STRIKING THE BALANCE: PRIVACY AND SPATIAL PATTERN PRESERVATION IN MASKED GPS DATA

A Thesis

Presented to the

Faculty of

San Diego State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Geographic Information Science

---

by

Dara E. Seidl

Spring 2014

SAN DIEGO STATE UNIVERSITY

The Undersigned Faculty Committee Approves the

Thesis of Dara E. Seidl:

Striking the Balance: Privacy and Spatial Pattern Preservation in Masked GPS Data

____
Piotr Jankowski, Chair
Department of Geography

____
Ming-Hsiang Tsou
Department of Geography

____
Carl Eckberg
Department of Computer Science

____
Approval Date

Copyright © 2014

by

Dara E. Seidl

All Rights Reserved

DEDICATION

This work is dedicated to the family, friends, and colleagues who supported me in my journey to pursue graduate education in GIScience on the West Coast. I have been very fortunate to have such an encouraging network of loved ones in my life. This is also dedicated to the new friends and colleagues who have taken a New Yorker under their wings and made San Diego feel like home.

The intensity and complexity of life, attendant upon advancing civilization, have rendered necessary some retreat from the world…

--Warren and Brandeis
"The Right to Privacy," *Harvard Law Review*

ABSTRACT OF THE THESIS

Striking the Balance: Privacy and Spatial Pattern Preservation in Masked GPS Data
by
Dara E. Seidl
Master of Science in Geographic Information Science
San Diego State University, 2014

Volunteered location and trajectory data are increasingly collected and applied in analysis for a variety of academic fields and recreational pursuits. As access to personal location data increases, issues of privacy arise as individuals become identifiable and linked to other repositories of information. While the quality and precision of data are essential to accurate analysis, there is a tradeoff between privacy and access to data. Obfuscation of point data is a solution that aims to protect privacy and maximize preservation of spatial pattern. This study explores two methods of location obfuscation for volunteered GPS data: grid masking and random perturbation. These methods are applied to travel survey GPS data in the greater metropolitan regions of Chicago and Atlanta in the first large-scale GPS masking study of its kind.

TABLE OF CONTENTS

PAGE

ABSTRACT vi
LIST OF TABLES ix
LIST OF FIGURES x
ACKNOWLEDGEMENTS xi
CHAPTER
1 INTRODUCTION 1
Justification 2
2 BACKGROUND 4
Conceptualization of Privacy 4
Evolving Privacy Concerns in GIS 5
Motivations for Volunteering Geographic Information 6
Tradeoffs between Privacy and Data Access 7
Approaches to Privacy Preservation 8
Obfuscation Techniques 9
Gaps in Obfuscation Research 10
3 CONCEPTUALIZATION 11
Hypotheses 11
Study Area 13
Chicago 13
Atlanta 14
GPS Data Access 15
4 METHODS 17
Obfuscation 17
Grid Masking 17
Random Perturbation 18
k-Anonymity 21
Point Anonymity 21
Route Anonymity 22
Spatial Pattern Preservation 25

5 RESULTS AND DISCUSSION 27
k-Anonymity 28
Point Anonymity 28
Route Anonymity 31
Spatial Pattern Preservation 35
6 CONCLUSIONS 38
REFERENCES 40
A IRB APPROVAL DOCUMENT 46

LIST OF TABLES

PAGE

Table 1. Summary of GPS data points in the study regions 28

Table 2. Mean occupied housing density for masked home locations, CMAP 29

Table 3. Percent of home locations relocated to different block group, CMAP......................29

Table 4. Mean occupied housing density for masked home locations, ARC 30

Table 5. Percent of home locations relocated to different block group, ARC.........................30

Table 6. Mean k-anonymity index by trip, CMAP 32

Table 7. Mean k-anonymity index by trip, ARC 33

Table 8. Results of a linear regression predicting route collocation index value, CMAP 34

Table 9. Results of a linear regression predicting route collocation index value, ARC 35

Table 10. Pearson's correlation coefficients between original and masked KDE 36

LIST OF FIGURES

PAGE

Figure 1. Hypotheses for privacy and pattern preservation. 12
Figure 2. CMAP GPS travel survey area. 14
Figure 3. ARC GPS travel survey area. 15
Figure 4. Example GPS obfuscation results with a 100-foot distance threshold. 20

ACKNOWLEDGEMENTS

This research was made possible by the National Renewable Energies Laboratory (NREL) and the Transportation Secure Data Center (TSDC), which provided access and support for the data utilized in this analysis.

I am very grateful to Dr. Piotr Jankowski for his support of this topic and in his insightful comments and questions on my drafts. I am also grateful to Dr. Ming-Hsiang Tsou for bringing new ideas in privacy and GIS to my attention and his support in this pursuit. I am looking forward to continuing to work with you both as I pursue my doctorate. Thank you also to Dr. Carl Eckberg for your advice and feedback in this analysis.

CHAPTER 1

INTRODUCTION

In this digital age of volunteered geographic information (VGI), sensor data, and robust computing power, there is improved access to vast quantities of location data. Powerful spatial data visualization tools are freely available online, and a willing public liberally shares location data on social media platforms such as Foursquare, Twitter, and Facebook. VGI is put to use in disaster response on platforms such as Ushahidi, which allows users to plot information on incident locations or sites where help is needed. Location-based services (LBS) on smartphones allow consumers to locate nearby products and services, often in exchange for their location information. The numerous forms of VGI and their precision, often at the point level, provide a rich data source for analyses in social science, disease etiology, transportation studies, and market research.

Of concern in this age of big data is the potential for privacy to be breached through the disclosure of location information. Locational privacy, or geoprivacy, is a person's right to protect their identifying location information from disclosure, or to determine how and to what extent that information is shared with others (Kwan et al. 2004; Elwood and Leszczynski 2011; Kar et al. 2013). Location information is a strong personal identifier, particularly if multiple locations are provided for an individual. A study of fifteen months of location data for 1.5 million European residents finds that just four location points over fifteen hours are required to determine not only residences, but the unique identities of 95% of individuals (de Montijoye et al. 2013). In addition, the combination of birth date and zip code alone is found to be a unique identifier for 87% of the population (Sweeney 2002). Home and work locations can be inferred with ease using geotagged tweets and land use data (Li and Goodchild 2013). While the precision of data provided makes it more valuable for analysis, a fine degree of precision also increases the risk of revealing identities. Obfuscation, or masking, is a solution for mapping that aims to balance between the integrity of the data and the preservation of privacy. The goal of this study is to determine which masking techniques and associated distance thresholds are best applied to GPS data across different geographies.

JUSTIFICATION

Compelling evidence of the importance of locational privacy comes from scenarios in which the right has been violated, leading to harm. The U.S. Justice Department estimates that in 2009, more than 34,000 adults were victims of stalking using GPS (Baum et al. 2009). A website called pleaserobme.com highlights the dangers of posting personal location data on social media, underscoring how textual information about event, vacation, and restaurant plans can be cross-referenced with Google Maps and Street View to provide burglars with an itinerary for crime. While it is possible that robbers used pleaserobme.com to target empty houses, documented harassment cases include those from several users whose content was posted and cross-referenced on the site. These victims received calls at restaurants and events they checked in at with disturbing anonymous messages that they should keep their data private (Riordan Seville 2010; Herzog 2010). This website was used on multiple occasions by perpetrators to harass contributors of geotagged content by placing calls to restaurant and event staff to reach victims. A teenage girl who died from use of an acne medication was identified by media through the linking of a prescription data set with newspaper obituaries (Malheiros 2009). In 2012, a Los Angeles man tweeted the incorrect Florida home address he thought to belong to George Zimmerman, an address which was later reposted by celebrity Spike Lee (Jacobson 2012). The misappropriated home location in reality belonged to an elderly couple that felt compelled to relocate following an inundation of hate mail and reporters.

Many of these examples rely on VGI, which can be more difficult to regulate than location information coming from authorities (Li and Goodchild 2013). In February 2014, it was revealed that the Tinder mobile dating application allowed users to track each other’s locations for several months in 2013 (Burns 2014). Tinder provided exact distances from other Tinder users, which allowed for triangulation of user distances to obtain the exact coordinates of any person on the application. In online crime maps, which are often available to the public, comprehensive documentation of sexual assaults can place where rape occurs in the home, thereby violating non-disclosure regulations for sex crimes (Monmonier 2003).

GPS data are increasingly collected and shared, heightening vulnerability to identification. A Seattle company called Placed recruited 125,000 users who agreed to provide GPS data from their cell phones in exchange for occasional $5 gift cards (Robison 2014). The term cybercasing has emerged in the literature, which refers to the planning of physical attacks based on online geotagged data (Friedland and Sommer 2010). Privacy International raised concerns about the now-retired Google Latitude, which shared updated location information with friends on a continuous basis. The application once made it easy to enable tracking on another person's device, a technology which could have been employed by jealous spouses, overbearing employers, or stalkers (Gaudin 2009).

