compared to all frequented locations of individuals in Geolife and the number of transport facilities compared to all frequented locations of individuals in T-Drive.

Table 2.1. Percentage of frequented locations per land use. This table shows that the majority of frequented locations in Geolife are related to residential parcels (home places), followed by educational parcels, while in T-Drive the majority of frequent locations are related to transport facilities (workplaces), followed by residential parcels.

Land use	**Frequent locations in Geolife**			**Frequent locations in T-Drive**		
	Freq=0.3	*Freq*=0.5	*Freq*=0.8	*Freq*=0.3	*Freq*=0.5	*Freq*=0.8
Residential	44%	58%	67%	23%	28%	30%
Transport Facility	9%	7%	0%	37%	40%	49%
Educational	24%	23%	22%	2%	4%	6%
Commercial	9%	7%	5%	4%	0%	0%
Green Space	5%	2%	0%	8%	5%	0%
Firms	3%	3%	4%	5%	3%	2%
Other	5%	3%	2%	21%	20%	13%

2.6.4. Testing the location-aware anonymization framework

So far, the framework has converted GPS points into sets of semantic trajectories and detected personal frequent locations with the frequency of their re-occurrence. The output of the whole process is shown for a sample user (user A) in the Geolife dataset in Figure 2.7. Details about trajectories of user A and the number of stops for this user and a sample user in the T-Drive dataset are listed in Table 2.2.

Table 2.2. Summary of trajectories for two sample users from the Geolife and T-Drive datasets.

Dataset	**User**	**Duration**	**#Daily trajectories**	**# Stops**	**#Distinct Stops**	**#Frequented locations**
Geolife	A	2008/02/02-2009/09/13	295	492	32	3
T-Drive	B	2008/02/02-2009/09/08	7	47	6	1

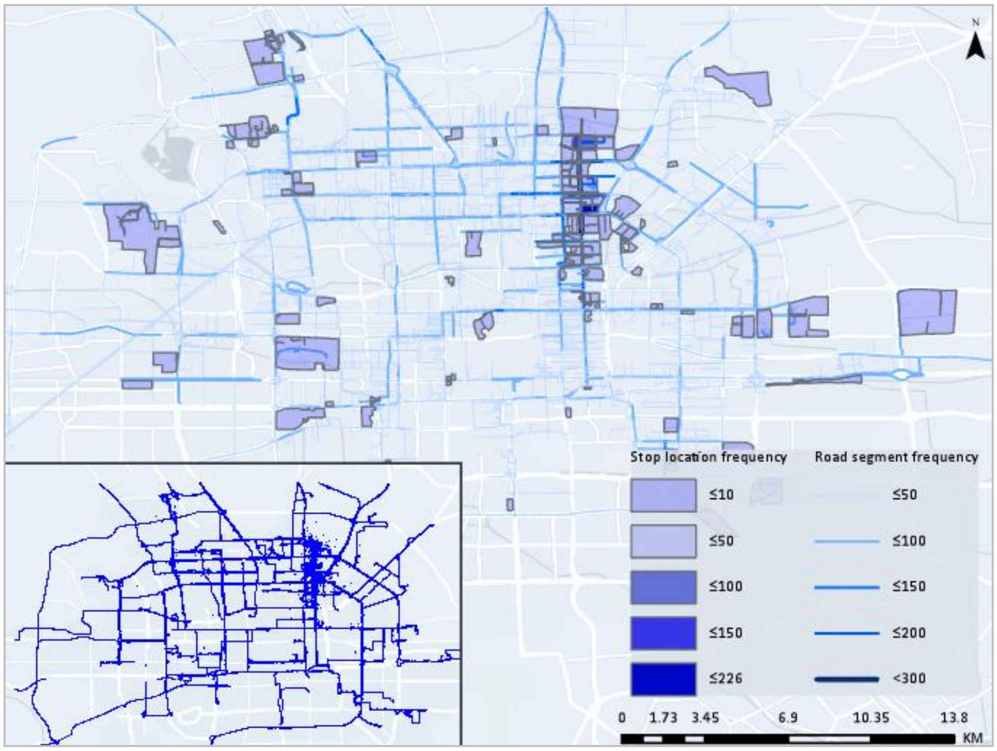
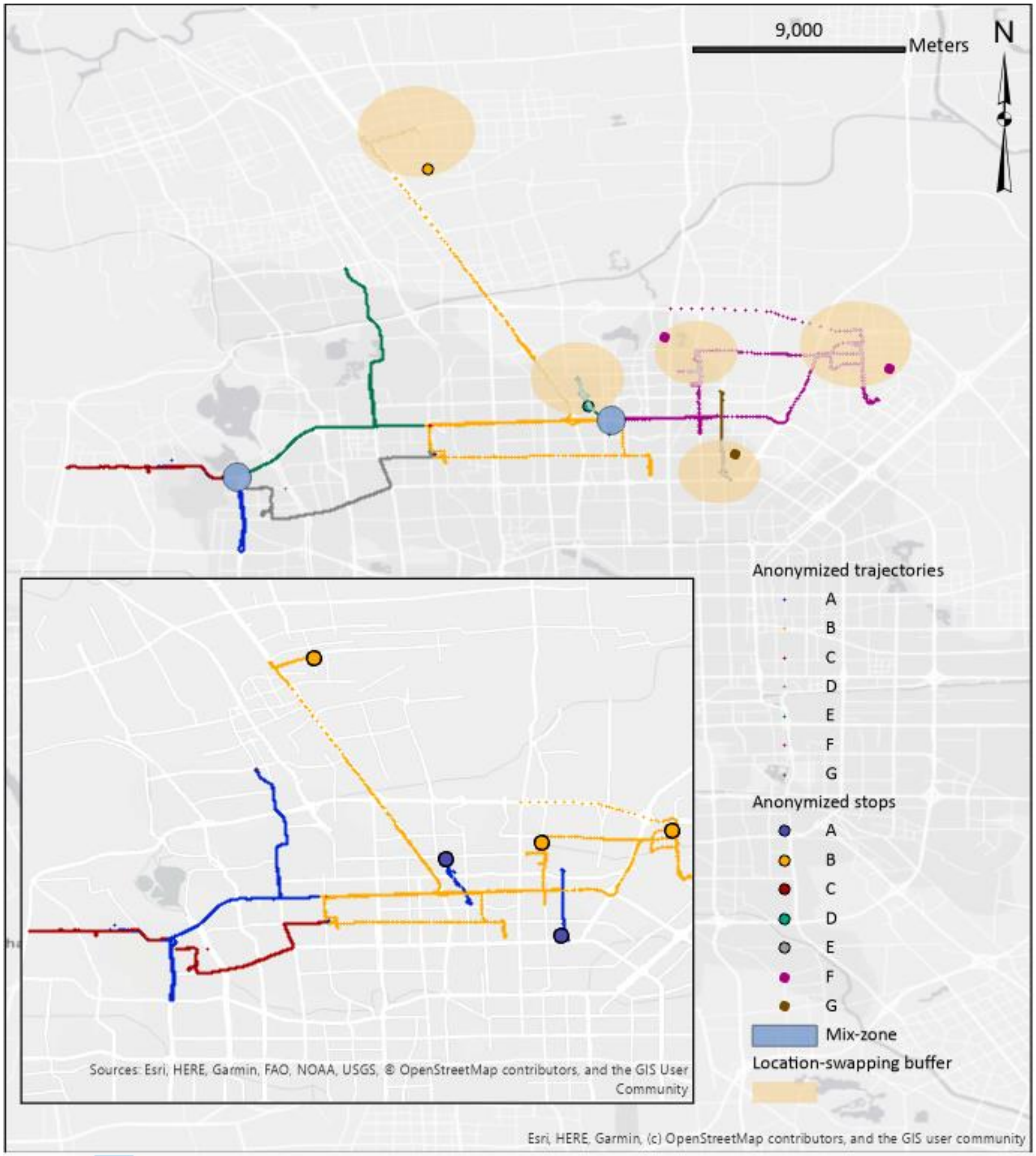


Figure 2.7. **Example of raw trajectories (inset maps) vs semantic trajectories (main maps) for a sample user of Geolife. Showing the frequency of stops provides a realistic and practical way for privacy protection.**

The output of this process was then used as the input for the anonymization framework. Anonymization started by relocating personal frequent locations to satisfy a user-defined privacy level sk, followed by temporal generalization. Sequentially, a pseudonym-swapping method was used to break the links between locations of the trajectories. The output of the whole process is shown in Figure 2.8 for user A and other co-located users on the same day. The radius of displacement buffer in the location swapping process varied to achieve an acceptable user-defined privacy level for personal frequent locations. To carry out the pseudonym swapping process, the time resolution of the dataset was generalized, and candidate co-occurred trajectories that passed close enough with regard to a spatial threshold, were renamed to a new random pseudonym. Finally, corresponding move points reaching stops within their displacement buffer and the corresponding move points of pseudonymized trajectories within mix-zone were removed.



Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community

Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

Figure 2.8. Three co-occurred original trajectories (inset map) on a specific day vs anonymized trajectories (main map) using when time resolution=6 hours, mix-zone radius=50 m, SKA=12.

2.6.5. Evaluation of proposed framework in terms of data utility and privacy

To assess the efficiency of the anonymization framework, estimating the relationship between privacy level (inversely proportional to the risk of user re-identification) and data utility based on average displacement distance and dissimilarity between original and anonymized datasets is vital. The following sub-sections measure privacy levels and data utility after anonymization.

Evaluating location privacy and data utility after location swapping

We started by investigating the balance between privacy level and data utility after the location swapping process. To assess the level of privacy, in this case, the probability of inferring locations related to users' habits, SKA was used and to assess the amount of lost data utility, average displacement distance was measured. To that end, after creating displacement buffers around personal frequent locations, when the *frequency* threshold was equal to 0.5, and swapping them to one of k potential locations, we measured the value of SKA. The value of SKA represents the number of candidate locations that fall within a displacement buffer centered at the masked location with a radius equal to the distance from the masked location to the original location.

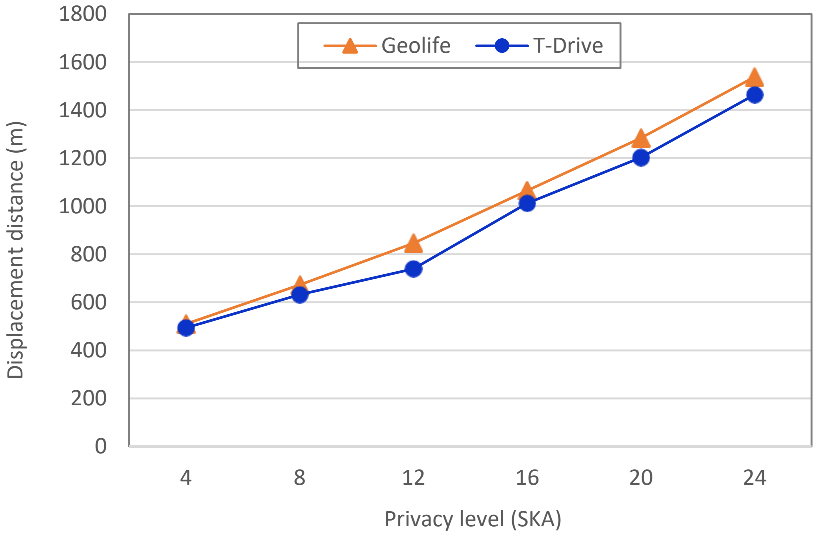


Figure 2.9. The relationship between SKA and average displacement distance of personal frequent locations after location swapping in Geolife and T-Drive.

Figure 2.9 highlights the average displacement distance required to achieve a certain level of privacy (SKA) which was dependant on local location density. However, there has not been a universal agreement between privacy protection researchers over a minimally acceptable level of privacy yet. Zhang et al. (2017), satisfied 20-anonymity for approximately 75% of residential addresses in Travis County, Texas, USA, and Wake County, North Carolina, USA considered as high-density areas, by swapping them less than 200 m. In their study, the local population density was used as an indicator for the number of residential addresses. In this study 44% and 23% of personal frequent locations in Geolife and T-Drive were related to residential parcels and the Beijing land-use dataset was not up-to-date and high-resolution. As described in Figure 2.9, an average displacement of 1285 m in Geolife and 1200 m in T-Drive resulted in 70% of locations achieving 20-anonymity. The increase of SKA from 12 to 24 with the increment of 4, resulted in increasing average displacement distance starting from 846 m by approximately 250 m for 70% of personal frequent locations in Geolife. In T-Drive, to achieve SKA greater than 12 with an increment of 4, an average displacement of about 730 m with an increment of 240 m was required for 70% of personal frequent locations. We chose 12-anonymity as an acceptable level of privacy in order to preserve data utility somehow.

Evaluating trajectory uniqueness and data utility after pseudonym swapping

To evaluate the effectiveness of the pseudonym swapping process, we examined the uniqueness of anonymized trajectories based on the number of exact spatiotemporal points required to uniquely characterize an entire trajectory. Although, this scenario, which is considered as the worst-case scenario, is less likely to happen in the real world and an attacker usually knows instead important locations of a user. Figure 2.10 shows the number of spatiotemporal points to identify a trajectory at a given time interval and distance threshold for both datasets. Reasonably, the number of points for characterizing a trajectory grows exponentially with trajectory length and the number of crossing trajectories.

Rossi et al., (2015) investigated the uniqueness of GPS trajectories in Geolife and showed that all users in the dataset were located with only two random spatiotemporal points. According to de Montjoye et al. (2013), only four spatiotemporal points were enough to uniquely identify 95% of the trajectories in an aggregated dataset of one and a half million users with a temporal resolution of 1 hour and a spatial resolution equal to that given by the carrier’s antennas (each antenna serves approximately 2000 inhabitants), and the most unique trajectories could be identified with only 11 points. They showed that the uniqueness of trajectories degrades as around the 1/10 power of their resolution through the spatial and temporal aggregation of the dataset. Therefore, in a greatly aggregated dataset with a temporal resolution of 7 hours and a spatial resolution of 5 cellular antennas (each antenna covers areas of approximately 0.15 km^2in cities), having 10 spatiotemporal points could re-identify 80% of trajectories.

In T-Drive, we assumed that two candidate trajectories were co-occurred if they passed at a distance of at most 50m in a 20-min interval, which is about six times bigger than the average sampling interval for time. In T-Drive Knowing 10 spatiotemporal points, cannot still disclose 55% of trajectories. The most unique trajectories could be identified with 50 points and only 3% of trajectories were unique in the dataset.

In Geolife, which is a dataset with a smaller number of users and more sparse trajectories, one of the time or spatial resolutions had to be generalized extremely in order to better reduce the uniqueness of trajectories. Since the focus of this paper was to preserve spatial patterns, we generalized time resolution to 6 hours and set the radius of the mix-zone to 50 m, which is about six times bigger than the average sampling interval of the dataset for distance (5~10 m). After pseudonym swapping, with having four spatiotemporal points, 60% of trajectories could not still be uniquely identified and only 10% of trajectories were unique in the dataset. Having 10 spatiotemporal points could uniquely reveal 70% of trajectories.

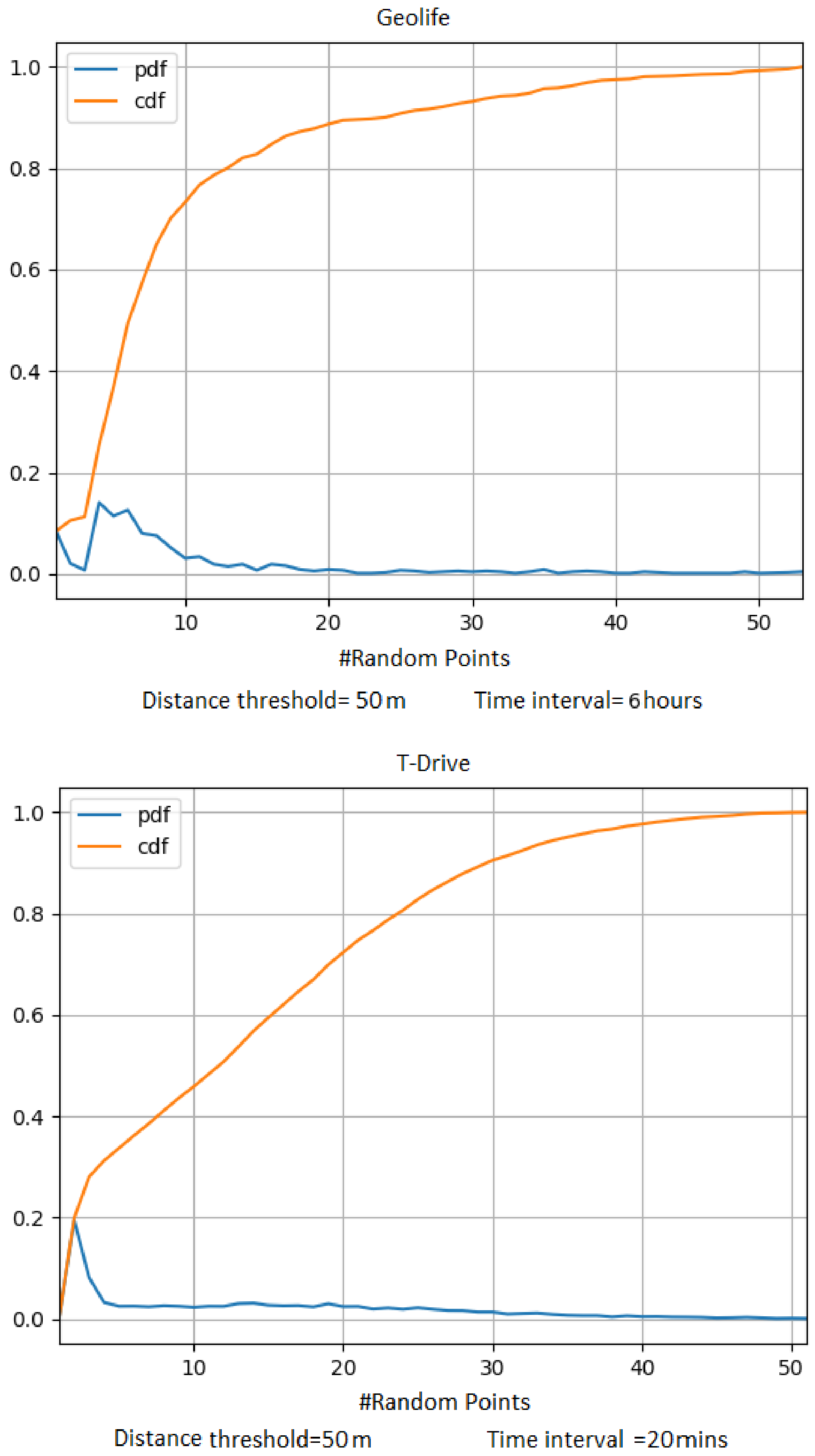


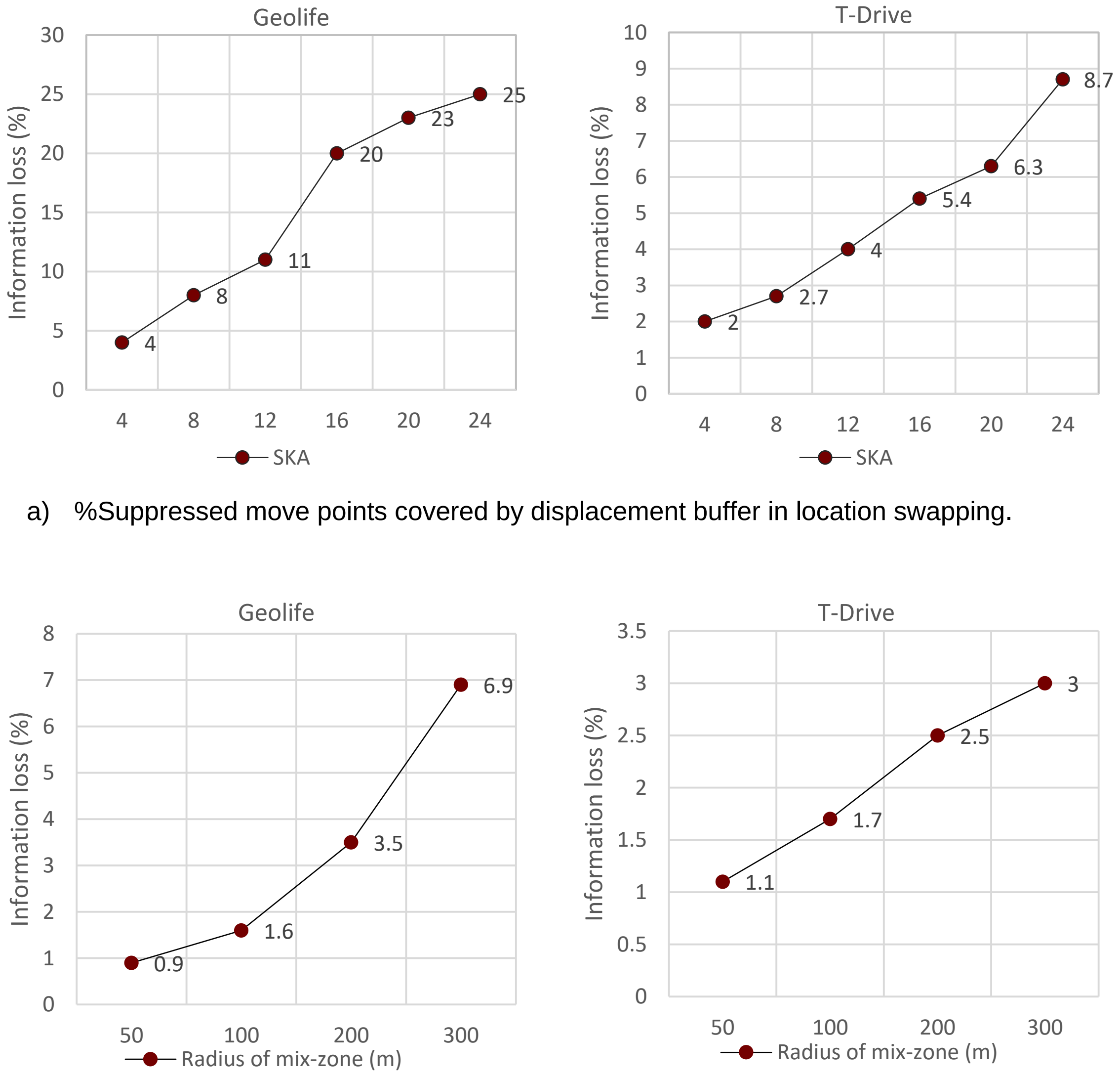
Figure 2.10. The number of required spatiotemporal points to uniquely re-identify trajectories after the pseudonym swapping.

Information loss

We also measured information loss caused by anonymization in terms of the percentage of suppressed move points, either covered by dynamic mix-zones in the pseudonym swapping process or placed within the displacement buffers in the location swapping process at different privacy levels as follows:

$$Suppression_{move\ points} = \frac{\sum number\ of\ suppressed\ move\ points}{\sum number\ of\ move\ points} \qquad (1)$$

Figure 2.11a highlights the percentage of suppressed move points covered by displacement buffers while varying the value of SKA from 4 to 24 in location swapping for both datasets. Seemingly, to satisfy a higher value of SKA, a larger displacement buffer was required and an increase in suppression of move points was resulted. It can be seen that in both datasets when SKA ≤12, there is little difference with regard to suppression, with approximately 10% of Geolife move points and 5% of T-Drive move points were suppressed. When SKA reaches 16, nearly twice the number of move points in Geolife were suppressed, and information loss sharply increased. Hence, in this paper SKA=12 is considered to be an acceptable value for Geolife that protects a high level of privacy while maintaining data utility to a good extent. In T-Drive when SKA=24, less than 10% of points were removed which is a result of high sampling interval in distance.



a) %Suppressed move points covered by displacement buffer in location swapping.

b) %Suppressed move points covered by mix-zones in pseudonym swapping.

Figure 2.11. Percentage of suppressed move points in the anonymization process.

Figure 2.11b shows the percentage of suppressed move points covered by mix-zones while varying the radius of mix-zone or distance threshold to 50, 100, 200, and 300 m in both datasets. Seemingly, a larger radius caused an increase in the suppression of covered moves. It can be seen that when the radius was equal to 300 m, approximately 7% of Geolife move points suppressed, while in T-Drive nearly 3% of move points were removed since the sampling interval of Geolife (5~10 m) was smaller than T-Drive (623 m). To reduce the amount of information loss and preserve spatial pattern, the radius of mix-zone was set to 50 m for both datasets while suppressing only 1.1% of T-Drive move points and 0.9% of Geolife move points. To reduce the uniqueness of Geolife trajectories, we instead generalized temporal resolution extremely.

In summary, this method results in alternative semantic trajectories that parallel the meaning but not the exact position of the locations related to users' habits while de-personifying moves. By using this combined method, we offer a means of 1) protecting both location and trajectory privacy, 2) preserving the original pattern in locations and movements without disclosing individual continuous paths, and 3) sharing de-personalized micro-trajectories instead of aggregated trajectories for the benefit of researchers and the benefit of the public at large.

2.7. Discussion and conclusion

Anonymizing trajectory data at publishing time is an ongoing problem, and this is evident from the numerous complex techniques devised so far. Many studies have shown that users' re-identification can still happen when coarsening spatiotemporal resolution (Chang et al., 2018; de Montjoye et al., 2013; Gambs et al., 2014; Kondor et al., 2018; Zang & Bolot, 2011), geomasking (Fiore et al., 2020), and trimming (Abul et al., 2008; Andrienko et al., 2009; Baik Hoh et al., 2006), of such person-specific data. While many anonymization methods exist, trajectories can still be uniquely identified using few spatiotemporal points and roughly half of the methods treat all locations similarly regardless of the fact that locations related to users' habits have higher identity disclosure risk [89].

Our study aimed to prevent true detection of personal frequent locations and unlink different activities of a user and reduce the uniqueness of the user's trajectories with a little modification to a dataset while preserving the semantics of visited locations and shape of trajectories. The results show that the combination of pseudonym swapping and temporal generalization can reduce the trajectory uniqueness, and applying location swapping on only frequent locations can maintain the integrity of trajectories to a great extent while effectively preserving important locations' privacy.

The proposed framework is resistant against reverse-engineering attacks when publishing statistics required for map anonymization, due to the randomness during the anonymization process. This anonymization can successfully preserve the spatial resolution and produce de-personalized micro-trajectories instead of the aggregated trajectories produced by the grid-masking method (Y. Wang & McArthur, 2018). It can also reduce information loss compared to methods that use the suppression of sensitive location (Sila-Nowicka & Thakuriah, 2016). The anonymized dataset can be efficiently used for many policy-making applications including location-based analysis (since non-frequent locations remain in the same positions and frequent locations parallel their meaning after relocation) or mobility-based analysis that examines users' mobility patterns such as transportation network modeling, urban planning.

Of course, this approach is not without shortcomings. One shortcoming of this approach is that the stop detection and the level of privacy in the location swapping process rely upon the availability of a high-resolution and up-to-date land-use or building map. Given the poor state of OpenStreetMap, especially in Beijing 10 years ago, the land-use dataset used in this paper may result in inaccurate evaluation. Another shortcoming of this framework is that we presume that false identification is not of concern, although in a sensitive dataset one can attribute a sensitive attribute to a wrong household. Finally, the output dataset cannot be used for personalized trajectory analysis which requires the movements of a particular person because we unlink users from their trajectories.

Our work points to avenues for future work. The stop detection algorithm used in this study was based on the same value of time, distance, and speed thresholds in different locations, which may not be realistic since, for example, minimum possible stop times are different in different locations (restaurant vs cinema). Other algorithms for producing micro-trajectories while obfuscating the mobility patterns of users in order to prevent attacks when knowing the mobility model and movement patterns of users are worth examining for future work.

In closing, our combined privacy protection framework has advanced privacy protection by addressing certain vulnerabilities in previous methods. Our approach of protecting personal frequent locations and unlinking locations of a user makes an important contribution to addressing the critical need to protect users' privacy.