Despite privacy concerns, VGI and the ability to link to other repositories of information based on location have great utility for emergency response, navigation, disaster relief, health, and social research purposes (Duckham and Kulik 2007). Of critical importance is the balance of the utility of the data for research with the safeguarding of confidentiality (Vicente et al. 2011). Overall, the literature on locational privacy for participatory GIS leaves a sizeable gap for implementable guidelines and technology solutions for locational privacy infringement (Krumm 2007). Obfuscation, or masking, of point data by altering accuracy or precision is one means by which sensitive location data can be processed and displayed to protect identities. Obfuscation is preferable to aggregation to larger polygon boundaries such as census tracts or zip codes because it is better able to preserve spatial pattern (Allshouse et al. 2010). Due to the modifiable areal unit problem (MAUP), attempting to draw conclusions about a phenomenon operating at a finer level of geography than the one being analyzed is an error-prone pursuit. Many spatial phenomena, including transportation and disease patterns, cross administrative boundaries and would be distorted in aggregation. This study examines the balance between privacy and spatial pattern preservation in GPS data sets under the application of two masking methods: grid masking and Gaussian noise perturbation.

CHAPTER 2

BACKGROUND

This chapter provides context for historical and present-day debates on privacy, the development of the concept of geoprivacy, and how these concepts are negotiated in the landscape of big data. Motivations for volunteered geographic information (VGI) contributions are discussed, as well as tradeoffs between privacy and data access. Regulation strategies for preserving privacy are reviewed with a particular focus on GIScience-based technology solutions. Successful studies in masking and obfuscation for point maps are identified along with gaps in the current literature on GPS masking.

Conceptualization of Privacy

In 1994, Onsrud et al. introduced one of the capstone overviews of privacy in geographic information systems, citing United States legal definitions, including “the right of the individual to be let alone” and “the right to one’s personality,” stemming from Supreme Court Justices Warren and Brandeis (1890). More recently from the legal realm, Solove (2007) posits privacy as a set of “family resemblances” of persons and information, drawing from a taxonomy of four categories. These four classifications include information collection, information processing, information dissemination, and invasion, the last portion of which refers to intrusion into an individual’s life. From a geography perspective, Goss (1995) conceptualizes a privacy infringement as the unnecessary or unjust revelation of individual identity through the release of personal records. The specific concept of location privacy is referenced in more recent articles regarding GIScience (Zhong et al. 2007; Krumm 2009; AbdelMalik et al. 2008; Elwood and Leszczynski 2011; Kar et al. 2013) as the right of individuals to determine how and the extent to which their location information is shared with other parties. Similarly, there is an increasing number of references to the term geoprivacy, or the right to prevent the undesired disclosure of personal locations and activities (Kwan et al. 2004; Nouwt 2008; Li and Goodchild 2013). Given the applicability of this definition to obfuscation, this definition of geoprivacy is employed for this study.

EVOLVING PRIVACY CONCERNS IN GIS

Research from the mid-1990s discusses the emerging threat to privacy from the practice of geodemographics and the cross-referencing of data from large databases. Goss (1995) describes the challenges to privacy caused by cross-referencing data between public and private spheres and the stereotyped characterization of the individual based on geographic location. Similarly, Curry (1997) calls for a rethinking of the concept of privacy based on the evolution of the digital individual, or the profile created of individuals grounded on data matching and marketing through geodemographics. Both Goss (1995) and Curry (1997) attribute the dangers of privacy infringement as being aided particularly by GIS, which allows the tracking of individuals through new geographic data matching, especially with a constant stream of data.

The GIS and Society debates over privacy (Goss 1995; Curry 1997; Crampton 1995; Pickles 1995) invoked a series of articles on the concept of the surveillance society. These influential studies include riveting terms such as the "panopticon," a symbol of total surveillance and control (Dobson and Fisher 2007) and "geoslavery," in which an entity monitors and exerts control over the location of an individual (Dobson and Fisher 2003). Shilton (2012) echoes that the collection of individual data by authoritative entities has been typically labeled surveillance. Today, the bitterness with which privacy violations in GIS had been debated has lessened, while privacy concerns are growing in an active research field (Goodchild 2011).

A related theme is the commoditization of location data. Sui (2004) writes that the emergence of location-based services (LBS) signifies the commoditization of location. This means that location has attained a value and is exchanged for trade. Monmonier (2003) cites location-based restaurant-finders and the sale of coordinates of wireless subscribers as examples of the commodification of location. Sui (2007) elaborates on this, citing that while GIS previously represented media in a primitive stage according to McLuhan's stages of media, LBS has moved GIS to the mature stage of media. This is because LBS primarily functions through wireless connections and can be configured to access previous forms of media. This process has imbued a sense of an "information commons" and furthered the annihilation of space by time with mobile services making distance less important (Harvey 1990). The development of a capitalist system of exchange for location data may contribute to the exploitation of individual location data in the drive for profit. This poses a particular challenge for locational privacy.

MOTIVATIONS FOR VOLUNTEERING GEOGRAPHIC INFORMATION

Despite the darker concepts of exploitation by data commoditization and the idea that a knowledgeable entity could be tracking and targeting them, some individuals are contributors of copious amounts of geographic data. Volunteered Geographic Information (VGI) may include personal location data, such as geotagged tweets, or the submission of the locations of physical entities, such as with OpenStreetMap. A subcomponent of VGI may be considered citizen science, which typically refers to the contributions of data volunteers with more advanced skills, just as in OpenStreetMap (Goodchild 2007). Another example of this is bird watchers who participate in a volunteered Christmas Bird Count. Goodchild (2007) indicates that self-promotion, personal satisfaction or self-fulfillment, and connecting with friends are key motivations for locational contributions in citizen science. Self-promotion is evident as a motivator in a study of fitness-related VGI. With MapMyRide.com and MapMyRun.com, fitness and GIS enthusiasts can create maps of their favorites exercise paths and share them with online groups or the public. The functions of the two sites include map creation, a "route genius" that suggests new user routes, training plans and tools, as well as a social element, encouraging group events and city exploration. Kessler (2011) finds that the user motivations of self-promotion and mapping interests hinder the utility of MapMyRide.com and others to bicycling communities because community interest does not come first. He concludes that there is little proof that such data can be applied for urban road planning.

Still, numerous studies make use of "citizens as sensors" or citizen science to draw insights for transportation, health, politics, and other social science research questions. The European Union has funded five major projects based on Citizens' Observatories, in which citizens act as sensors and volunteer their location data, often from cell phones and GPS, in pursuit of the greater good. One such project is CITI-SENSE, which uses participatory sensing of environmental components and air quality to support decisions on environmental regulations (CITI-SENSE 2014). Citizens as sensors on social media platforms have also been leveraged to predict and respond to immediate emergencies, such as earthquakes

(Crooks et al. 2013). Motivations for contributing VGI are likely dependent on the platform, the type of information contributed, user skills, and attitudes toward technology.

TRADEOFFS BETWEEN PRIVACY AND DATA ACCESS

There is a growing body of literature on the competing values of privacy and data access. In a survey of health professionals, AbdelMalik et al. (2008) find that locational privacy concerns are overwhelmingly viewed as an impediment to proper health research, because data aggregation and barriers to access for fine-level data impede accurate spatial analysis. Allhouse et al. (2010) echo that analyses based on aggregated data in health research make it difficult to effectively allocate resources. Phenomena that move across administrative boundaries, such as disease outbreaks, are not successfully captured due to the modifiable areal unit problem (MAUP). In such research, point-level data are requisite. Precise location data sharing is also crucial to many popular location-based services, some of which could not exist without exact coordinate locations (Vicente et al. 2011). Most of the projects described in the previous section, including MapMyRide, CITI-SENSE, and geotagged tweets for earthquake response, would be of little use if aggregated to a larger geography.

The disclosure of point-level data, especially in interactive online mapping platforms, is potentially objectionable for several reasons. First, past or real-time location data can uniquely identify individuals, their homes, and their workplaces, potentially leading to crime (Friedland and Sommer 2010). The disclosure of sensitive information, such as disease, can make those identified more susceptible to harassment. Kwan et al. (2004) write that releasing individual data is unethical due to promises of confidentiality in most studies and hidden human subjects who may be impacted by the identification. Some laws are in place, especially in health research for protecting individuals. In particular, online crime maps showing rape that occurs at a victim’s home would violate non-disclosure laws for sex crimes (Monmonier 2003). The Privacy Act of 1974 and the Public Health Service Act of 1946 offer some protections of patient confidentiality that would preclude the posting of identifiable health information (Hampton et al. 2010). Most regulations in place that would extend to geoprivacy are in the realm of health. John Edwards introduced the Locational Privacy

Protection Act of 2001 to ban wireless providers from releasing customer location data without permission, but it did not pass (Monmonier 2003).