So far, a privacy protection framework has been designed and developed. However, adoption and utilization of this framework in practice might be difficult and offer challenges for the average user, since it requires knowledge of coding and an extended workflow that may be time-consuming, and more importantly prone to errors, preventing data owners from publishing their sensitive data. Rather than the contribution to the design of a new privacy protection technique, our effort is directed at making existing techniques more accessible and usable by a larger pool of data analysts. Therefore, in the next chapter, an easy-to-use toolbox is developed for novice GIS users to protect the privacy of GPS trajectory data without knowledge of coding and anonymization.

Chapter 3. PrivacyProtection: An ArcGIS toolbox for processing and protecting the privacy of semantic trajectories

3.1. Abstract

Recently, monitoring the spread of the SARS-CoV-2 coronavirus which causes COVID-19 has become an emergent and crucial task, undertaken through tracking individuals' movements and physical distance. However, publishing individuals' movement data, known as "trajectory data," leads to potential privacy breaches when trajectories are combined with additional information from other external sources. To protect privacy, an anonymization process, which transforms data to remove sensitive information, can be performed but this can be at the cost of a decrease in data utility. Generating tools to ease this process is crucial but poorly utilized in academic geography. In this paper, we describe the development of a privacy protection toolbox in an ArcGIS Pro environment with an easy-to-use interface, that performs trajectory data processing and anonymization. The main objective of this toolbox is to construct semantic trajectories from raw GPS points based on user-defined values, anonymize the trajectories using the combination of suppression with generalization or grid-masking techniques, and then evaluate the resulting balance between privacy and utility by adjusting anonymization parameters. It executes each of these processes on thousands of points in seconds. This toolbox offers two advantages to users: 1) GIS users can download and use the easy-to-use toolbox and graphical user interface without deep knowledge of anonymization methods; and 2) due to being accessible and re-usable, it enables privacy protection researchers to reproduce and compare their empirical works.

3.2. Keywords

Privacy protection, ArcGIS Python toolbox, Semantic trajectory construction, Generalization, Suppression, Grid-masking

3.3. Introduction

The growing collection and use of GPS data have provided an opportunity for proximity and contact tracking (Kapa et al., 2020) to monitor physical distancing and to aid in the control of the COVID-19 pandemic (Demirag & Ayday, 2020). However, due to the inherent vulnerability of this data, publishing GPS data can constitute a major privacy violation, since it can disclose critical information about people's lives (Basu et al., 2014; Monreale et al., 2010; Y. Wang & McArthur, 2018). For instance, using easy-to-use maps and reverse geocoding tools, it is easy to reveal the identity and home addresses of COVID-19 patients on a map representing their movements. However, previous research has studied the significance of outbreak and disease surveillance without giving (enough) attention to this privacy issue (Bhatia et al., 2019; Carneiro & Mylonakis, 2009). Kapa et al. (2020) emphasized the importance of creating a balance between individuals' privacy and public health benefits when using contact tracking tools and analyzing the movements of patients and healthy users. Consequently, there is a need to provide a better understanding of movements to be able to warn healthy users who may have contacted with an infected individual of potential exposure, and one that is balanced in favour of individual privacy protection.

Increased concern about identity and sensitive information disclosure when publishing data directly was the impetus for the development of a set of methods, called privacy-preserving data publishing (PPDP), which provides a set of approaches for the publication of data that preserves privacy while ensuring the availability of valid data for analysis (Fung et al., 2010). There is a wide range of PPDP methods described in the academic literature, ranging from simply removing unique identifiers to transforming/coarsening semantic, spatial, and temporal accuracy of data (Adrienko & Adrienko, 2011; Andrienko et al., 2009; Nergiz et al., 2008; Samarati & Sweeney, 1998) and integrating contextual information for better preservation (Han & Tsai, 2015; Monreale et al., 2011; Tu et al., 2019; Y. Wang & McArthur, 2018).

However, there is limited adoption and utilization of these techniques in practice due in part to the difficulty of interpretation and implementation (Swanlund, Schuurman, & Brussoni, 2020), resulting in publishing users' data without taking action for privacy protection. For example, Hasan et al., (2017) stated that many bike-sharing companies have made their user data publicly available after removing user identity that led to the disclosure of user’s privacy and movement patterns. This leaves a significant gap in the current body of scientific knowledge, where privacy protection is discussed without considering how it might be utilized by a wider range of users who deal with spatial data. Boeing (2020) demonstrated the fundamentally important role that accessible tools play in progressing science and theory. Given the importance of increasing the accessibility of these tools, we elected to focus our attention on making existing techniques more accessible and usable by a larger pool of data analysts, rather than designing a new PPDP technique.

ArcGIS is a well-known software for GIS and spatial analysis (Conolly et al., 2006). However, using ArcGIS to perform spatial data processing, interpretation, and privacy protection offer challenges for an average user. The use of ArcGIS entailing privacy protection requires knowledge of coding and an extended workflow (see Figure 3.1) that may prove time-consuming, and more importantly, may be difficult to implement and prone to errors. This work describes the process of developing an ArcGIS python toolbox designed to provide several tools for novice GIS users to protect the privacy of GPS trajectory data. The toolbox processes GPS data by constructing a meaningful structure of trajectories using a stop detection method. It also infers users’ habits and preferences from their movements through finding frequent locations of users, a pattern that serves as their mobility signature. Lastly, it protects the privacy of personal frequent locations and trajectories by making use of suppression, spatiotemporal generalization, and grid-masking techniques.

The paper begins by reviewing the existing PPDP techniques in the area of trajectory data, as well as the relevant existing GIS tools and applications. Next, the paper turns to the design and architecture of the toolbox, illustrating its performance when dealing with real-world GPS data. Finally, the paper discusses the toolbox's strengths and limitations, as well as future research directions that would ease the real-world adoption of privacy protection techniques, before providing a brief conclusion.

3.4. Related works

3.4.1. Privacy protection of GPS trajectories at publishing time

The concept of PPDP has been well studied in the area of location-based services (Andrés et al., 2013; Ishikawa & Yoshiura, 2020; C. Xu et al., 2020), where the goal is real-time anonymization of a location or query of an issuer. However, the purpose of this paper is anonymization of an entire trajectory dataset at once. In general, there are five main techniques of PPDP methods in the area of trajectory publishing including perturbation and clustering, generalization, suppression, adding noise, and pseudonym changing/swapping. The common purpose of these methods is reducing user identity and sensitive information disclosure by satisfying a privacy criterion such as k-anonymity (Sweeney, 2002) and its extensions (Poulis, Skiadopoulos, et al., 2013; R. C.-W. Wong et al., 2006), l-diversity (Machanavajjhala et al., 2006), t-closeness (N. Li et al., 2007), and differential privacy (Dwork, 2006). Although finding an optimal k-anonymization is NP-hard (Aggarwal et al., 2005; Meyerson & Williams, 2004), it remains a realistic privacy model with practical relevance.

Perturbation consists of displacing trajectory points to new spatial coordinates at a random distance and random angle (either uniformly or using Gaussian noise) within a buffer with radius r centered on the original point. It can further be combined with clustering to merge co-located trajectories within a specific time window (Abul et al., 2008; Lin et al., 2010; Nergiz et al., 2008). However, they cannot preserve the semantics of trips and association with road links, especially when providing a high level of privacy (Yin et al., 2015). Adding random or restricted noise to a dataset adhere to differential privacy and merging trajectories to confuse an attacker about the presence of a particular trajectory in the dataset while providing statistics for specific data mining tasks such as frequent sequential pattern mining (Chen et al., 2012; Deldar & Abadi, 2018, 2019; Dwork, 2006; Jiang et al., 2013; Kellaris et al., 2014). However, differential privacy only focuses on preserving the privacy of trajectories regardless of the relation between different locations of a trajectory and privacy of locations. Pseudonym changing/swapping was inspired by the concept of Mix-zones (Freudiger et al., 2007). Mix-zones are spatial areas where the locations of individuals are lost and such that each trajectory that exits the mix-zone will have a different pseudonym than the one when entering a mix-zone. Based on mix-zone, trajectory swapping was introduced with the aim of exchanging trajectories of two different individuals/pseudonyms that pass within mix-zones for a certain period of time (Salas et al., 2018). Suppression eliminates those spatiotemporal samples that do not satisfy privacy constraints such as k-anonymity. Some researchers used only suppression techniques (Chen et al., 2013; Mohammed et al., 2010; Terrovitis & Mamoulis, 2008) to iteratively suppresses violating spatiotemporal samples until meeting the privacy constraint. However, using suppression alone may cause a huge information loss and needs to be combined with other techniques such as generalization for better protection. Generalization partitions the underlying area of a trajectory into appropriate sub-areas which are temporally and spatially disjoint by coarsening the spatial and temporal resolution, and merges co-located and co-occurred users into the sub-areas. Although, the combination of generalization and suppression might cause information loss, e.g., loss of data granularity or loss of data samples, the produced data can be useful for a wider range of applications (Gramaglia et al., 2017) such as pandemic controlling, transportation research, and urban planning. On the other hand, visualizing trajectories in their raw form, with overlapping points, makes the display excessively cluttered (Adrienko & Adrienko, 2011), and using generalization for transforming atomic data into a more simplified one can be helpful.

There is a rich history of literature on leveraging generalization to preserve the privacy of trajectory data. In spatial generalization, locations are replaced by the closest street, postal code, census area, and so on. For example, some studies have worked on partitioning a geographical area of a trajectory into same-sized grids using a fixed grid hierarchy (Yarovoy et al., 2009), different-sized Voronoi tessellations (Andrienko et al., 2009), or different-sized clusters of semantically diverse locations around stops (Huo et al., 2012). Monreale et al. (2011) applied a spatial-semantic generalization using predefined location taxonomy to restrict the probability of extracting sensitive locations from non-sensitive ones adhere to a proposed privacy criterion named c-safety. Cicek, Nergiz, and Saygin (2014) proposed a similar notion, named p-confidentiality, to upper-bound the probability of extracting a sensitive location from a trajectory by replacing sensitive locations with sensitive regions which contain a group of POIs around sensitive locations and ensuring location diversity within sensitive regions.

Importantly, it has been demonstrated that despite data generalization, repeated moving objects within an anonymity region can still be re-identified (Chang et al., 2018;

F. Xu et al., 2017; Zang & Bolot, 2011) when having background knowledge about users' home addresses (Krumm, 2007), home-workplaces (Golle & Partridge, 2009), and most frequently visited locations (Zang & Bolot, 2011). What is certain is that some locations can easily reveal users' identity and sensitive information—even in very large populations—due to the regularity of movement patterns (Golle & Partridge, 2009; Krumm, 2007; F. Xu et al., 2017; Zang & Bolot, 2011). These realities all further emphasize the importance of frequent locations and the necessity of effective protection for these locations rather than all sample points. Given this regularity, a wide range of more sophisticated anonymization techniques focus on distinguishing important locations of users according to their frequency of occurrence and staying time to treat these important locations differently which can enhance both privacy protection and data utility. For instance, Sila-Nowicka & Thakuriah, (2016) suppressed important locations and kept the remaining data intact for road network-based analysis. Similarly, Wang & McArthur (2018) only cloaked top N frequented and long-stayed locations, such that home-workplaces were blurred with k nearest buildings and other important locations were blurred with k nearest POIs. Recently, Rajesh et al. (2019) extracted the important locations of users' trajectories based on the amount of time spent and created Minimum Bounding Rectangular zones of user-defined size with at least one important location and ''n'' number of non-important locations.

However, trajectory anonymization is a relatively young discipline and the anonymization techniques that have been developed to date can preserve privacy with minimal information loss. Their biggest shortage is not in technical innovation, but rather in knowledge translation and there remains a need for an accessible and easy-to-use toolbox with comprehensive documentation that is specialized for trajectory anonymization.

3.4.2. GIS tools and applications for geoprivacy protection

Despite a variety of privacy protection techniques available in academic literature, limited adoption of these techniques has been observed due to users' lack of knowledge and expertise in their interpretation and implementation in a GIS framework. However, there has recently been attention to the need to develop more accessible tools and applications with a focus on privacy protection.

For instance, the Geographic Privacy-aware Knowledge Discovery project (GeoPKDD) (Giannotti, Nanni, Pedreschi, & Renso, 2009) was designed to integrate geoprivacy protection and knowledge mining from spatial data into a unified framework. They adopted the notion of k-anonymity and formalized the concept of user re-identification in the context of patterns, instead of data, and identified all potential attacks that might happen when releasing extracted patterns. Their main focus was privacy-preserving data mining (PPDM) rather than PPDP. In PPDP the original data after anonymization are published without considering the final usage of the data, while in PPDM the discovered patterns and data mining results are published.

In the context of PPDP, Gambs, Killijian, and Cortez (2010) developed an open-source software named GEPETO to anonymize trajectories of individuals using simple anonymization techniques such as pseudo-anonymization, downsampling, random perturbation, and clustering. GEPETO used downsampling to summarize a set of close GPS points within a time window with a short Euclidean distance from each other rather than Manhattan distance. However, this method significantly reduced the data utility of the dataset. Then it applied random perturbation on all recorded points without maintaining the semantics of trips and the association of move points to the road segments. Their methods were simple and not sufficient for trajectory data which have complexity in terms of structure. Recently, Swanlund et al. (2020) developed a client-side application called MaskMy.XYZ to mask location point data such as health records and crime data by applying the donut masking method (Hampton et al., 2010). This application cannot be used for trajectory data since it randomly relocates single points without considering the semantics of locations and their links with the road segments.

3.5. Method: toolbox process and results

This paper proposes a new toolbox, called the PrivacyProtection toolbox, that mitigates the difficulty of trajectory anonymization to a good extent and is run locally to reduce security concerns when transferring data. PrivacyProtection toolbox works in the ArcGIS environment to construct semantic trajectories, derive important locations of individuals, and protect their privacy and the privacy of entire trajectories while maintaining low computational cost. This toolbox provides privacy for trajectories otherwise being published without anonymization. We offer different anonymization techniques including suppression, generalization, and grid-masking, allowing users to determine their desired relationship between the level of privacy and retaining data utility. End-users of this toolbox can be data publishers who do not preserve the privacy of trajectory data due to a lack of time, anonymization knowledge, and/or coding skills. By developing this toolbox, we aim to enable the better preservation of privacy when publishing trajectories. In this section, we provide an overview of the required extensions, design, and architecture of the toolbox and consider its merits and limitations.

3.5.1. Required extensions

To distribute a technology, there have been numerous discussions over the advantages of developing open-source and platform-independent tools and applications. However, implementing solutions within a popular commercial environment is beneficial because these environments can facilitate customization, allow easy preparation of data, and provide a user-friendly graphical user interface (GUI) (Charleux, 2015). PrivacyProtection toolbox was developed in the ArcGIS Pro 2.4 environment and was written by Python, which is a general-purpose programming language that provides excellent packages in dealing with spatial data such as Arcpy. Developing a Python Toolbox enabled us to develop additional modules in numerous capacities of Python such as NumPy and pandas for data manipulation and analysis.

3.5.2. Design and architecture

One of the most important merits of an ArcGIS toolbox is that it operates completely on the user’s local system. Therefore, data publishers can safely analyze, anonymize, and visualize their sensitive data without any privacy or security concerns. Another merit of the ArcGIS toolbox is that it provides a user-friendly GUI, enabling users to determine the input data and adjust parameters until achieving a satisfying privacy level and data utility.

A flowchart outlining the ideal workflow for our PrivacyProtection toolbox is shown in Figure 3.1. The PrivacyProtection toolbox is subdivided into three main processes performed by two toolsets and six tools. The first step consists of processing raw GPS trajectory data and extracting the meaningful structure of daily trips that delineate the starting point, stops, and moves, using a stop detection method by means of the *Trajectory processing* tool. The next step consists of two tools under the *Location interpretation* toolset, which can be used to identify the Most Frequented Locations (MFLs) of each user. The last step is assigned to *Trajectory anonymization,* which comprises three tools: suppression, generalization, and grid-masking. Using the suppression tool, a number of MFLs and their neighbouring points within a buffer of a user-defined radius are removed and the amount of information loss is measured. Using the generalization tool, the spatial and temporal resolution of the dataset is coarsened by replacing the exact timestamp with an approximate time interval and replacing stop points with larger stay zones with k-locations and move points with larger travel zones within a user-defined radius. Adapted from the anonymization framework designed by Wang & McArthur (2018), the grid-masking tool generalizes the trajectories from points to raster cells and aggregates either traveling time or the number of users within a specific time interval to each grid cell. Tools in the *Location interpretation* and *Trajectory anonymization* toolsets can be run independently from one another.

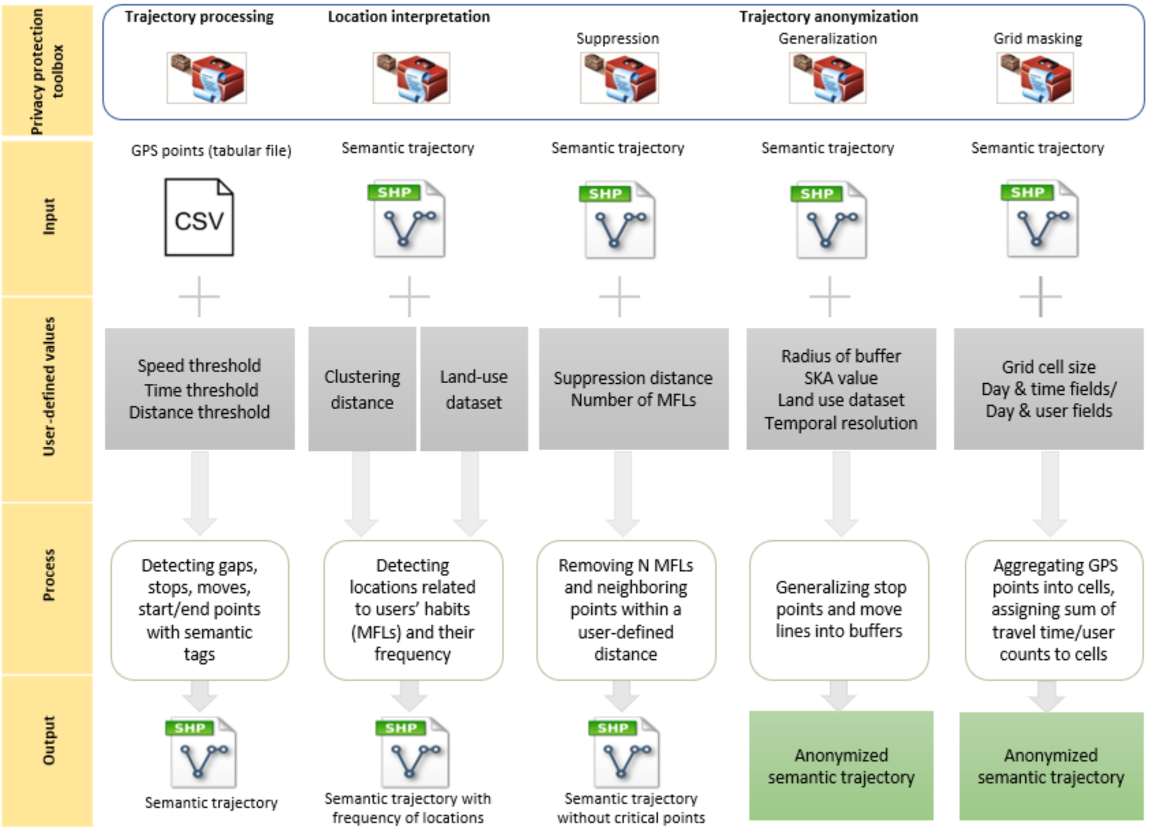
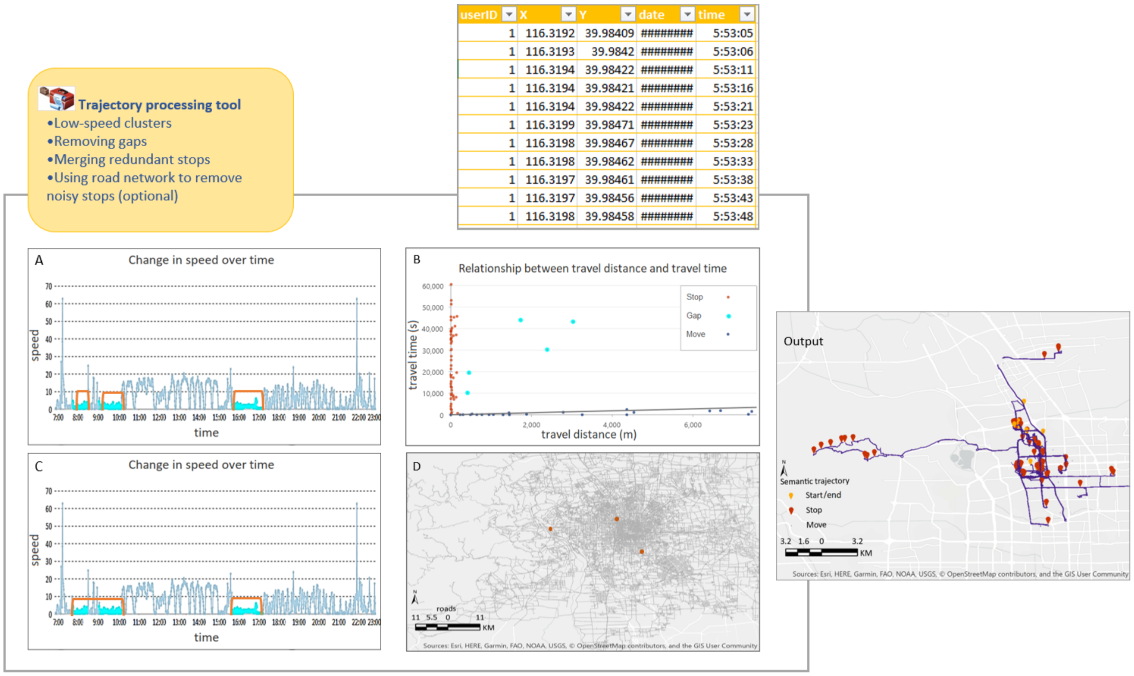

Privacy protection toolbox
Input
User-defined values
Process
Output
Trajectory processing
Location interpretation
Trajectory anonymization
Suppression
Generalization
Grid masking
GPS points (tabular file)
Semantic trajectory
Semantic trajectory
Semantic trajectory
Semantic trajectory
CSV
SHP
Speed threshold
Time threshold
Distance threshold
Clustering distance
Land-use dataset
Suppression distance
Number of MFLs
Radius of buffer
SKA value
Land use dataset
Temporal resolution
Grid cell size
Day & time fields/
Day & user fields
Detecting gaps, stops, moves, start/end points with semantic tags
Detecting locations related to users' habits (MFLs) and their frequency
Removing N MFLs and neighboring points within a user-defined distance
Generalizing stop points and move lines into buffers
Aggregating GPS points into cells, assigning sum of travel time/user counts to cells
Semantic trajectory
Semantic trajectory with frequency of locations
Semantic trajectory without critical points
Anonymized semantic trajectory
Anonymized semantic trajectory

Figure 3.1. Example of semantic trajectory extraction and anonymization through the PrivacyProtection toolbox. This theoretical framework involves all the tools of the toolbox.

Semantic trajectory construction tool

As a first step, the tool named *Trajectory processing,* as shown in Figure 3.2 is used to structure raw GPS points and assign semantic tags to them to create semantic trajectories. Semantic trajectories have two main components—stops and moves. Constructing semantic trajectories normally starts with detecting stops. To detect stops, it is necessary to define parameters and thresholds depending on specific purposes. The tool enables users to load a tabular file (e.g., CSV file) containing raw GPS points with X, Y, userID, date, and time fields for extracting the daily trips of each user; to define possible values of speed, distance, and time thresholds for a stop; and choose either a geographic or a projected coordinate system. The default values for speed, distance, and time thresholds are 1.2 m/s, 300m, and 600s, respectively. An additional feature is the ability to add road network data as an optional input to remove noisy stops, such as the stops at traffic lights or in traffic jams that appear on the road network. This feature can be useful for finding only meaningful stops where a user may perform an activity.

Once processed, trajectory data with semantic tags indicating their structure as start-stop-move, coupled with information about speed and traveling/staying time, are exported as a shapefile in the user-defined projection. One of the advantages of this tool is that it only requires a tabular file.



userID	X	Y	date	time
1	116.3192	39.98409	#########	5:53:05
1	116.3193	39.9842	#########	5:53:06
1	116.3194	39.98422	#########	5:53:11
1	116.3194	39.98421	#########	5:53:16
1	116.3194	39.98422	#########	5:53:21
1	116.3199	39.98471	#########	5:53:23
1	116.3198	39.98467	#########	5:53:28
1	116.3198	39.98462	#########	5:53:33
1	116.3197	39.98461	#########	5:53:38
1	116.3197	39.98456	#########	5:53:43
1	116.3198	39.98458	#########	5:53:48

Trajectory processing tool
- Low-speed clusters
- Removing gaps
- Merging redundant stops
- Using road network to remove noisy stops (optional)

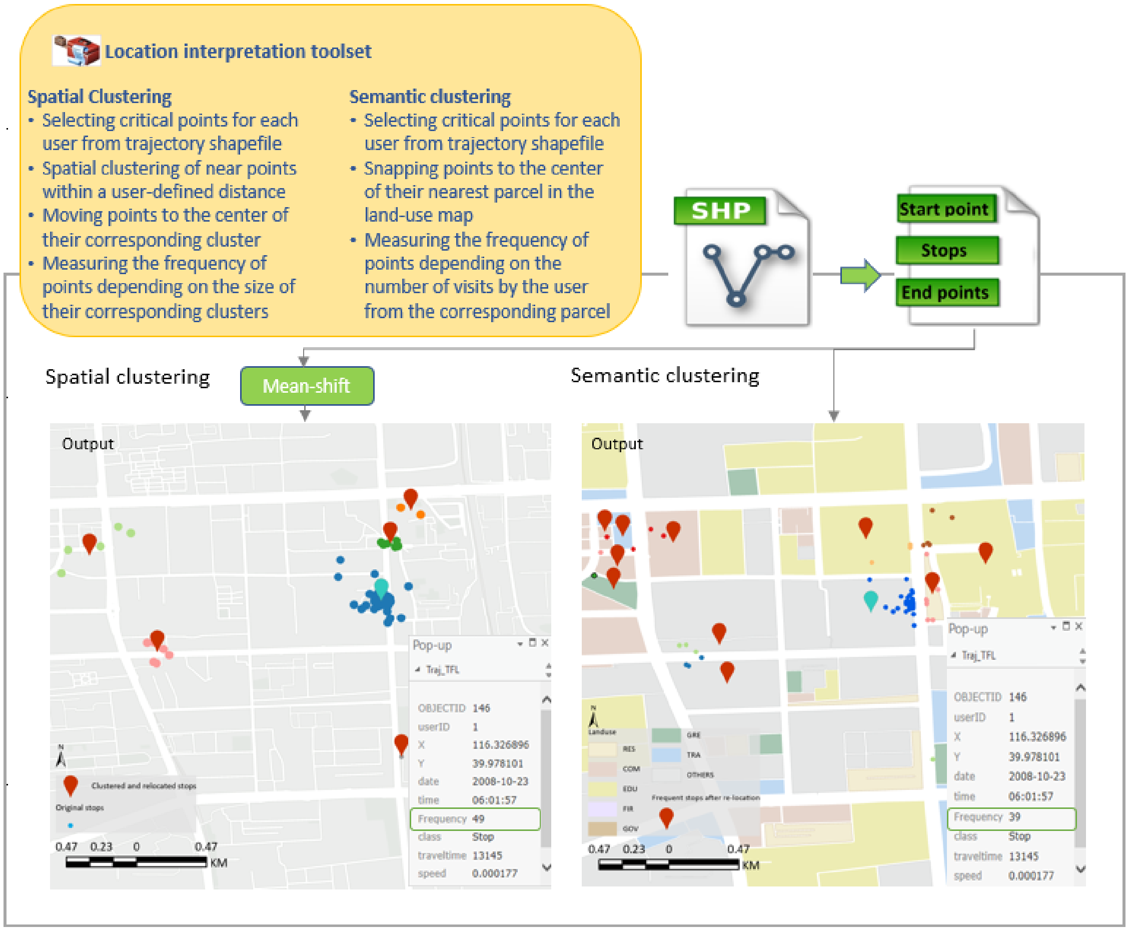
A
Change in speed over time
speed
time
B
Relationship between travel distance and travel time
travel time (s)
travel distance (m)
Stop
Gap
Move
C
Change in speed over time
speed
time
D
roads
KM
Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community
Output
Semantic trajectory
Start/end
Stop
Move
KM
Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community

Figure 3.2. The process of semantic trajectory construction using *trajectory processing* tool through four interconnected process.