On the opposite side of privacy protection are the motivations of the public to contribute data (discussed in the previous section), the motivations of mediating entities to make the data accessible, and the motivations of researchers and hobbyists who want to access the data. Sweeney (2002) writes that the survival of the database itself depends on the publishing of anonymous data. His argument is that the promise of anonymity is precisely what makes data volunteers most likely to contribute. As the volume of available data sources increase through citizen science and other forms of VGI, however, the promise of anonymity may be difficult to sustain. Shilton (2012) argues that privacy is not an absolute right, but relative value, and that in participatory research, privacy must be weighed with other values, including accuracy and the greater good. Elwood and Leszczynski (2011) note that the aptness of locational privacy as well as its sustainability continue to be publically negotiated.

Approaches to Privacy Preservation

Given these tradeoffs and vulnerabilities, there are three categories of solutions for privacy protection: regulation, education, and technology. Regulation could include laws similar to the Privacy Act of 1974 or the proposed Locational Privacy Act of 2001, or include directives for posting locational data online. Onsrud et al. (1994) maintain that privacy guidelines within the GIS community should be established to avoid overreaction by the public and secure what has already been invested in geographic data collection. Friedland and Sommer (2010) suggest the solutions of further education for users of location-based services and enforcing privacy rules for public databases, rather than imposing limitations on the contributor side. The National Science Foundation (NSF) is currently funding a project to build ethics and privacy into GIS curricula across education levels (Carr 2013).

On the technology side, the National Research Council (2007) released an influential set of guidelines for preserving the privacy of point data, which includes recommendations for secure data enclaves with restricted access. Kar et al. (2013) discuss the history of regulatory strategies and policy mechanisms to protect privacy, but also highlight the technology solutions of anonymity and obfuscation. Anonymity (or pseudonymity) involves the separation of identifying personal information from location data (Kar et al. 2013; Krumm 2007). Obfuscation, which is the focus of this study, degrades the quality of spatial data in one of three methods, including introducing inaccuracy, increasing imprecision, and maintaining vagueness in descriptive terms, such as “far from” (Duckham and Kulik 2007). Obfuscation is intended to be applied both in published research reports as well as in online volunteered geographic information.

Obfuscation Techniques

Several categories of data masking techniques have been tested on both discrete location and GPS point data sets. Options for obfuscation include grid masking (Leitner and Curtis 2006), affine transformation (Kwan et al. 2004), and random perturbation (Kwan et al. 2004; Hampton et al. 2010; Gambs et al. 2010). Leitner and Curtis (2006) introduce grid masking as an obfuscation measure, which generates inaccuracy by transforming location points within grid cells of a given size. The authors conclude that there is a threshold cell size for both privacy and masking, above which larger cell sizes cause the unmasked pattern to be perceived differently. Affine transformations translate, expand, or contract a point pattern, maintaining relative positions (Kwan et al. 2004). Random perturbation involves relocating each point in an original data set a random distance in a random direction within a set distance threshold (Kwan et al. 2004; Hampton et al. 2010). Perturbation is the most frequently cited method of obfuscation for discrete point data, particularly in health research (Allshouse et al. 2010; Shi et al. 2009; Kwan et al. 2004; Hampton et al. 2010).

Preliminary masking work has also been conducted on GPS data. Krumm (2007) tests the efficacy of obfuscation in identifying residences in GPS data tracks through the addition of Gaussian noise in random perturbation and snapping points to the centroids of grid cells. Krumm calls for future work to elaborate assessment of obfuscation techniques with stronger algorithms to detect home locations in GPS data. Gambs et al. (2010) discuss a system designed for privacy preservation in GPS data sets that includes measures for perturbation. In this system, the user must set a distance threshold for perturbation. Neither study makes recommendations for masking thresholds to be implemented for GPS trajectories.

Gaps in Obfuscation Research

Overall, the literature on locational privacy in participatory GIS leaves a sizeable gap for implementable guidelines and technology solutions to privacy infringement. Krumm (2007) and Leitner and Curtis (2006) recommend that future scientific inquiry examine privacy protection mechanisms for volunteered or participatory spatial data. Shilton (2012) specifically calls for geographic privacy researchers to design and test privacy-aware participatory systems and the sensitivity of different forms of participatory data to privacy infringement. Kwan et al. (2004) and Hampton et al. (2010) note that aggregation of point data is currently the benchmark for privacy protection, but with better documentation of point obfuscation techniques and results, masking can bridge the divide between privacy preservation and accuracy of analysis. Some research has been conducted on random perturbation and grid masking distance thresholds for discrete point data (Kwan et al. 2004; Leitner and Curtis 2006; Curtis et al. 2011). These findings are discussed in detail in the methods section. No conclusions have yet been drawn on distance thresholds for masking GPS data.

Another gap in the literature relates to how underlying geography impacts the success of masking for privacy preservation. Kwan et al. (2004) and Allshouse et al. (2010) implement adaptive obfuscation distance thresholds that are weighted by population and housing density. The Kwan et al. (2004) study, however, does not measure privacy preservation once the mask is implemented. Furthermore, the Kwan et al. study finds that the obfuscation measures that are not weighted by population density yield better results, noting that different configurations can lead to different results. None of the GPS obfuscation studies have tested the success of obfuscation with regard to privacy preservation as a function of underlying population or road density.

CHAPTER 3

CONCEPTUALIZATION

As a response to the gaps in obfuscation research for GPS data and in acknowledgement of the tradeoffs between data quality and participant protection, the purpose of this study is to test the preservation of both privacy and spatial pattern in masked GPS trajectory data. The guiding research questions for this study are:

1. How effective are two obfuscation methods, a) grid masking and b) random perturbation, in protecting privacy and spatial patterns in trajectory data?
2. How does privacy preservation in trajectory masking vary by housing unit density, road density, and mode of travel?

This study examines two methods of location obfuscation, which are a) grid masking and b) random perturbation. Both types of masking are implemented in the greater metropolitan regions of Chicago and Atlanta with distance thresholds of 30 meters, 100 meters, and 250 meters. Spatial pattern preservation is measured using kernel density estimation on the original and masked data sets and calculating the Pearson's correlation coefficient for each permutation. The privacy metric is calculated using concepts of k-anonymity of both home location and route. Home k-anonymity is determined by the occupied housing unit density where the masked point is placed. Route k-anonymity is determined via colocation percentage with other persons in the database. These techniques are discussed in detail in Chapter 4.

HYPOTHESES

The first hypothesis for this study is that the higher the distance threshold implemented for both grid masking and random perturbation, the greater the privacy protection offered by k-anonymity for both home locations and routes. An individual in a given data set is "k-anonymous" when he or she cannot be distinguished from at least k-1 other individuals (Sweeney 2002; Krumm 2007; Zhong et al. 2007). Particularly in regions with low household density, a higher magnitude of point location displacement through masking would be beneficial in moving a home location away from the original sparsely populated region. The higher distance threshold is also beneficial for the anonymity of the

GPS trip as a unit, since while the general travel pattern is preserved, it becomes more difficult to decipher which road a person travelled on. This is particularly true in regions of high road density. At least two previous studies (Kwan et al. 2004; Allshouse et al. 2010) implement weighted masking based on population density and household density, increasing the magnitude of perturbation as the population density decreased. While their studies build k-anonymity into the obfuscation design, this study assesses the outcome of the masking thresholds in terms of k-anonymity.

The second principal hypothesis is that as the distance threshold increases in masking, pattern preservation and accuracy will decline across the overall GPS data in the study regions. This expectation is well-supported in previous studies of masking residence location data as well as GPS trajectory data. As the radius for random perturbation in the Kwan et al. (2004) study increases, the cross-k function used to compare clustering patterns results in lower correlations. Aside from spatial pattern degradation, Gambs et al. (2010) find that the average speed of GPS trips calculated with time kept constant and distance altered from data perturbed with a standard deviation of 50 meters was 5% higher than for the original unmasked GPS data. It would be unexpected to find a comparatively higher coefficient of correlation with a higher distance threshold for obfuscation. Such a finding would suggest that the sample size is not large enough or that random perturbation was not fully or uniformly random. These first two hypotheses are summarized in Figure 1.

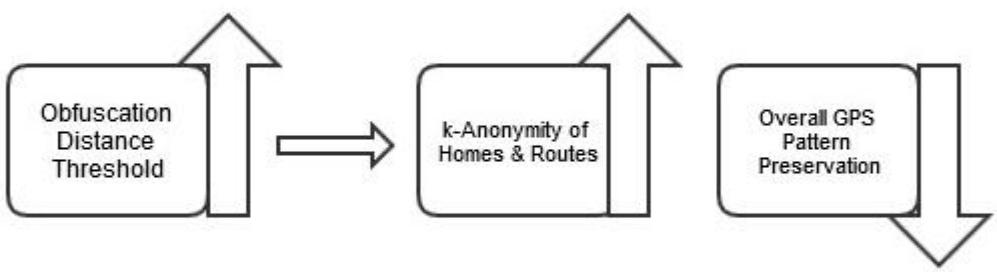

Figure 1. Hypotheses for privacy and pattern preservation.

A third hypothesis is that areas of lower road density will be correlated with GPS trips that exhibit a higher degree of k-anonymity. This is because a higher density of roads suggests that there are more unique routes an individual can take in that area compared to neighbors (Pingley et al. 2009). If a GPS waypoint is recorded in an area with very low road density, it is likely that no matter the degree of obfuscation, the point will still be associated with the same roadway. Low densities of road connecting regions of high population are likely to reduce route uniqueness, with many participants taking the same route.

Fourth, the trajectories of different modes of travel are expected to exhibit varying levels of route k-anonymity. Non-vehicle traffic, including bicycle and foot traffic, is expected to result in greater route uniqueness with waypoints closer to individual homes and workplaces. The data sets employed in this study are GPS points from wearable GPS units and have recorded trips of multiple travel modes. Since points are collected for every one second of travel, a greater concentration of waypoints is found closer together for walking trips compared to vehicle trips. Therefore, trips with slower average speeds are more uniquely identifiable and attributable to individuals than vehicle traffic along highways.

Study Area

The study area selected for this research is comprised of the greater metropolitan regions of Chicago and Atlanta. The greater regions Chicago and Atlanta have internally varying but comparable population densities and adequate ranges for home density and road density analysis. The advantage of employing broader metropolitan regions rather than strict city boundaries for this study is the opportunity to compare results across varying levels of urban and rural composition and household density. As Metropolitan Planning Organizations, the Chicago Metropolitan Agency for Planning (CMAP) and the Atlanta Regional Commission (ARC) conduct large-scale travel behavior studies across their regions to model and analyze travel data for planning purposes. GPS data loggers are increasingly deployed in travel activity surveys to supplement and replace traditional travel diary collection methods (Wolf et al. 2001). The 2007 CMAP Travel Tracker Survey and the 2011 ARC Household and Activity Travel Survey both incorporated GPS data logger technology to collect everyday travel information from residents in the regions.

Chicago

The city of Chicago is the third largest in the United States, with 2.7 million residents, and the population of the Chicago metropolitan area is close to 10 million (Census Bureau 2012). This speaks to the importance of this region for representation in volunteered geographic information and social media. The Chicago study area in the 2007 CMAP travel survey encompasses eight counties in Illinois, as shown in Figure 2 (NuStats 2008). The mean population density of this region is 5,423 persons per square mile with a standard deviation of 1,550. CMAP's GPS travel survey included both vehicle GPS loggers and wearable loggers that were intended to capture all modes of travel. The wearable GPS study was conducted between September 2007 and January 2008, resulting in 209 participants travelling during 7-day study periods.

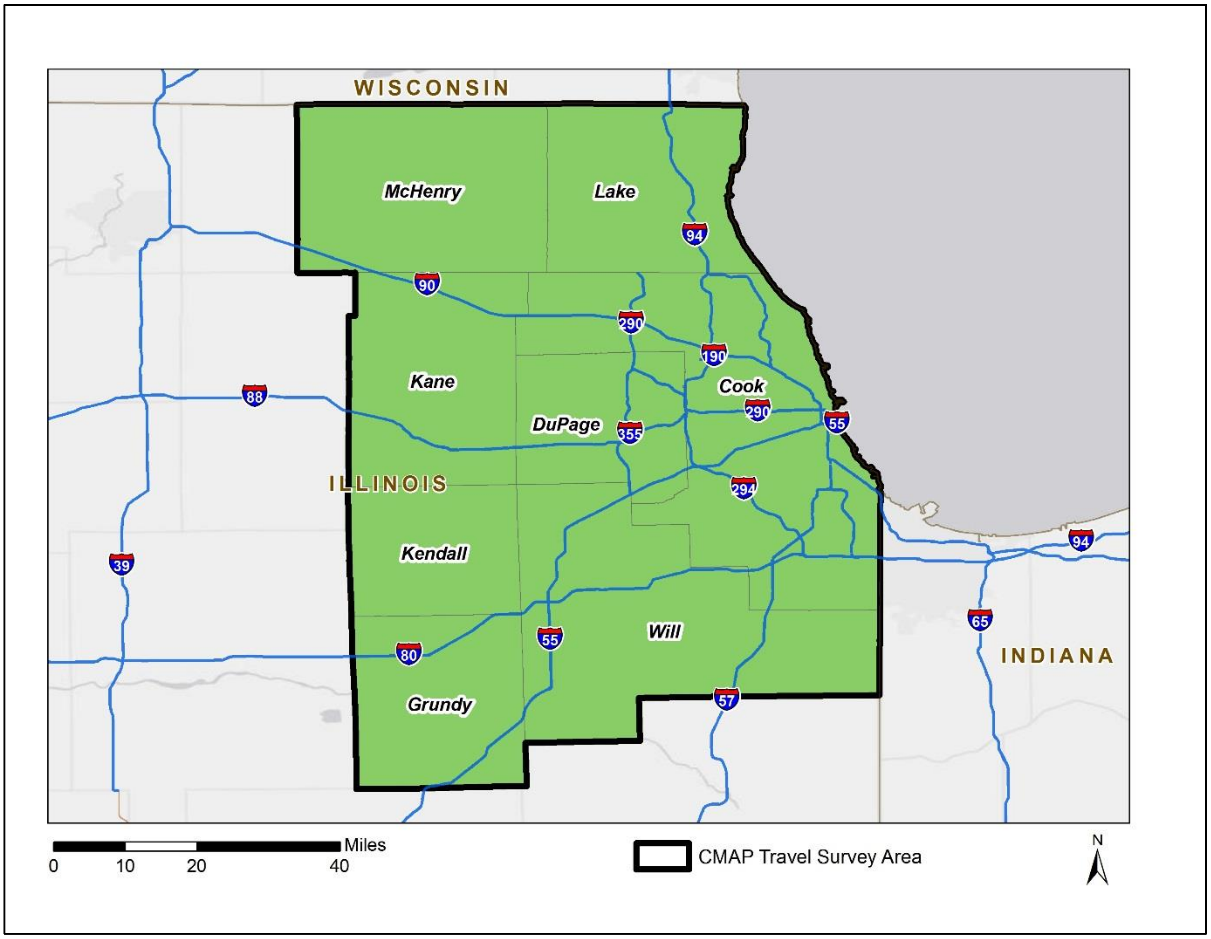

Figure 2. CMAP GPS travel survey area.

Atlanta

Atlanta has a much smaller city population at approximately 400,000 inhabitants, but the metropolitan region has over 5 million residents (Census Bureau 2012). The extent of the 2011 ARC travel survey is shown in Figure 3 and includes twenty counties in Georgia (PTV NuStats 2011). These counties are considered to be the commute-shed for the city of Atlanta. The mean population density of this twenty-county region is 826 persons per square mile with a standard deviation of 698. The population density is thus much lower than that of the

CMAP region. These cities of differing population characteristics and removed from each other by approximately 750 miles present a unique opportunity for GPS data privacy comparisons. The wearable GPS data logger survey was conducted between March 2011 and September 2011 with travel periods of 7 days, just as in the 2007 CMAP study. The wearable GPS resulted in data collected from 797 individuals. Recruitment for the wearable GPS study was targeted in the four counties of DeKalb, Gwinnett, Cobb, and Fulton.