The *Trajectory processing* tool detects stops in four steps. The first step is to filter out the dataset by removing noisy GPS points which move at a speed greater than 45 m/s. The second step is to split trajectories into user-level daily trips and for each trip, to detect candidate stops by finding clusters of low-speed points that match user-defined spatiotemporal windows. In this step, there are gaps that might be mistaken as stops. Gaps are the amount of time and distance when GPS points are lost and can be caused by interrupting GPS satellite signals in dense areas or when a vehicle enters a tunnel or due to the GPS logger battery running out. Two consecutive GPS points that have a time difference of five times larger than the average sampling time interval of the dataset and spatial distance greater than the staying distance threshold are considered as the start and end of a gap. The third step is to merge redundant stops, stops at neighbouring time windows that belong to the same stop; in our process stops that are 100 meters and 10 minutes apart, are merged to a single stop. Once stops are detected, the last step is optionally conducted to remove noisy stops where the user does not perform an activity by removing stops that are within a distance of 5 meters from the road centerline in the road network dataset, which must be provided as an input. We found that the Trajectory processing tool is able to reconstruct a semantic trajectory dataset from 55,000 points over 22 days in 1 minute and 22 seconds and from 110,000 points over 43 days in 2 minutes and 15 seconds. A general trend was that doubling of point counts caused execution times to double. Performance testing was repeated, and we found that when a road network dataset to remove noisy stops was added, the results respectively increased by 15 and 30 seconds.

Location interpretation toolset

Once the semantic trajectories have been extracted, they can be analyzed by the second toolset, named *Location interpretation.* A particular characteristic of human spatial behaviour is being periodic and regular such that individuals frequently visit the same location and have the same movement patterns. To better preserve the privacy of individuals it is important to identify personal MFLs, which can disclose personal preferences and habits with higher probability, and impose stronger privacy protection on them. To find MFLs, we focused on measuring the frequency of occurrence of start and stop points, as critical points. Obviously, the start points reveal sensitive locations of a user such as their home address, and combined with stops, this can expose sensitive information about the purpose of a trip such as home-work commuting. As detailed in Figure 3.3, this toolset provides two tools—spatial clustering and semantic clustering—to identify the meaningful locations with the highest frequency for each user.



Location interpretation toolset

Spatial Clustering
- Selecting critical points for each user from trajectory shapefile
- Spatial clustering of near points within a user-defined distance
- Moving points to the center of their corresponding cluster
- Measuring the frequency of points depending on the size of their corresponding clusters

Semantic clustering
- Selecting critical points for each user from trajectory shapefile
- Snapping points to the center of their nearest parcel in the land-use map
- Measuring the frequency of points depending on the number of visits by the user from the corresponding parcel

Pop-up
Traj_TFL
OBJECTID 146
userID 1
X 116.326896
Y 39.978101
date 2008-10-23
time 06:01:57
Frequency 49
class Stop
traveltime 13145
speed 0.000177

Pop-up
Traj_TFL
OBJECTID 146
userID 1
X 116.326896
Y 39.978101
date 2008-10-23
time 06:01:57
Frequency 39
class Stop
traveltime 13145
speed 0.000177

Figure 3.3. **The process of detecting frequent critical points and measuring the frequency of their occurrence for each user using the spatial clustering and semantic clustering tools. These are the outputs that would be seen by the user.**

The spatial clustering tool was developed to cluster neighbour start and stop points of each user, using the mean-shift algorithm, and to move corresponding points of a cluster to the center of the cluster. The mean-shift algorithm was used since it only requires one input including cluster distance. The spatial clustering tool allows users to load a shapefile containing GPS points with semantic tags (the output of the previous step) to extract the starts and stops of each user and to define a clustering distance (200m as a default value). In the next step, the semantic clustering tool is used to snap the center of clusters to the nearest meaningful locations/parcels in the land-use dataset. In this case, the frequency of occurrence of a location is measured according to the number of visits by the user from the corresponding parcel. This tool requires a land-use dataset as an input. The output of this tool is a shapefile, the same as the input, in which frequent critical points are displaced to a location in the land-use map with an additional field indicating the frequency of their occurrence for each user.

During testing, the spatial clustering tool was able to do the process for a dataset with 100,000 GPS points and 200 starts/stops in particular, in just 11 seconds when the radius of clustering was selected as 200m. The performance test was repeated using the semantic clustering tool and adding a land-use dataset, instead of clustering distance, and it could run the same process in 2 minutes and 20 seconds. It is important to note that higher processing time resulting from the second tool is worthwhile since it ensures that personal frequent locations have the same activity and semantics.

Trajectory anonymization toolset

Suppression tool

For the core anonymization procedure, the first tool in the *Trajectory* anonymization toolset—the suppression tool—was developed to remove a number of personal frequent locations, as well as their neighbouring points within a buffer of user-defined radius (see details in Figure 3.4). The tool enables users to load a shapefile of the semantic trajectory, determine userID and location frequency fields that show the frequency of re-occurrence of a location in a user's trajectories, and define a suppression buffer radius and N number of personal frequent locations to be suppressed. After the suppression process is complete, the tool provides the assessment of information loss, by measuring the proportion of removed points to the total number of points in the input data and is a useful and straightforward measure of potentially undesirable suppression made due to a large value of N or buffer radius, which can be adjusted to result in an acceptable information loss.

Generalization tool

The second tool—the generalization tool—was developed to produce a spatially and temporally generalized dataset according to user-defined spatiotemporal resolution. This tool requires a shapefile containing GPS points with four fields indicating semantic tags, userID, date, and timestamp; time resolution (1 hour is the default value); spatial k-anonymity value (12 is the default value) and land-use dataset for stop generalization; and gap distance (300 m is the default value) and radius of buffer (60 m is the default value) for move generalization. This tool converts GPS points into a sequence of stay zones and travel zones while assigning average stay or travel duration within a specific time window [t1, t2], where t1 is the entering time and t2 is the leaving time (in general) and reports the population of each zone within a given time window.

To this end, the tool starts with temporal generalization by which timestamps of GPS points are replaced with an approximate time considering user-defined time resolution (e.g., 8:20 pm is generalized to 8:00 pm and 8:45 pm is generalized to 9:00 pm when time resolution is 1 hour).

After temporal generalization is complete, spatial generalization is achieved by generalizing stop points, while satisfying spatial k-anonymity (SKA) as a privacy criterion. Each stop point is replaced by a larger stay zone with k-locations in the form of a buffer centered at the original stop point. Then all other co-occurred points within the buffer are removed before assigning the enter/exit time (in general) and stay durations in addition to the total number of co-occurred users within a given time interval (one hour). The radius of buffers varies depending on the local location density. With regard to SKA, a higher value suggests greater privacy protection while reducing the spatial accuracy of the anonymized dataset by enlarging the radius of buffers

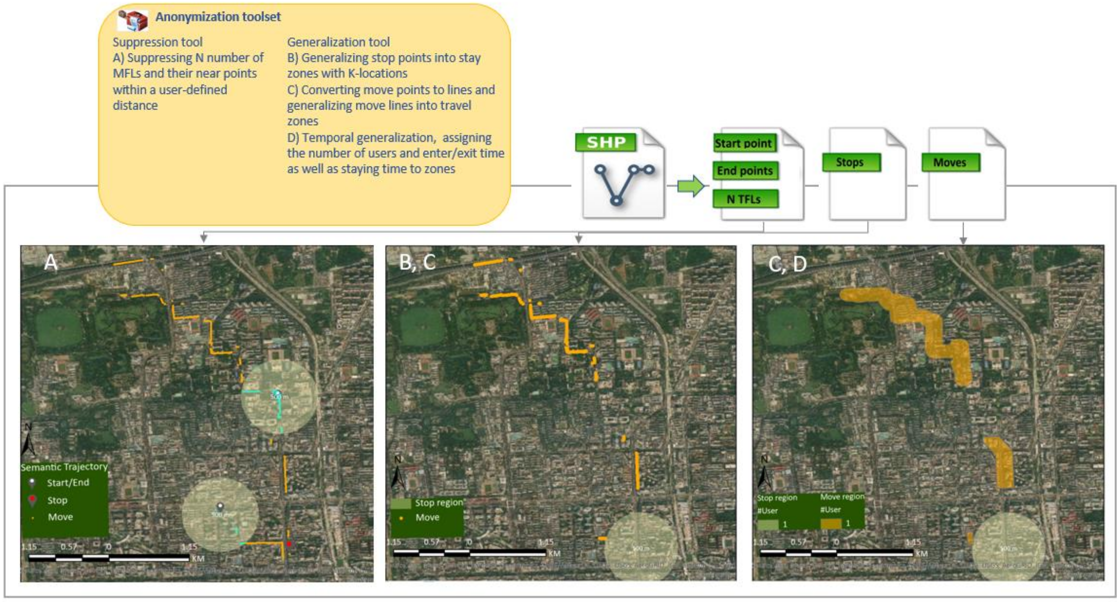

Anonymization toolset
Suppression tool
A) Suppressing N number of MFLs and their near points within a user-defined distance
Generalization tool
B) Generalizing stop points into stay zones with K-locations
C) Converting move points to lines and generalizing move lines into travel zones
D) Temporal generalization, assigning the number of users and enter/exit time as well as staying time to zones
SHP
Start point
End points
N TFLs
Stops
Moves
A
B, C
C, D
Semantic Trajectory
Start/End
Stop
Move
Stop region
Move
Stop region
#User
Move region
#User
1
1
1.15
0.57
0
1.15
KM

Figure 3.4. **The process of anonymizing semantic trajectories in two steps. 1) The suppression tool removes a given number of MFLs; and 2)the generalization tool replaces stops and move points by stay zones and travel zones, respectively, while assigning the number of users and enter/exit time (in general) as well as staying duration to them.**

To spatially generalize move points, the generalization tool combines points to line and buffer tools in ArcGIS. To this end, it first partitions each daily trip into move segments between every two consecutive stops and/or gaps. Then move points among them are converted into a polyline feature and subsequently replaced by a buffer centered at the polyline with a user-defined radius, while assigning enter/exit time (in general), the total number of travelers, average traveling time, and average speed of travelers within the travel zone. The whole process of converting trajectories to generalized zones is straightforward. Publishing travel zones traversed by a single user may reveal the user identity and satisfying k-anonymity can reduce the risk of user identity disclosure. However, it is hard to achieve for datasets with a low density of trajectories, especially within a short time window.

After the generalization process is complete, the tool computes and assigns the average value of k-anonymity, as the number of users in each stay or travel zones, as an indicator of the privacy level and allows a user to adjust spatiotemporal resolution until achieving acceptable privacy protection. During testing, the tool was able to generalize 100,000 points in just 3 minutes when setting the default values for input parameters.

Grid masking tool

Last but not least, an alternative anonymization technique was employed adopted from the framework proposed by Wang & McArthur (2018), to combine grid-masking and aggregation techniques and produce an aggregated dataset in a raster format with user-defined cell size (60 m is the default value). The benefit of producing aggregated dataset is that there is a low risk of user re-identification, even if an attacker has partial information about a user since average information about the population is released. This tool produces aggregated raster dataset in which grid cells indicate the sum of traveling/staying time by users within a specific time period without releasing personal information as shown in Figure 3.5. This dataset is useful for transport planning models which require aggregated travel information rather than user-specific trip information. In this case, the tool only requires GPS records of individuals in a shapefile format. The date field must be identified, and the time field must be selected as the cell value while the sum must be selected as the cell assessment type. Another output of this tool can be an aggregated dataset in which grid cells indicate the total number of users within a specific time period. This output can be achieved if the userID field is selected as the cell value instead of the time field. To better preserve privacy, enlarging the cell size can blur GPS points into larger areas, making it difficult to find the exact position of users, however, the tool restricts the maximum cell size to 200 m to maintain data utility.

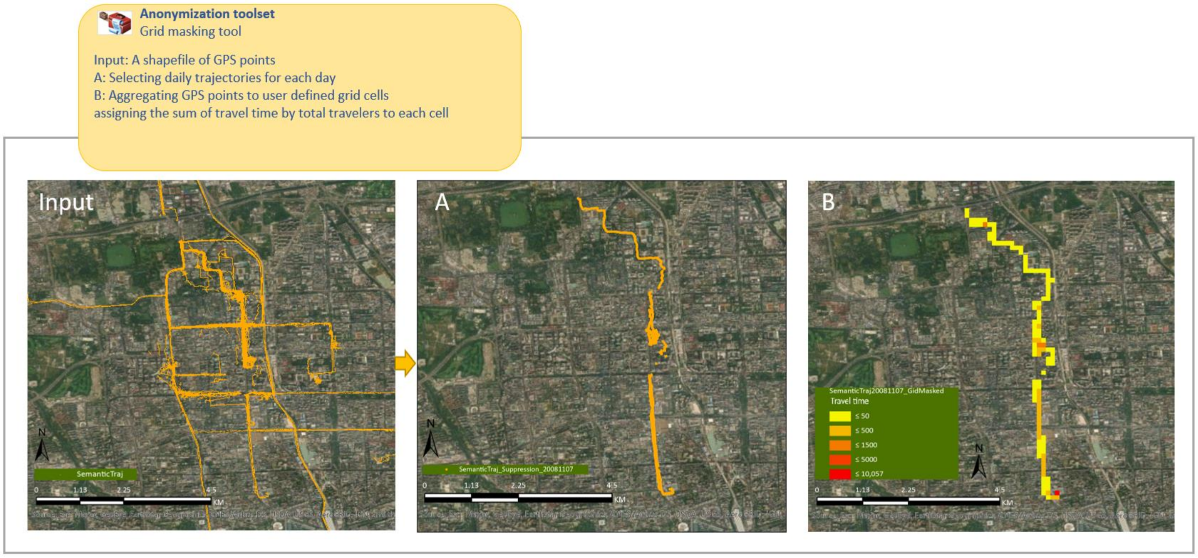

Anonymization toolset
Grid masking tool
Input: A shapefile of GPS points
A: Selecting daily trajectories for each day
B: Aggregating GPS points to user defined grid cells
assigning the sum of travel time by total travelers to each cell
Input
SemanticTraj
A
SemanticTraj_Suppression_20081107
B
SemanticTraj20081107_GidMasked
Travel time
≤ 50
≤ 500
≤ 1500
≤ 5000
≤ 10,057
N
0
1.13
2.25
4.5
KM

Figure 3.5. The process of grid masking and aggregating daily trips using the grid masking tool. The output of this tool is an aggregated raster dataset in which cells show the sum of travel time by travelers.