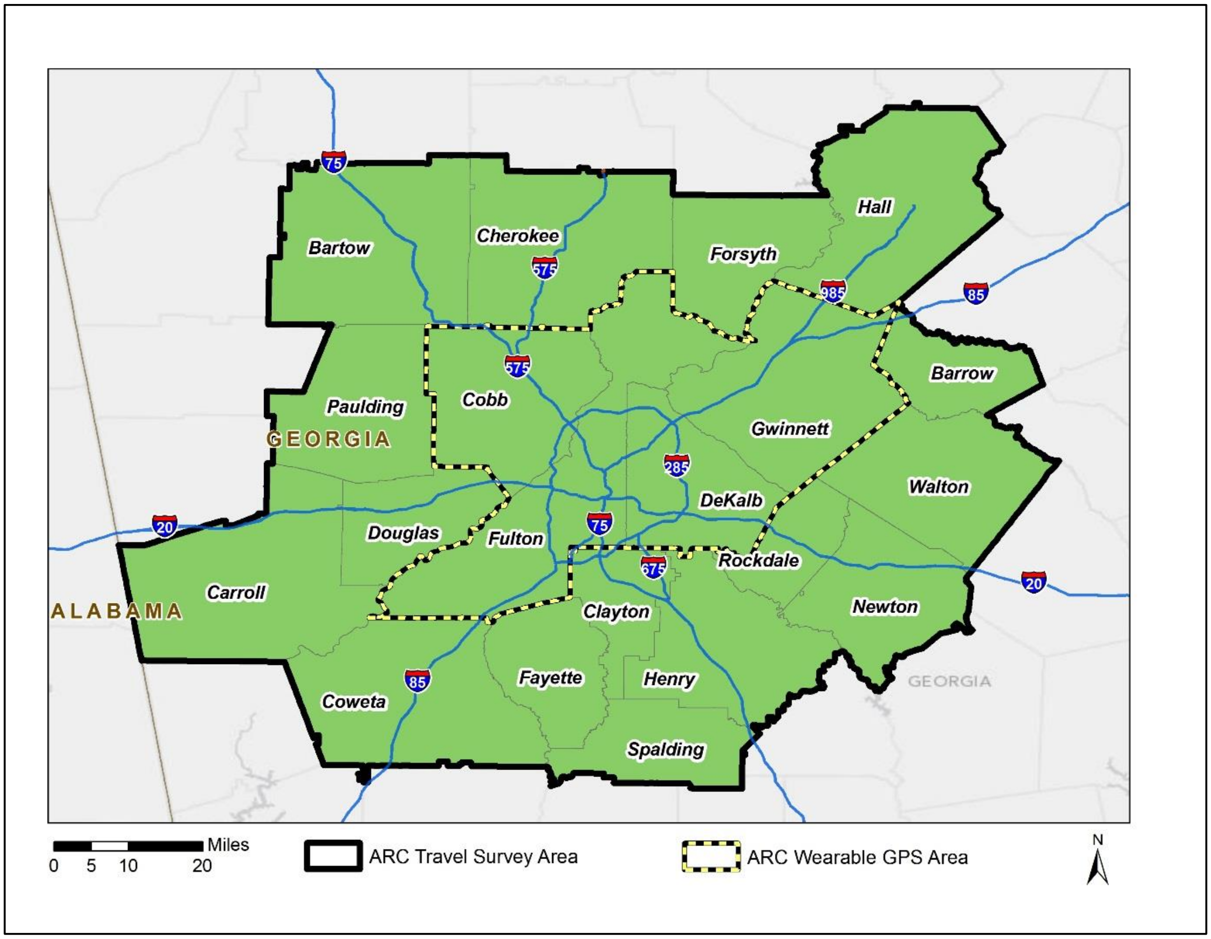

Figure 3. ARC GPS travel survey area.

GPS Data Access

Access to these GPS data sets is available through the National Renewable Energy Laboratory's (NREL) Transportation Secure Data Center (TSDC) (National Renewable Energy Laboratory 2013). The TSDC provides both cleansed summaries of the results of various regional travel behavior surveys to the public, as well as spatial data access through a secure remote environment following an application process. This restricted environment follows recommendations from the National Research Council's 2007 report on protecting confidentiality in spatial data and does not permit internet access or the addition of or copying files externally. Any data or software tools to be added must be reviewed and approved by an administrator, and only aggregated report summaries are approved for being sent out of the remote environment. The TSDC currently offers spatial travel behavior data from the California statewide, Atlanta, Texas, Minneapolis/St. Paul, Chicago, Puget Sound, and southern California regional travel surveys. The Chicago and Atlanta data sets are currently the only two that provide seven days of GPS data per participant.

The TSDC remote access environment stores the GPS data in PostGIS databases, which are directly accessible using the open-source software QGIS, available in the remote environment. Among the other tools available are R, python, ArcGIS, Spyder, and Microsoft Office. Due to limited availability of licenses for ArcGIS, as this environment is shared with other NREL employees, QGIS was used for all GIS analysis in this study. Since an ultimate goal of this study is to make obfuscation techniques readily deployable and implementable across a range of users and geographies, full reliance on open-source tools is advantageous. It is important to note that the storage of GPS data points is challenging due to large file sizes and limited space. Successfully completing this study entirely in the remote access environment meant that intermediate results had to be continually deleted to free up space. Overall, the TSDC user environment was fast, responsive, and provided an adequate number of tools for this analysis.

CHAPTER 4

METHODS

This section describes the methodology employed in masking the GPS data, determining levels of privacy preservation, and evaluating the maintenance of spatial patterns between the original and masked data sets. The obfuscation techniques employed in this study are grid masking and random perturbation. Privacy preservation is evaluated by adherence to the principle of k-anonymity, or the principle that each feature or route must be indistinguishable from that of k-1 other individuals. The preservation of spatial pattern is determined by Pearson's correlation coefficient run on kernel density estimations for the original and masked data sets.

OBFUSCATION

The GPS trajectories in this study are masked by one of two methods: grid masking and random perturbation. The goal of obfuscation is to strike a balance between data quality and privacy preservation. This subsection provides context on other masking studies and distance thresholds selected, informing the choice of obfuscation methods for this study.

Grid Masking

In grid masking, a grid of a specified cell size is overlaid with the point data to be masked, and each point is snapped to or transformed within its corresponding grid cell (Leitner and Curtis 2006; Krumm 2007). Snapping can either be to the centroid of the grid cell or to a corner point. The cell size of the grid varies according to the degree of privacy desired. Larger cell sizes allow for enhanced identity protection, but are more likely to alter spatial patterns. Leitner and Curtis (2006) first implemented grid masking on mortality data in Baton Rouge, LA, concluding that there is a threshold cell size of 30 meters by 30 meters, above which the larger cell size causes the spatial pattern to be perceived differently. Krumm (2007) conducts grid masking by snapping GPS waypoints to the nearest point on a 50 meter by 50 meter grid. Curtis et al. (2011) implement grid masking with cell sizes of 1,000 meters, 750 meters, and 500 meters, concluding that compared to control simulations, there is little risk with any of these cell sizes for manual identification of the original points. This suggests that cell sizes smaller than 500 meters should be tested for privacy preservation in the interest of maintaining spatial patterns for analysis. It is important to note that a constant 500-meter threshold can still lead to varying results by settlement pattern, and the results in one city are unlikely to reflect those in a small town or rural area. The Curtis et al. (2011) threshold of 500 meters is considerably larger than the Leitner and Curtis (2006) 30-meter threshold. This is because Curtis et al. focus more on preventing identification than on maintaining spatial pattern and seek to prove that these masking thresholds are better for privacy than aggregation to zip code. Leitner and Curtis, on the other hand, test for preservation of spatial pattern to balance with the intended privacy protection. In all of these studies, a blanket distance threshold for grid masking is set, regardless of the underlying regional attributes.

Bernheim Brush et al. (2010) also implement grid masking when testing public preferences for the masking of their own trajectory data. The study area is transformed into a grid, and only the grid cells that a given trajectory traverses are displayed, rather than snapping waypoints to centroids or vertices of the grid. This study does not test any specific cell sizes, but rather the preference of users for this masking technique compared to others. The cell sizes implemented in this study must not be so large that the trajectories are not coherent, and 500 meters is about the size of 9 city blocks. Therefore, the cell sizes tested for grid masking in this study are 30 meters, 100 meters, and 250 meters, most closely resembling the magnitudes selected by Leitner and Curtis (2007). Trajectory points in this study are snapped to the closest vertex of the grid cells.

Random Perturbation

The second method tested for masking the GPS data is random perturbation. In this type of obfuscation, each point is moved in a random direction and distance within the confines of a distance threshold. Kwan et al. (2004) refer to this as random perturbation, or the introduction of random error to each point, varying the magnitude and the direction of the location reassignment. The displacement can be thought of as setting a radius for a circle with each original point in the data set as the centroids. Each displaced point will fall somewhere within that circle. Krumm (2007) introduces the term Gaussian noise for this technique, because in his study, the randomization of GPS points follows a Gaussian distribution from 0 to σ^2 in magnitude. The radii Kwan et al. (2004) test in their random perturbation are 98 feet (30 meters), 915 feet (279 meters), and 4,273 feet (1,302 meters), which create areas that correspond to block groups and census tracts in their study area. Shi et al. (2009) refer to this method as dithering, using the radii of 100 meters, 500 meters, 1000 meters, and 2000 meters. The results of the Shi et al. (2009) study are that the preservation of spatial pattern with these degrees of dithering are highly dependent on the bandwidth used in kernel density estimation, which must be equal to or greater than the perturbation threshold. Gambs et al. (2010) also implement perturbation with Gaussian noise to mask the mobility traces of taxi drivers. Gambs et al. find that with a standard deviation of 50 meters, it is still easy to identify the homes of taxi drivers in San Francisco. With a standard deviation of 200 meters, the quality becomes too degraded to see the home location, but possible to see the neighborhood of the taxi drivers.

An alternative is to mask with the distance threshold weighted by population characteristics in the region. Kwan et al. (2004) test masking with distance thresholds weighted by the underlying population density, but the results for privacy and spatial pattern preservation are no better than for the “blanket” masking thresholds. This is potentially related to the weighting factors selected for population density. In this study, the blanket masking approach is applied in order to subsequently analyze any differences with respect to the original point distribution for housing density. Determining these differences will allow for more informed weighting figures to be developed for masking thresholds.