3.6. Discussion

Despite growing recognition of the need to protect individuals' privacy and an enhancement of privacy protection techniques, few readily available/usable means exist to allow data analysts to create a balance between privacy protection and data utility. We used Python to develop an ArcGIS toolbox, which enables an automated process of privacy protection. We created these freely available tools to overcome the logistical and coding hurdles that would be faced by non-specialists trying to anonymize their trajectory data.

The PrivacyProtection toolbox is an attempt to enhance the accessibility of movement data without privacy concerns. The immediate strength of the PrivacyProtection toolbox is that it provides data owners/publishers an opportunity to protect geoprivacy while enabling meaningful analysis of movement data in a way that facilitates adoption by users. Providing the GUI enables users to easily and flexibly achieve the desired stops within trajectories, identify locations related to users' habits, suppress a set of these locations, and vary the spatial and temporal resolutions to protect their privacy. For each process, parameter values can be adjusted by users and additional datasets can be used to enable comparison and analysis of different outputs and determine when an acceptable privacy level has been found. To assess privacy level, the application uses k-anonymity and measures the average k-anonymity (the number of users) of all anonymity regions in a specific time window. An alternative method to suppression and generalization is to perturb and cluster points. While perturbation can preserve privacy and data utility somewhat, it does not maintain the trip semantic and association to road segments.

A key feature of the PrivacyProtection toolbox is its user-friendly tooltips. Each tool includes a comprehensive descriptive summary to assist users in understanding how the tool works, as well as an explanatory picture and dialog and scripting explanations next to almost every input and output. The toolbox also provides a video tutorial that records the whole process of semantic trajectory construction and anonymization from start to finish. The PrivacyProtection toolbox is publicly available as a geoprocessing package on ArcGIS online.

It must be noted that trajectory anonymization is an onerous task that requires significant domain expertise. Easy-to-use toolsets are essential, but they are not a silver bullet as they do not replace expert decision-making when selecting anonymization parameters. Some may argue that such frictionless tools may lull ill-equipped users into a false sense of security, and it is for this reason that we have included a warning in the tool's interface about the importance of parameter selection when anonymizing spatial data. Nevertheless, while our tool cannot replace conceptual expertise, it significantly lowers the barrier to actually operationalizing anonymization methods and thus brings them within reach of a wider audience.

A few limitations apply to privacy protection with the PrivacyProtection toolbox. The anonymization tool simply blurs moves into larger travel zones without ensuring k-anonymity for moves in order to produce a dataset for road network-based analysis. Publishing k-anonymous moves require moves to be traversed by at least k co-occurred users which is not likely possible within a small time window, especially for datasets with a small number of users. This can result in significant information loss. Therefore, methods to satisfy k-anonymity for all parts of trajectories, and not only stops, will be interesting to explore in the future.

3.7. Conclusion

Trajectory data are particularly valuable, but also vulnerable because they reveal much about individual human lives while constituting a picture of people's habits and preferences. Privacy, therefore, is a pressing issue that needs to be strongly protected when dealing with trajectory data—however, it is often still ignored at data publishing time. PPDP techniques are a solution for protecting privacy while maintaining data utility for meaningful analysis. Unfortunately, these techniques have remained largely unrealized and are often out of reach in real-world applications, squirreled away in specialized academic literature. Our toolbox is a means to transform some of these techniques into an operationalized GIS workflow and make them accessible for a wide range of data owners who, to date, have had to withhold or suppress their data due to privacy concerns. The PrivacyProtection toolbox adopts a combination of spatiotemporal generalization and suppression techniques in addition to grid-masking and provides the assessment of information loss and level of individuals' privacy depending on user-defined parameter values. It does this by implementing the entire anonymization process in a form of a user-friendly ArcGIS toolbox. Along with being highly accessible, this toolbox importantly provides GPS data processing and analysis as pre-processing steps to convert raw data into meaningful data from a human point of view and extract locations related to users' habits considering their regularity.

Chapter 4. Conclusion

The ongoing uptake of trajectory data in academic research, business, and industry has caused geoprivacy to become a key issue for consideration, yet geoprivacy is often undervalued and neglected when publishing sensitive trajectory data in maps. Today, with the advancement of easy-to-use web maps and reverse geocoding tools, it is feasible to extract individuals’ home-work addresses and POIs from even aggregated trajectory data. Further, there is a high risk of uniquely re-identifying users and their sensitive information from trajectories, even with knowledge about as few as four random spatiotemporal points in those trajectories. This thesis has argued that because publishing trajectories can reveal personal sensitive information and raise serious privacy violations, it is imperative to develop an effective trajectory privacy protection framework as well as an easy-to-use toolbox for GIS users to use, even without deep knowledge of anonymization techniques. More specifically, it sought to answer these questions:

I. How do we represent trajectory data in a meaningful way from the human point of view?

II. How do we reduce the re-identification risk of users, locations related to their habits, and trajectory privacy violations when publishing users’ trajectories?

III. How do we make the process of semantic trajectory construction and privacy protection accessible in a form of an all-in-one toolbox for even non-expert GIS users?

4.1. Thesis summary

The main empirical findings are chapter-specific and were summarized within the respective chapters. Chapter 1 discusses the basic definitions and concepts regarding trajectory and geoprivacy. Chapter 2 propose a privacy protection framework for semantic trajectories that addresses ongoing geoprivacy issues by employing several anonymization techniques to protect both locations related to users’ habits and entire trajectories. It derives semantic components of trajectories which leads to providing more effective privacy protection. Significantly, since frequent and/or long-stayed locations of individuals are more related to their habits and identities, the framework identifies personal frequent locations and protects these key locations, rather than all sample points, to reduce information loss and preserve the integrity of trajectories. In order to protect personal frequent locations, it makes them indistinguishable from other k-1 locations with the same function. It also obfuscates the linkage between actual users and trajectories to protect entire trajectories with the purpose of providing unlinkability between locations of users. In the end, it compares different values of effective parameters in the anonymization process and their effects on the privacy level and attempts to show the acceptable balance between privacy and data utility.

Chapter 3 develops an easy-to-use toolbox in the ArcGIS Pro environment that makes the anonymization methods described in Chapter 2, accessible to everyone without needing a deep understanding of the methods themselves. It interprets the process of semantic trajectory construction and privacy protection into a GIS workflow. This toolbox is an attempt to address the problem of suppressing spatiotemporal data due to privacy concerns. It is comprised of six tools; each enables users to define parameter values and adjust the output to achieve acceptable results. Although the issue of privacy protection for semantic trajectory has been noted in previous literature, it has never been implemented as a plugin, toolbox, or standalone desktop application, accessible to all levels of users, that is so important to keep sensitive data local and safe.

4.2. Research Contributions

This thesis adds to the current body of trajectory privacy research by more deeply exploring the issues that were neglected by other privacy designers. Data publishing is always a limitation, but as more spatial data become available, geoprivacy research needs to be updated to address the possible complexities that might get hidden in black boxes. Two primary contributions can be distilled from this thesis. First, it offers a combined framework for privacy protection in trajectory data publishing scenarios to not only protect locations related to users' habits and entire trajectories but also to maintain the most semantics of visited locations and small modification to sample points. The anonymized dataset is beneficial for both mobility-based analysis and location-based analysis.

Second, and in light of the design of the proposed framework, it calls for more adoption and utilization of geoprivacy in practice, which thus far has been restricted to academic research and which it is imperative to make more accessible for users without advanced expertise and knowledge. The adoption of geoprivacy for trajectories, to date, has received little attention, preventing data owners from publishing their data due to privacy concerns. This thesis translates the process of trajectory privacy protection into a GIS framework, and designs and develops an all-in-one ArcGIS toolbox to bring the implementation of trajectory privacy into practice.

4.3. Future Work

Future work should continue to implement and make accessible available geoprivacy techniques, especially for trajectory data that contain more location samples of individuals and are, therefore, at higher risk of a privacy breach. More fundamentally, the privacy-preserving data mining algorithms that anonymize data for a specific data mining task deserve execution and more adoption. Finally, it is imperative to characterize potential attacks to trajectory datasets before data anonymization and use machine-learning algorithms to find out what information can be inferred from a dataset after anonymization.

References

Abul, O., Bonchi, F., & Nanni, M. (2010). Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, *35*(8), 884–910. https://doi.org/10.1016/j.is.2010.05.003

Abul, O., Bonchi, F., & Nanni, M. (2008). Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. *2008 IEEE 24th International Conference on Data Engineering*, 376–385. https://doi.org/10.1109/ICDE.2008.4497446

Acs, G., & Castelluccia, C. (2014). A case study: Privacy preserving release of spatio-temporal density in paris. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '14*, 1679–1688. https://doi.org/10.1145/2623330.2623361

Adrienko, N., & Adrienko, G. (2011). Spatial Generalization and Aggregation of Massive Movement Data. *IEEE Transactions on Visualization and Computer Graphics*, *17*(2), 205–219. https://doi.org/10.1109/TVCG.2010.44

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., & Zhu, A. (2005). Anonymizing Tables. In T. Eiter & L. Libkin (Eds.), *Database Theory—ICDT 2005* (pp. 246–258). Springer. https://doi.org/10.1007/978-3-540-30570-5_17

Alaggan, M., Gambs, S., Matwin, S., & Tuhin, M. (2015). Sanitization of Call Detail Records via Differentially-Private Bloom Filters. In P. Samarati (Ed.), *Data and Applications Security and Privacy XXIX* (pp. 223–230). Springer International Publishing. https://doi.org/10.1007/978-3-319-20810-7_15

Ali, S., Islam, N., Rauf, A., Din, I. U., Guizani, M., & Rodrigues, J. J. P. C. (2018). Privacy and Security Issues in Online Social Networks. *Future Internet*, *10*(12), 114. https://doi.org/10.3390/fi10120114

Alvares, L. O., Bogorny, V., Kuijpers, B., de Macedo, J. A. F., Moelans, B., & Vaisman, A. (2007). A model for enriching trajectories with semantic geographical information. *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems - GIS '07*, 1. https://doi.org/10.1145/1341012.1341041

Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K., & Palamidessi, C. (2013). Geo-Indistinguishability: Differential Privacy for Location-Based Systems. *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security - CCS '13*, 901–914. https://doi.org/10.1145/2508859.2516735

Andrienko, G., Andrienko, N., Giannotti, F., Monreale, A., & Pedreschi, D. (2009). Movement data anonymity through generalization. *Proceedings of the 2nd SIGSPATIAL ACM GIS 2009 International Workshop on Security and Privacy in GIS and LBS - SPRINGL '09*, 27. https://doi.org/10.1145/1667502.1667510

Arain, Q. A., Deng, Z., Memon, I., Zubedi, A., Jiao, J., Ashraf, A., & Khan, M. S. (2017). Privacy protection with dynamic pseudonym-based multiple mix-zones over road networks. *China Communications*, *14*(4), 89–100. https://doi.org/10.1109/CC.2017.7927579

Baglioni, M., Fernandes de Macêdo, J. A., Renso, C., Trasarti, R., & Wachowicz, M. (2009). Towards Semantic Interpretation of Movement Behavior. In M. Sester, L. Bernard, & V. Paelke (Eds.), *Advances in GIScience* (pp. 271–288). Springer Berlin Heidelberg.

Baik Hoh, & Gruteser, M. (2005). Protecting Location Privacy Through Path Confusion. *First International Conference on Security and Privacy for Emerging Areas in Communications Networks (SECURECOMM'05)*, 194–205. https://doi.org/10.1109/SECURECOMM.2005.33

Baik Hoh, Gruteser, M., Hui Xiong, & Alrabady, A. (2006). Enhancing Security and Privacy in Traffic-Monitoring Systems. *IEEE Pervasive Computing*, *5*(4), 38–46. https://doi.org/10.1109/MPRV.2006.69

Basu, A., Monreale, A., Corena, J. C., Giannotti, F., Pedreschi, D., Kiyomoto, S., Miyake, Y., Yanagihara, T., & Trasarti, R. (2014). A Privacy Risk Model for Trajectory Data. In J. Zhou, N. Gal-Oz, J. Zhang, & E. Gudes (Eds.), *Trust Management VIII* (Vol. 430, pp. 125–140). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-43813-8_9

Beresford, A. R., & Stajano, F. (2003). Location privacy in pervasive computing. *IEEE Pervasive Computing*, *2*(1), 46–55. https://doi.org/10.1109/MPRV.2003.1186725

Bettini, C., Wang, X. S., & Jajodia, S. (2005). Protecting Privacy Against Location-Based Personal Identification. In W. Jonker & M. Petković (Eds.), *Secure Data Management* (pp. 185–199). Springer Berlin Heidelberg.