Another variant of random perturbation is donut masking, which ensures that the masked points are moved some minimum distance from the original points (Hampton et al. 2010). This requires an input parameter as the inner radius for perturbation, along with the outer radius. Random perturbation within the confines of the donut would place the masked point a random distance and direction from the original point between the minimum and maximum radii specified. Hampton et al. (2010) test donut masking with the radii kept variable based on the underlying population density at the point. Likewise, Allshouse et al. (2010) vary the radii for their donut masking by the number of households and area of the block group in which the point is located. However, in some cases the implementation of an inner radius could hamper both privacy and spatial pattern preservation. First, it limits the potential area of displacement by making the ring smaller. Second, allowing for some points to randomly remain at their original positions maintains uncertainty in the masked data set and does not limit the mask to only areas where the original point is not located. Therefore, this study does not implement an inner radius for donut masking. Rather than favor pattern preservation over privacy preservation (Leitner and Curtis 2006) or identity protection over preservation of privacy (Curtis et al. 2011), this study selects middle ranges between 30 meters and under 500 meters hypothesized to balance between these two concerns. Furthermore, since GPS data differs in structure than standalone point data, the test thresholds are set to lower than those in other studies due to the necessity of a greater degree of coherence for proper measure of route collocation and anonymity. Given these considerations, the radii implemented here are 30 meters, 100 meters, and 250 meters, matching the distance thresholds for the grid masking component of this study. The lower threshold of 30 meters is tested in this study because of the high density of settlement patterns in the Chicago region. Examples of original unmasked GPS points and their obfuscated counterparts with a 100-foot distance threshold are shown in Figure 4.

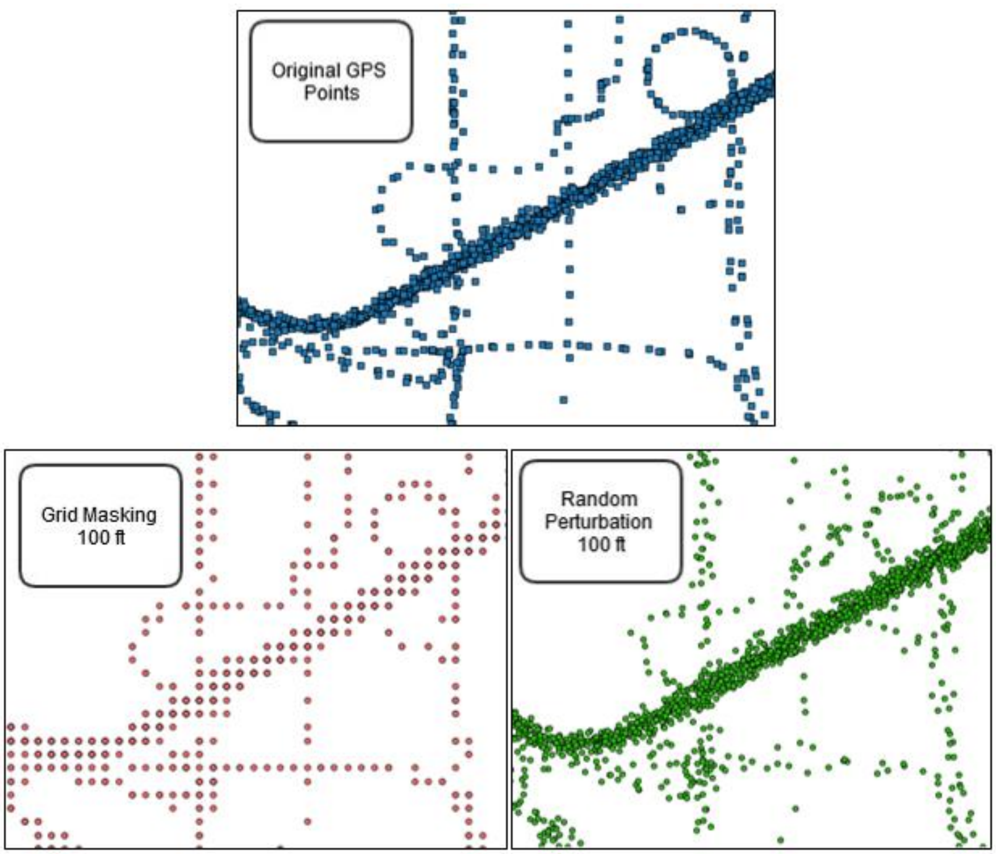

Figure 4. Example GPS obfuscation results with a 100-foot distance threshold.

k-Anonymity

The concept used in this study for evaluating the preservation of privacy between the original and masked GPS data sets is k-anonymity. According to Sweeney (2002), who introduces the concept, k-anonymity ensures that the data for an individual is indistinguishable from k-1 other individuals in a given data set. This concept has been increasingly applied in masking studies as a means of both measuring the anonymity generated or maintained in masked data sets and generating radii distances for perturbation. Hampton et al. (2010) engineer their donut masking study to create masking radii based on the concept of spatial k-anonymity, specifying a number of persons between which a health or disease cluster cannot be reversely identified to generate the radii for each point.

Point Anonymity

Research teams have measured the k-anonymity of persons and households of masking studies with varying metrics. Allshouse et al. (2010) measure k-anonymity in their donut masking study as the total number of households in the circle formed with the original point as the centroid and the radius as the distance to the masked point. A potential issue with this measurement is that the public would only be viewing the masked data set, and would make assumptions about where the masked point falls. It is important that the masked point then have an appropriate underlying household density. Hampton et al. (2010) calculate k-anonymity by extracting population density values for each point from a raster of population density.

Another option is to base point k-anonymity on the number of households, or household density, contained within a buffer of each masked point. This could be determined using a GIS layer of residential parcels with a buffer of a set distance from the masked point. If there is more than one residential parcel centroid within each masked home buffer, the masked point can be said to be indistinguishable from at least one other home. There are several issues with this approach, however. First, parcel layers may not include information on the type of residential land use. Second, the goal of masking for privacy preservation is not limited to the release of spatial data from academic and professional studies; it extends to other forms of VGI that would have otherwise been identity-revealing point data, such as check-ins at home locations. In order to encourage the widespread application of masking techniques in VGI, all the resources to server as input parameters for masking should be free to use and easily accessible. Complete parcel data are often made available only by request through municipalities. It can be challenging to ascertain information on the number of households residing in each parcel, and the accessibility of these data across the U.S. is inadequate at this time. A second issue is that buffers and intersections are computationally expensive operations in GIS and are not advisable for the large quantities of point data this study evaluates.

An alternative approach is to use household density or population density as a proxy for k-anonymity for home locations. In designing a k-anonymity-weighted masking threshold, Allshouse et al. (2010) set the displacement distance as dependent on the number of households within the area of the block group where the household resides. Kwan et al. (2004) also assign k-anonymity levels from census block groups to each point in the weighted masking. Based on the consistency of this approach across the geography (Kwan et al. 2004) and health disciplines (Allshouse et al. 2010), this study opts for the less computationally expensive technique of evaluating the privacy preservation of masked points from the density of occupied housing units in the block group of the masked home locations. In this study we are interested in maintaining the confidentiality of only home locations, though we acknowledge that work, school, and other frequented locations may serve as identifiers. The same methods applied to home locations here may also be applied to other origins and destinations in the GPS data.

Route Anonymity

This study also tests the levels of privacy maintained in trajectory data, which each trip considered a separate entity. Nergiz et al. (2009) introduce the concept of trajectory k-anonymity, which means that every trajectory released in a set of data can be tied to at least k participants in the database. In this case, k is representative of travelers of routes, rather than of routes. Just as for individual locations, which should be attributable to more than one person or household, each trajectory should be attributable to at least one other person in the data set. The goal in this part of the analysis is to identify collocated travel patterns and group trajectories by similarity in geography and time. The literature reveals multiple options for clustering similar routes together.

Multiple categories of clustering algorithms are available, including partitioning, density-based, grid-based, model-based, and hierarchical clustering, as well as clustering based on constraints (Braga et al. 2012). Density-based Spatial Clustering of Applications with Noise (DBSCAN) is one commonly applied density-based clustering algorithm, which discovers spatial clusters of arbitrary shapes and requires the input parameters of a minimum number of neighbors for each cluster and a spherical neighborhood threshold (Ester et al. 1996; Richardson and van Oosterom 2002). Another common density-based clustering algorithm is Ordering Points to Identify Clustering Structure (OPTICS), which has been successfully implemented in the clustering of trajectories (Braga et al. 2012; Andrienko et al. 2007). Braga et al. (2012) apply OPTICS to trajectory data, storing core distances and reachability distances, in combination with minimum bounding rectangles and the Hausdorff distance to group trajectories. The OPTICS thresholds Braga et al. (2012) apply for their study area are a distance of 1,000 meters and 3 minimum neighbors. The minimum bounding rectangles are established from the most extreme values (i.e. northernmost, westernmost) of each trajectory, and each rectangle is compared to those of other trajectories in the data set. For where the minimum bounding rectangles overlap, the Hausdorff distance is calculated between the trajectories. The Hausdorff distance is used to compare one linear feature to another, and has been used elsewhere in applications to assess the positional accuracy of digitized polylines (Goodchild and Hunter 1997). A disadvantage of this distance is that it is sensitive to outliers.