Bhatia, S., Carrion, M., Cohn, E., Cori, A., Nouvellet, P., Lassmann, B., Madoff, L., & Brownstein, J. (2019). Big brother is watching—Using digital disease surveillance tools for near real-time forecasting. *International Journal of Infectious Diseases*, *79*, 27. https://doi.org/10.1016/j.ijid.2018.11.080

Boeing, G. (2020). The right tools for the job: The case for spatial science tool-building. *Transactions in GIS*, *24*(5), 1299–1314. https://doi.org/10.1111/tgis.12678

Buttyán, L., Holczer, T., & Vajda, I. (2007). On the Effectiveness of Changing Pseudonyms to Provide Location Privacy in VANETs. In F. Stajano, C. Meadows, S. Capkun, & T. Moore (Eds.), *Security and Privacy in Ad-hoc and Sensor Networks* (pp. 129–141). Springer. https://doi.org/10.1007/978-3-540-73275-4_10

Cáceres, M. D., Coll, L., Legendre, P., Allen, R. B., Wiser, S. K., Fortin, M.-J., Condit, R., & Hubbell, S. (2019). Trajectory analysis in community ecology. *Ecological Monographs*, *89*(2), e01350. https://doi.org/10.1002/ecm.1350

Carneiro, H. A., & Mylonakis, E. (2009). Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. *Clinical Infectious Diseases*, *49*(10), 1557–1564. https://doi.org/10.1086/630200

Cecaj, A., Mamei, M., & Zambonelli, F. (2016). Re-identification and information fusion between anonymized CDR and social network data. *Journal of Ambient Intelligence and Humanized Computing*, *7*(1), 83–96. https://doi.org/10.1007/s12652-015-0303-x

Chang, S., Li, C., Zhu, H., Lu, T., & Li, Q. (2018). Revealing Privacy Vulnerabilities of Anonymous Trajectories. *IEEE Transactions on Vehicular Technology*, *67*(12), 12061–12071. https://doi.org/10.1109/TVT.2018.2871745

Charleux, L. (2015). A GIS Toolbox for Measuring and Mapping Person-Based Space-Time Accessibility. *Transactions in GIS*, *19*(2), 262–278. https://doi.org/10.1111/tgis.12115

Chen, R., Acs, G., & Castelluccia, C. (2012). Differentially private sequential data publication via variable-length n-grams. *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, 638–649. https://doi.org/10.1145/2382196.2382263

Chen, R., Fung, B. C. M., Mohammed, N., Desai, B. C., & Wang, K. (2013). Privacy-preserving trajectory data publishing by local suppression. *Information Sciences*, *231*, 83–97. https://doi.org/10.1016/j.ins.2011.07.035

Cicek, A. E., Nergiz, M. E., & Saygin, Y. (2014). Ensuring location diversity in privacy-preserving spatio-temporal data publishing. *The VLDB Journal — The International Journal on Very Large Data Bases*, *23*(4), 609–625. https://doi.org/10.1007/s00778-013-0342-x

Clifton, C., & Tassa, T. (2013). On syntactic anonymity and differential privacy. *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, 88–93. https://doi.org/10.1109/ICDEW.2013.6547433

Conolly, J., Ontario), J. (Trent U. C., Peterborough, & Lake, M. (2006). *Geographical Information Systems in Archaeology*. Cambridge University Press.

Dai, H., Li, H., Meng, X., & Wang, Y. (2020). On the Vulnerability and Generality of K–Anonymity Location Privacy Under Continuous LBS Requests. In X. Wang, R. Zhang, Y.-K. Lee, L. Sun, & Y.-S. Moon (Eds.), *Web and Big Data* (pp. 351–359). Springer International Publishing. https://doi.org/10.1007/978-3-030-60290-1_28

Dai, Y., Shao, J., Wei, C., Zhang, D., & Shen, H. T. (2018). Personalized semantic trajectory privacy preservation through trajectory reconstruction. *World Wide Web*, *21*(4), 875–914. https://doi.org/10.1007/s11280-017-0489-2

Dalumpines, R., & Scott, D. M. (2011). GIS-based Map-matching: Development and Demonstration of a Postprocessing Map-matching Algorithm for Transportation Research. In S. Geertman, W. Reinhardt, & F. Toppen (Eds.), *Advancing Geoinformation Science for a Changing World* (pp. 101–120). Springer. https://doi.org/10.1007/978-3-642-19789-5_6

Dankar, F. K., & Emam, K. E. (2013). *Practicing Differential Privacy in Health Care: A Review*. 33.

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, *3*, 1376. https://doi.org/10.1038/srep01376

Deldar, F., & Abadi, M. (2018). PLDP-TD: Personalized-location differentially private data analysis on trajectory databases. *Pervasive and Mobile Computing*, *49*, 1–22. https://doi.org/10.1016/j.pmcj.2018.06.005

Deldar, F., & Abadi, M. (2019). PDP-SAG: Personalized Privacy Protection in Moving Objects Databases by Combining Differential Privacy and Sensitive Attribute Generalization. *IEEE Access*, *7*, 85887–85902. https://doi.org/10.1109/ACCESS.2019.2925236

Demirag, D., & Ayday, E. (2020). Tracking the Invisible: Privacy-Preserving Contact Tracing to Control the Spread of a Virus. *ArXiv:2003.13073 [Cs]*. http://arxiv.org/abs/2003.13073

Duckham, M., & Kulik, L. (2006). Location privacy and location-aware computing. *Dynamic and Mobile GIS; CRC Press*, 20. https://doi.org/9781420008609-11

Duckham, M., & Kulik, L. (2005). A Formal Model of Obfuscation and Negotiation for Location Privacy. In H.-W. Gellersen, R. Want, & A. Schmidt (Eds.), *Pervasive Computing* (pp. 152–170). Springer Berlin Heidelberg.

Dwork, C. (2006). Differential Privacy. In M. Bugliesi, B. Preneel, V. Sassone, & I. Wegener (Eds.), *Automata, Languages and Programming* (pp. 1–12). Springer. https://doi.org/10.1007/11787006_1

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In S. Halevi & T. Rabin (Eds.), *Theory of Cryptography* (pp. 265–284). Springer Berlin Heidelberg.

Elwood, S., & Leszczynski, A. (2011). Privacy, reconsidered: New representations, data practices, and the geoweb. *Geoforum*, *42*(1), 6–15. https://doi.org/10.1016/j.geoforum.2010.08.003

Eom, C. S.-H., Lee, C. C., Lee, W., & Leung, C. K. (2020). Effective privacy preserving data publishing by vectorization. *Information Sciences*, *527*, 311–328. https://doi.org/10.1016/j.ins.2019.09.035

Fiore, M., Katsikouli, P., Zavou, E., Cunche, M., Fessant, F., Hello, D. L., Aivodji, U., Olivier, B., Quertier, T., & Stanica, R. (2020). Privacy in trajectory micro-data publishing: A survey. *Transactions on Data Privacy*, *13*, 91–149.

Freudiger, J., Raya, M., Félegyházi, M., Papadimitratos, P., & Hubaux, J.-P. (2007). *Mix-Zones for Location Privacy in Vehicular Networks*. ACM Workshop on Wireless Networking for Intelligent Transportation Systems (WiN-ITS). https://infoscience.epfl.ch/record/109437

Freudiger, J., Shokri, R., & Hubaux, J.-P. (2012). Evaluating the Privacy Risk of Location-Based Services. In G. Danezis (Ed.), *Financial Cryptography and Data Security* (pp. 31–46). Springer Berlin Heidelberg.

Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, *42*(4), 14:1-14:53. https://doi.org/10.1145/1749603.1749605

Gambs, S., Killijian, M.-O., & Cortez, M. N. (2010). GEPETO: A GEoPrivacy-Enhancing TOolkit. *2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops*, 1071–1076. https://doi.org/10.1109/WAINA.2010.170

Gambs, S., Killijian, M.-O., & Núñez del Prado Cortez, M. (2014). De-anonymization attack on geolocated data. *Journal of Computer and System Sciences*, *80*(8), 1597–1614. https://doi.org/10.1016/j.jcss.2014.04.024

Ghinita, G. (2009). Private Queries and Trajectory Anonymization: A Dual Perspective on Location Privacy. *Cyber Center Publications*. https://docs.lib.purdue.edu/ccpubs/129

Giannotti, F., Nanni, M., Pedreschi, D., & Renso, C. (2009). *GeoPKDD Geographic Privacy-aware Knowledge Discovery*. 8.

Giannotti, F., Nanni, M., Pedreschi, D., Renso, C., & Trasarti, R. (2009). Mining Mobility Behavior from Trajectory Data. *2009 International Conference on Computational Science and Engineering*, *4*, 948–951. https://doi.org/10.1109/CSE.2009.542

Golle, P., & Partridge, K. (2009). On the Anonymity of Home/Work Location Pairs. In H. Tokuda, M. Beigl, A. Friday, A. J. B. Brush, & Y. Tobe (Eds.), *Pervasive Computing* (pp. 390–397). Springer. https://doi.org/10.1007/978-3-642-01516-8_26

Gramaglia, M., & Fiore, M. (2015). Hiding Mobile Traffic Fingerprints with GLOVE. *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, 26:1-26:13. https://doi.org/10.1145/2716281.2836111

Gramaglia, M., Fiore, M., Tarable, A., & Banchs, A. (2017). Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories. *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, 1–9. https://doi.org/10.1109/INFOCOM.2017.8056979

Gruteser, M., & Grunwald, D. (2003). *Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking*. 13.

Guo, N., Ma, L., & Gao, T. (2018). Independent Mix Zone for Location Privacy in Vehicular Networks. *IEEE Access*, *6*, 16842–16850. https://doi.org/10.1109/ACCESS.2018.2800907

Hampton, K. H., Fitch, M. K., Allshouse, W. B., Doherty, I. A., Gesink, D. C., Leone, P. A., Serre, M. L., & Miller, W. C. (2010). Mapping Health Data: Improved Privacy Protection With Donut Method Geomasking. *American Journal of Epidemiology*, *172*(9), 1062–1069. https://doi.org/10.1093/aje/kwq248

Han, P., & Tsai, H. (2015). SST: Privacy Preserving for Semantic Trajectories. *2015 16th IEEE International Conference on Mobile Data Management*, *2*, 80–85. https://doi.org/10.1109/MDM.2015.18

Hasan, A. S. M. T., Jiang, Q., & Li, C. (2017). An Effective Grouping Method for Privacy-Preserving Bike Sharing Data Publishing. *Future Internet*, *9*(4), 65. https://doi.org/10.3390/fi9040065

Hay, M., Machanavajjhala, A., Miklau, G., Chen, Y., & Zhang, D. (2016). Principled Evaluation of Differentially Private Algorithms using DPBench. *Proceedings of the 2016 International Conference on Management of Data*, 139–154. https://doi.org/10.1145/2882903.2882931

Hern, A. (2018, January 28). Fitness tracking app Strava gives away location of secret US army bases. *The Guardian*. https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases

Huo, Z., Meng, X., Hu, H., & Huang, Y. (2012). You Can Walk Alone: Trajectory Privacy-Preserving through Significant Stays Protection. In S. Lee, Z. Peng, X. Zhou, Y.-S. Moon, R. Unland, & J. Yoo (Eds.), *Database Systems for Advanced Applications* (pp. 351–366). Springer Berlin Heidelberg.

Hwang, S., Hanke, T., & Evans, C. (2013). Automated Extraction of Community Mobility Measures from GPS Stream Data Using Temporal DBSCAN. In B. Murgante, S. Misra, M. Carlini, C. M. Torre, H.-Q. Nguyen, D. Taniar, B. O. Apduhan, & O. Gervasi (Eds.), *Computational Science and Its Applications – ICCSA 2013* (pp. 86–98). Springer Berlin Heidelberg.

Ishikawa, M., & Yoshiura, N. (2020). Privacy Protection in Location Based Service by Secure Computation. In N. T. Nguyen, K. Jearanaitanakij, A. Selamat, B. Trawiński, & S. Chittayasothorn (Eds.), *Intelligent Information and Database Systems* (pp. 493–504). Springer International Publishing. https://doi.org/10.1007/978-3-030-42058-1_41

Jiang, K., Shao, D., Bressan, S., Kister, T., & Tan, K.-L. (2013). Publishing trajectories with differential privacy guarantees. *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, 1–12. https://doi.org/10.1145/2484838.2484846

Kalnis, P., & Ghinita, G. (2009). Spatial Anonymity. *Cyber Center Publications*. https://docs.lib.purdue.edu/ccpubs/152

Kapa, S., Halamka, J., & Raskar, R. (2020). Contact Tracing to Manage COVID-19 Spread—Balancing Personal Privacy and Public Health. *Mayo Clinic Proceedings*, *95*(7), 1320–1322. https://doi.org/10.1016/j.mayocp.2020.04.031

Kellaris, G., Papadopoulos, S., Xiao, X., & Papadias, D. (2014). *Differentially private event sequences over infinite streams*. https://doi.org/10.14778/2732977.2732989

Kenig, B., & Tassa, T. (2012). A practical approximation algorithm for optimal k-anonymity. *Data Mining and Knowledge Discovery*, *25*(1), 134–168. https://doi.org/10.1007/s10618-011-0235-9

Keßler, C., & McKenzie, G. (2018). A geoprivacy manifesto. *Transactions in GIS*, *22*(1), 3–19. https://doi.org/10.1111/tgis.12305

Kiukkonen, N., J, B., Dousse, O., Gatica-Perez, D., & Laurila, J. (2010). *Towards rich mobile phone datasets: Lausanne data collection campaign*.