A similar clustering procedure comes from Andrienko et al. (2007). In this case, OPTICS is implemented to determine points of interest from raw GPS data, and the clustering of routes is approached using a “common route” function, in which pairs of trajectories are scanned for the closest pair of points. The mean distance between positions and a penalty distance are calculated for each pair during scanning. Clustered routes or trips in the Andrienko et al. study are created with a distance threshold of 250 meters. A variant of this function accepts time as a parameter so that the routes can be clustered spatiotemporally. A disadvantage of this procedure is that each point of every GPS trip must be compared to every other, which can be computationally intensive.

Another similarity measure is to determine the longest common subsequence (LCSS) for each pair of trajectories (Meratnia and de By 2002; Vlachos et al. 2002). The LCSS is appropriate for trajectory data where there would be outliers and noise, such as with GPS data, and a similarity function is run to determine the length of the LCSS. The dynamic comparisons of each trajectory pair again make this technique computationally expensive. Nergiz et al. (2008) minimize the log cost metric of two trajectories in order to group trajectories and determine a group representative from which to determine nearest neighbors of trajectory points. Minimum bounding boxes are then formed around each group of trajectory points.

A less computationally expensive method for calculating trajectory k-anonymity is preferable given the quantity of iterations necessary for each masking technique and threshold. Another solution comes from Meratnia and de By (2002), who demonstrate the use of a spline raster representation of trajectories. Meratnia and de By propose the generation of five rasters, one defining the number of hits per cell, and the other four referring to the number of movements out of each cell in the up, down, right, and left directions. Cell size in this aggregation must be carefully selected and is dependent on the maximum speed obtained in the GPS trajectory. Meratnia and de By propose that where cell size in meters is λ, maximum velocity is υ_{max}, and ρ is the re-sampling rate for the trip from a spline representation,

$$\frac{\upsilon\mathrm{max}}{\rho} \ll \frac{\lambda}{\sqrt{2}}$$

While Lin and Su (2008) critique this method for its focus on aggregation of routes, rather than similarity, our study is focused on finding out which routes fail to obtain threshold levels of k-anonymity. Therefore, an extension of this grid-based approach would be to calculate dissimilarity of routes based on relative trip length traveled through grid cells without k other trajectories. Another extension of this method is to aggregate the trajectories spatiotemporally in a specified time interval. This raster method is more efficient than clustering techniques that compare each trajectory to each other trajectory. In this case, each trajectory is only compared to the reference grid. This method also is less sensitive to variations in trajectory length and frequency of data collection within a trip. Trajectory k-anonymity for our study therefore relies on the raster methodology proposed by Meratnia and de By (2002). Kernel density estimation (KDE) is applied to the original and masked points to create reference grids of collocation values. The maximum average velocity in the data sets was 105 kilometers per hour, and the sampling rate of each point in the data set is one per second. Based on the Meratnia and de By cell size recommendation for these parameters, a cell size of 150 meters is chosen for the output kernel density rasters of collocation values. These collocation index values are then extracted from each corresponding raster cell to the GPS points. The extracted point density values then correspond to each second of travel. This process can be disassembled into multiple reference grids for each target time period. The average collocation of each trip with other trips for the duration of the trip in question is calculated to produce a mean collocation value for the trip. A challenge of this approach is that points from the same trip or person can occur in the same grid cell more than once. This is particularly true with GPS data collected of various modes of travel. Results will be discussed in the context of this limitation. The strong advantage of this approach is that the conversion to raster speeds up processing compared to a vector-based method.

Spatial Pattern Preservation

While optimizing a masking technique for privacy preservation, there is often a decline in consistency of the spatial pattern. Some studies appear to favor the preservation of privacy at the expense of the utility of the masked data produced. Leitner and Curtis (2006) favor a human-centered approach, asking university students in GIS to evaluate the similarity of two subsets of original and masked point data and rank them from very similar to very dissimilar. This perspective is useful, since it is often the human brain making decisions from viewing perceptible cluster patterns, rather than the use of statistics in all cases. Considering the large data sets in this study, a comprehensive human-based evaluation of spatial patterns would not be possible.

Instead, most masking studies test the preservation of spatial pattern using a clustering statistic. Kwan et al. (2004) implement the cross-k function, which examines the clustering of one point pattern compared to another. The cross-k function provides the expected number of points of a certain pattern within a set distance of an arbitrary point of another point pattern. The function runs in 100 simulations for 51 distances for each of the 3 sets of radii Kwan et al. use in their perturbation study. It is unclear why these numbers of simulations and distances are chosen for the cross-K function, and it appears that a different combination of simulations could be applied. Similarly, Hampton et al. (2010) implement a spatial scan statistic test with a circular scanning window in their donut masking study. The researchers tested the sensitivity in their Monte Carlo simulations by dividing the number of simulated cluster cases by the number of cluster cases they injected in their test. Wieland et al. (2008) also use a spatial scan statistic within SaTScan circular cluster detection software and record the p values for each cluster.

Shi et al. (2009) utilize an approach based on kernel density estimation (KDE). Instead of running simulations for cluster analysis, the authors create kernel density surfaces of varying bandwidths for the original and masked data points with no distance decay, and then calculate Pearson's correlation coefficient for each of the masked density surfaces from the original. Shi et al. (2009) find that when the bandwidth is less than or equal to the threshold used in perturbation, the density surface from the masked points is very different from that of the original points. When the bandwidth applied in the kernel density estimation was five times that of the masking distance threshold, the resulting patterns are found to be identical to those of the original points. This speaks more to the applicability of the KDE parameters in this research than to the underlying differences in spatial pattern.

All three of the statistical techniques outlined above rely on discrete point data as an input, rather than the linked points of trajectories. Thus, a method that introduces cluster simulations would have to mimic trajectory clusters. Eliminating the need for simulated clusters and relying on the spatial arrangement of the data is possible, as shown in the Shi et al. (2009) study. A density analysis for the GPS trajectory data, such as kernel density estimation (KDE), provides general surface trends, and is equally appropriate for GPS trajectories and standalone point data. This study therefore also implements KDE, and Pearson's correlation coefficient is used to test for statistical difference between the original and masked data sets. Following the guidelines of Shi et al., bandwidths of two times the lowest masking threshold distances, or 200 feet, are utilized in the kernel density estimations with no distance decay.

CHAPTER 5

RESULTS AND DISCUSSION

This chapter highlights the results of grid masking and random perturbation tests conducted on GPS data collected in the greater Chicago and Atlanta metropolitan regions. Descriptive statistics on the data collected in each region are summarized in Table 1. For Chicago, there were 209 total persons for whom GPS data were recorded with a wearable device. This resulted in 5,671 total trips within the study area and 856,465 total points of verified second-by-second trip location data. The mean population density for the block groups in which the Chicago home points reside is 10,500 persons per square mile with a median of 6,248 persons per square mile. The mean occupied housing units in these block groups is 4,706, and the median is 2,288 occupied housing units per square mile. The average size of the Chicago block groups for the unmasked home points is 0.5 square miles.

In the original Atlanta wearable GPS data results, there were 797 total persons with GPS data generating 11,308 total trips. Out of these participants, all trips from 200 randomly selected persons were chosen for inclusion in this masking study. This study thus masks 2,773 total trips in the ARC region. The trip rate per person over seven days of travel is thus higher in the Chicago region than in the greater Atlanta region. The total number of waypoints comprising these Atlanta trips is 603,318. Thus, despite trips represented in the CMAP data set, similar totals of waypoints are examined in both greater metropolitan regions. A reason for this is that the average trip taken in the ARC region is longer than the average trip taken in the CMAP region. Lower overall population density in the Atlanta region is likely linked to necessity to travel farther for work and amenities. The mean population density of home points of Atlanta GPS participants is 2,517 persons per square mile, much lower than the mean population density of selected homes in the CMAP region. The mean density of occupied housing units per square mile is also much lower than in the Chicago region at 995 units per square mile. The mean size of the Atlanta block groups is 1.5 square miles, three times the average size of the Chicago home block groups. The differences in population and housing density between the two regions are expected to impact the masking results for privacy preservation.