Kondor, D., Hashemian, B., de Montjoye, Y.-A., & Ratti, C. (2018). Towards matching user mobility traces in large-scale datasets. *IEEE Transactions on Big Data*, 1–1. https://doi.org/10.1109/TBDATA.2018.2871693

Kounadi, O., & Leitner, M. (2014). Why Does Geoprivacy Matter? The Scientific Publication of Confidential Data Presented on Maps. *Journal of Empirical Research on Human Research Ethics*, *9*(4), 34–45. https://doi.org/10.1177/1556264614544103

Krumm, J. (2009). A survey of computational location privacy. *Personal and Ubiquitous Computing*, *13*(6), 391–399. https://doi.org/10.1007/s00779-008-0212-5

Krumm, J. (2007). Inference Attacks on Location Tracks. In A. LaMarca, M. Langheinrich, & K. N. Truong (Eds.), *Pervasive Computing* (pp. 127–143). Springer. https://doi.org/10.1007/978-3-540-72037-9_8

Kwan, M.-P., Casas, I., & Schmitz, B. (2004). Protection of Geoprivacy and Accuracy of Spatial Information: How Effective Are Geographical Masks? *Cartographica: The International Journal for Geographic Information and Geovisualization*, *39*(2), 15–28. https://doi.org/10.3138/X204-4223-57MK-8273

Leszczynski, A. (2017). *Geoprivacy*. SAGE Publications. https://researchspace.auckland.ac.nz/handle/2292/33851

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. *2007 IEEE 23rd International Conference on Data Engineering*, 106–115. https://doi.org/10.1109/ICDE.2007.367856

Li, T., Jackie, Yang, Faklaris, C., King, J., Agarwal, Y., Dabbish, L., & Hong, J. I. (2020). Decentralized is not risk-free: Understanding public perceptions of privacy-utility trade-offs in COVID-19 contact-tracing apps. *ArXiv:2005.11957 [Cs]*. http://arxiv.org/abs/2005.11957

Lin, D., Gurung, S., Jiang, W., & Hurson, A. (2010). Privacy-Preserving Location Publishing under Road-Network Constraints. In H. Kitagawa, Y. Ishikawa, Q. Li, & C. Watanabe (Eds.), *Database Systems for Advanced Applications* (pp. 17–31). Springer. https://doi.org/10.1007/978-3-642-12098-5_2

Ma, C. Y. T., Yau, D. K. Y., Yip, N. K., & Rao, N. S. V. (2013). Privacy Vulnerability of Published Anonymous Mobility Traces. *IEEE/ACM Trans. Netw.*, *21*(3), 720–733. https://doi.org/10.1109/TNET.2012.2208983

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). L-diversity: Privacy beyond k-anonymity. *22nd International Conference on Data Engineering (ICDE'06)*, 24–24. https://doi.org/10.1109/ICDE.2006.1

Meyerson, A., & Williams, R. (2004). On the complexity of optimal K-anonymity. *Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 223–228. https://doi.org/10.1145/1055558.1055591

Mohammed, N., Fung, B. C. M., & Debbabi, M. (2010). *Preserving Privacy and Utility in RFID Data Publishing* [Monograph]. N/A. https://spectrum.library.concordia.ca/6850/

Monreale, A., Trasarti, R., Pedreschi, D., & Renso, C. (2011). *C-safety: A framework for the anonymiza- tion of semantic trajectories*. 29.

Monreale, A., Trasarti, R., Renso, C., Pedreschi, D., & Bogorny, V. (2010). Preserving Privacy in Semantic-rich Trajectories of Human Mobility. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*, 47–54. https://doi.org/10.1145/1868470.1868481

Narayanan, A., & Shmatikov, V. (2006). How To Break Anonymity of the Netflix Prize Dataset. *ArXiv:Cs/0610105*. http://arxiv.org/abs/cs/0610105

Nergiz, M. E., Atzori, M., & Saygin, Y. (2008). Towards trajectory anonymization: A generalization-based approach. *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS*, 52–61. https://doi.org/10.1145/1503402.1503413

Palanisamy, B., & Liu, L. (2011). MobiMix: Protecting location privacy with mix-zones over road networks. *2011 IEEE 27th International Conference on Data Engineering*, 494–505. https://doi.org/10.1109/ICDE.2011.5767898

Palma, A. T., Bogorny, V., Kuijpers, B., & Alvares, L. O. (2008). A clustering-based approach for discovering interesting places in trajectories. *Proceedings of the 2008 ACM Symposium on Applied Computing - SAC '08*, 863. https://doi.org/10.1145/1363686.1363886

Pfitzmann, A., & Köhntopp, M. (2001). Anonymity, Unobservability, and Pseudonymity—A Proposal for Terminology. In H. Federrath (Ed.), *Designing Privacy Enhancing Technologies: International Workshop on Design Issues in Anonymity and Unobservability Berkeley, CA, USA, July 25–26, 2000 Proceedings* (pp. 1–9). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44702-4_1

Pinto, A. M. (2012). A Comparison of anonymization protection principles. *2012 IEEE 13th International Conference on Information Reuse Integration (IRI)*, 207–214. https://doi.org/10.1109/IRI.2012.6303012

Poulis, G., Loukides, G., Gkoulalas-Divanis, A., & Skiadopoulos, S. (2013). Anonymizing Data with Relational and Transaction Attributes. In H. Blockeel, K. Kersting, S. Nijssen, & F. Železný (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 353–369). Springer Berlin Heidelberg.

Poulis, G., Skiadopoulos, S., Loukides, G., & Gkoulalas-Divanis, A. (2013). Distance-Based k^m-Anonymization of Trajectory Data. *2013 IEEE 14th International Conference on Mobile Data Management*, *2*, 57–62. https://doi.org/10.1109/MDM.2013.66

Radley-Gardner, O., Beale, H., & Zimmermann, R. (Eds.). (2016). *Fundamental Texts On European Private Law*. Hart Publishing. https://doi.org/10.5040/9781782258674

Richter, W. (2018). The verified neighbor approach to geoprivacy: An improved method for geographic masking. *Journal of Exposure Science and Environmental Epidemiology*, *28*(2), 109–118. https://doi.org/10.1038/jes.2017.17

Rossi, L., & Musolesi, M. (2014). It's the way you check-in: Identifying users in location-based social networks. *Proceedings of the Second ACM Conference on Online Social Networks*, 215–226. https://doi.org/10.1145/2660460.2660485

Rossi, L., Walker, J., & Musolesi, M. (2015). Spatio-temporal techniques for user identification by means of GPS mobility data. *EPJ Data Science*, *4*(1), 11. https://doi.org/10.1140/epjds/s13688-015-0049-x

Rudenko, A., Palmieri, L., Herman, M., Kitani, K. M., Gavrila, D. M., & Arras, K. O. (2019). Human Motion Trajectory Prediction: A Survey. *ArXiv:1905.06113 [Cs]*. http://arxiv.org/abs/1905.06113

Salas, J., Megías, D., & Torra, V. (2018). SwapMob: Swapping Trajectories for Mobility Anonymization. In J. Domingo-Ferrer & F. Montes (Eds.), *Privacy in Statistical Databases* (pp. 331–346). Springer International Publishing. https://doi.org/10.1007/978-3-319-99771-1_22

Samarati, P., & Sweeney, L. (1998). *Protecting Privacy when Disclosing Information: K-Anonymity and Its Enforcement through Generalization and Suppression*. 19.

Sankar, L., Rajagopalan, S. R., & Poor, H. V. (2010). A theory of utility and privacy of data sources. *2010 IEEE International Symposium on Information Theory*, 2642–2646. https://doi.org/10.1109/ISIT.2010.5513684

Seidl, D. E., Jankowski, P., & Clarke, K. C. (2018). Privacy and False Identification Risk in Geomasking Techniques. *Geographical Analysis*, *50*(3), 280–297. https://doi.org/10.1111/gean.12144

Shin, K. G., Ju, X., Chen, Z., & Hu, X. (2012). Privacy protection for users of location-based services. *IEEE Wireless Communications*, *19*(1), 30–39. https://doi.org/10.1109/MWC.2012.6155874

Sila-Nowicka, K., & Thakuriah, P. (2016, June 8). *The trade-off between privacy and geographic data resolution. A case of GPS trajectories combined with the social survey results* [Conference Proceedings]. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. https://doi.org/10.5194/isprs-archives-XLI-B2-535-2016

Skinner, C. J., & Elliot, M. J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 855–867. https://doi.org/10.1111/1467-9868.00365

Song, Y., Dahlmeier, D., & Bressan, S. (2014). *Not So Unique in the Crowd: A Simple and Effective Algorithm for Anonymizing Location Data*. 6.

Srivatsa, M., & Hicks, M. (2012). Deanonymizing mobility traces: Using social network as a side-channel. *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, 628–637. https://doi.org/10.1145/2382196.2382262

Sui, K., Zhao, Y., Liu, D., Ma, M., Xu, L., Zimu, L., & Pei, D. (2016). Your trajectory privacy can be breached even if you walk in groups. *2016 IEEE/ACM 24th International Symposium on Quality of Service (IWQoS)*, 1–6. https://doi.org/10.1109/IWQoS.2016.7590444

Swanlund, D., Schuurman, N., & Brussoni, M. (2020). MaskMy.XYZ: An easy-to-use tool for protecting geoprivacy using geographic masks. *Transactions in GIS*, *24*(2), 390–401. https://doi.org/10.1111/tgis.12606

Swanlund, D., Schuurman, N., Zandbergen, P., & Brussoni, M. (2020). Street masking: A network-based geographic mask for easily protecting geoprivacy. *International Journal of Health Geographics*, *19*(1), 26. https://doi.org/10.1186/s12942-020-00219-z

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570.

Terrovitis, M., & Mamoulis, N. (2008). Privacy Preservation in the Publication of Trajectories. *The Ninth International Conference on Mobile Data Management (Mdm 2008)*, 65–72. https://doi.org/10.1109/MDM.2008.29

The World Bank. (2016). https://data.worldbank.org/indicator/IT.CEL.SETS.P2

Tu, Z., Zhao, K., Xu, F., Li, Y., Su, L., & Jin, D. (2019a). Protecting Trajectory From Semantic Attack Considering k -Anonymity, l -Diversity, and t -Closeness. *IEEE Transactions on Network and Service Management*, *16*(1), 264–278. https://doi.org/10.1109/TNSM.2018.2877790

Tu, Z., Zhao, K., Xu, F., Li, Y., Su, L., & Jin, D. (2019b). Protecting Trajectory From Semantic Attack Considering k-Anonymity, l -Diversity, and t-Closeness. *IEEE Transactions on Network and Service Management*, *16*(1), 264–278. https://doi.org/10.1109/TNSM.2018.2877790

Tu, Z., Zhao, K., Xu, F., Li, Y., Su, L., & Jin, D. (2017). Beyond K-Anonymity: Protect Your Trajectory from Semantic Attack. *2017 14th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 1–9. https://doi.org/10.1109/SAHCN.2017.7964921

Unnikrishnan, J., & Naini, F. M. (2013). De-anonymizing private data by matching statistics. *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1616–1623. https://doi.org/10.1109/Allerton.2013.6736722

Wang, J., & Kwan, M.-P. (2020). Daily activity locations k-anonymity for the evaluation of disclosure risk of individual GPS datasets. *International Journal of Health Geographics*, *19*(1), 7. https://doi.org/10.1186/s12942-020-00201-9

Wang, Y., & McArthur, D. (2018). Enhancing data privacy with semantic trajectories: A raster-based framework for GPS stop/move management. *Transactions in GIS*, *22*(4), 975–990. https://doi.org/10.1111/tgis.12334

Wong, R. C.-W., Li, J., Fu, A. W.-C., & Wang, K. (2006). (α, K)-anonymity: An Enhanced K-anonymity Model for Privacy Preserving Data Publishing. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 754–759. https://doi.org/10.1145/1150402.1150499

Wong, R., Li, J., Fu, A., & Wang, K. (2009). (α, k)-anonymous data publishing. *Journal of Intelligent Information Systems*, *33*(2), 209–234. https://doi.org/10.1007/s10844-008-0075-2

Xu, C., Luo, L., Ding, Y., Zhao, G., & Yu, S. (2020). Personalized Location Privacy Protection for Location-Based Services in Vehicular Networks. *IEEE Wireless Communications Letters*, *9*(10), 1633–1637. https://doi.org/10.1109/LWC.2020.2999524

Xu, F., Tu, Z., Li, Y., Zhang, P., Fu, X., & Jin, D. (2017). Trajectory Recovery From Ash: User Privacy Is NOT Preserved in Aggregated Mobility Data. *Proceedings of the 26th International Conference on World Wide Web*, 1241–1250. https://doi.org/10.1145/3038912.3052620

Xue, M., Kalnis, P., & Pung, H. K. (2009). Location Diversity: Enhanced Privacy Protection in Location Based Services. In T. Choudhury, A. Quigley, T. Strang, & K. Suginuma (Eds.), *Location and Context Awareness* (pp. 70–87). Springer Berlin Heidelberg.

Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S., & Aberer, K. (2013). *Semantic trajectories: Mobility data computation and annotation*. Association for Computing Machinery. https://doi.org/10.1145/2483669.2483682

Yarovoy, R., Bonchi, F., Lakshmanan, L. V. S., & Wang, W. H. (2009). Anonymizing moving objects: How to hide a MOB in a crowd? *Proceedings of the 12th International Conference on Extending Database Technology Advances in Database Technology - EDBT ’09*, 72. https://doi.org/10.1145/1516360.1516370

Yin, L., Wang, Q., Shaw, S.-L., Fang, Z., Hu, J., Tao, Y., & Wang, W. (2015). Re-Identification Risk versus Data Utility for Aggregated Mobility Research Using Mobile Phone Location Data. *PLoS ONE*, *10*(10). https://doi.org/10.1371/journal.pone.0140589

Yuan, J., Zheng, Y., Xie, X., & Sun, G. (2011). Driving with knowledge from the physical world. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 316–324. https://doi.org/10.1145/2020408.2020462

Zang, H., & Bolot, J. (2011). Anonymization of location data does not work: A large-scale measurement study. *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, 145–156. https://doi.org/10.1145/2030613.2030630

Zhang, S., Freundschuh, S. M., Lenzer, K., & Zandbergen, P. A. (2017). The location swapping method for geomasking. *Cartography and Geographic Information Science*, *44*(1), 22–34. https://doi.org/10.1080/15230406.2015.1095655

Zheng, Y., Li, Q., Chen, Y., Xie, X., & Ma, W.-Y. (2008). Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on Ubiquitous computing* (pp. 312–321). Association for Computing Machinery. https://doi.org/10.1145/1409635.1409677

Zheng, Y., Xie, X., & Ma, W. (2010). *GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory*.

Zheng, Y., Zhang, L., Xie, X., & Ma, W.-Y. (2009). Mining Interesting Locations and Travel Sequences from GPS Trajectories. *Proceedings of the 18th International Conference on World Wide Web*, 791–800. https://doi.org/10.1145/1526709.1526816

Zheng, Y., & Zhou, X. (Eds.). (n.d.). *Computing with spatial trajectories*. Springer, c2011.