Table 1. Summary of GPS data points in the study regions

Region	Persons	Trips	Total points	Mean population density of home points	Mean occupied housing unit density of home points
CMAP	209	5,671	856,465	10,500 pers/sq mile	4,706 units/sq mile
ARC*	200	2,773	603,218	2,517 pers/sq mile	995 units/sq mile

* Randomly sampled within wearable GPS participant results

k-Anonymity

This section reviews the success of the masking techniques in both regions for preserving privacy under the metric of k-anonymity, which is affectionately nicknamed "safety in numbers." The preservation of home anonymity is reviewed first, followed by the results for route anonymity through collocation with other trajectories in the data set. Home k-anonymity is calculated by the comparative occupied housing density of the block groups containing the home locations of the GPS data before and after obfuscation. Route k-anonymity is determined using a density raster of all GPS waypoints in the data sets and calculating collocation with other waypoints over the entire length of each trip.

Point Anonymity

The point anonymity of the home locations was measured using the occupied housing density of the block groups where the points and masked points reside. Overall, fewer than half of home points were moved from one block group to another for each of the masking thresholds for both Chicago and Atlanta. Of the 209 homes in the CMAP region, 96 home locations (45.9%) remained in the same block group as the original home location for all masking techniques and thresholds. Of the 200 homes in the ARC region, 139 home locations (69.5%) remained in the same block group as the unmasked home location. The higher percentage of homes remaining in the same block group for the Atlanta data is partially explained by the larger average size of the Atlanta region block groups, which is 1.5 square miles, compared to 0.5 square miles in Chicago.

Table 2 depicts the mean occupied housing density for the masked home locations in the Chicago region. On average, the occupied housing density of the block groups for all masking methods and thresholds is higher than that for the original points. Perturbation as an obfuscation method resulted in home points being moved to block groups of higher k-anonymity than for grid masking. Table 3 illustrates the percentage of home locations relocated to new block groups due to obfuscation in the Chicago area. Random perturbation resulted in a higher percentage of homes being relocated to different block groups than grid masking. This helps to explain the higher average occupied housing density reached for all of the random perturbation thresholds in the Chicago region.

Table 2. Mean occupied housing density for masked home locations, CMAP

Region	Technique	30 meters	100 meters	250 meters
CMAP	*Grid*	4,951.28	4,747.53	4,516.88
	Perturbation	5,013.85	5,006.46	5,166.11
Unmasked	*4,705.83*			

Table 3. Percent of home locations relocated to different block group, CMAP

Region	Technique	30 meters	100 meters	250 meters
CMAP	*Grid*	4.78%	14.35%	26.79%
	Perturbation	8.13%	14.83%	36.84%

Table 4 summarizes the results for mean occupied housing density under obfuscated conditions in the Atlanta region. The unmasked mean occupied housing density is 994.9 units per square mile, which is a much lower density than for the Chicago region, where it is 4,705.9 units per square mile. Just as in the Chicago region, as the distance threshold for perturbation increases, the k-anonymity of the home points increases. The highest k-anonymity is at the 250-meter perturbation threshold with an average occupied housing unit density of 1,067.9 units per square mile. The only iteration for the Atlanta region where the average k-anonymity decreased was with the 30-meter random perturbation threshold.

Similar to Chicago, as the distance thresholds increased for both grid masking and random perturbation in the Atlanta region, the percentage of home locations relocated to different block groups increased between 5.5% (grid masking, 30 meters) and 19.5% (random perturbation , 250 meters). These results are shown in Table 5. The lower percentages of block group reassignments compared to the Chicago results is explained by the higher average size of the block groups in the Atlanta region. Applying identical distance thresholds in masking points results in fewer block group reclassifications for each point in the larger block groups compared to those originally in the smaller block groups.

Table 4. Mean occupied housing density for masked home locations, ARC

Region	Technique	30 meters	100 meters	250 meters
ARC	*Grid*	1,004.52	1,016.26	1,014.75
	Perturbation	989.84	1,009.45	1,067.85
Unmasked	*994.85*			

Table 5. Percent of home locations relocated to different block group, ARC

Region	Technique	30 meters	100 meters	250 meters
ARC	*Grid*	5.50%	11.00%	17.00%
	Perturbation	7.00%	12.00%	19.50%

Overall, k-anonymity for the home points may be better assessed by measuring the number of other homes within a radius of each home location. Block groups may be too large of administrative areas to measure occupied housing density for these purposes. The average size of the Chicago block groups is 0.5 square miles and, the average size of the Atlanta block groups is 1.5 square miles. Even with these small areas, housing units are not likely to be distributed equally throughout. In these data sets, at least half of all home locations did not switch block groups under any of the masking conditions. This suggests that changes in the underlying settlement density were not fully portrayed in the block group characterization. Therefore, the weighting scheme used by Kwan et al. (2004) and Allshouse et al. (2010) for masking based on block group population density does not fully capture the underlying settlement patterns.

Route Anonymity

Route anonymity was determined using the concept of k-anonymity, meaning that each route in the GPS data set is indistinguishable from at least k other routes. The uniqueness of the route was determined by the proportion of the trip traveled without collocation with another point from another route. The collocation index was developed by creating a kernel density estimation with a cell size of 150 meters by 150 meters. This is in adherence to the cell size selection formula posited by Meratnia and de By (2002) for a maximum average velocity of 65 miles per hour. To ensure that only points within the grid cells themselves were included in the calculation, a search radius of 150 meters was also applied. There are more precise methods of computing collocation described in the methods section, but this more efficient method was selected based on the volume of data being analyzed and the fact that the application of this standardized measure to all masked data is what is generating meaningful statistics. These raster values from the kernel density estimation were then extracted to the masked points to calculate the mean collocation of each trip by the duration in seconds.

The mean collocation index value generated for the original CMAP data set is 115.09. With the same index applied to all obfuscation methods with the same cell sizes, the highest collocation value achieved was with the 250-meter masking threshold at 133.39. This is likely due to more points being snapped to the same grid centroid as the size of the cells increased. It was expected that the mean collocation index would also increase between the 30-meter and 100-meter threshold for grid masking, because the larger distance would snap more neighboring points to the centroid of the same grid cell. However, the resulting decrease in k-anonymity between these two grid masking distance thresholds is likely due to there still being small enough grid cells to expose variation. The 100-meter grid mask would still relocate neighboring points to adjacent grid cells rather than the same grid cell. The reasons for the decrease in collocation values between the 30-meter and 100-meter grid mask distances are not fully understood and should be explored further. The decrease in k-anonymity at this grid masking threshold is consistent across both study areas.

For random perturbation in the CMAP study area, the results demonstrate a decrease in collocation and therefore in k-anonymity as the size of the distance threshold increases. This is because perturbation introduces more variation between the points as they are moved in random distances and directions. The increase in variation captured at the 50-meter by 50-meter cell size used in kernel density estimation caused the perturbed points not to appear collocated with other routes. A larger kernel density cell size could demonstrate greater rates of k-anonymity in the perturbed points.

Table 6. Mean k-anonymity index by trip, CMAP

Region	Technique	30 meters	100 meters	250 meters
CMAP	*Grid*	105.42	88.65	133.39
	Perturbation	110.89	88.89	56.83
Unmasked	*115.09*			

Table 7 illustrates the mean k-anonymity index by trip in the greater Atlanta region. Overall, the collocation or k-anonymity values are lower than for the Chicago region. While there are more waypoints considered in Chicago, the lower values in Atlanta still suggest that routes are more unique within the Atlanta data set. A similar pattern to the Chicago region is encountered between the obfuscation techniques and distance thresholds. As the distance threshold of masking increases for both methods, the collocation value decreases, with the exception of the 250-meter grid masking results. The 250-meter grid masking collocation value is 71.49, higher than any of the other masked results and the original results for k-anonymity. As the size of the grid cells increased for grid masking, more point neighbors are snapped to the same grid cell, giving their associated trips a higher collocation value. For both study regions, the 250-meter grid masking threshold is most preferable for maintaining route anonymity at the resolution of 150 meters used for creating the route density surface. Random perturbation at 250-meter distance threshold is least preferable for route anonymity with this measurement. Again, if a larger cell size is used to estimate collocation, the smaller distance thresholds may show better results for protecting privacy with route k-anonymity. A larger cell size would group more GPS waypoints together for collocation.

Table 7. Mean k-anonymity index by trip, ARC

Region	Technique	30 meters	100 meters	250 meters
ARC	*Grid*	56.67	48.45	71.49
	Perturbation	56.52	45.56	28.01
Unmasked	*58.52*			

The other two hypotheses in this study are that areas of lower road density will be correlated with GPS trips that exhibit a higher degree of k-anonymity and that the trajectories of different modes of travel will exhibit varying levels of route k-anonymity. Lower road density lends itself to greater route anonymity because of the limited choice of routes one is able to take and the greater probability of collocation with other routes. Generally, as the average speed of a trip increases, the anonymity of the route is also expected to rise. This is because walking trips of lower speeds are expected to be closer to home locations and thus more unique within a data set compared to higher-velocity trips along highways. Highway trips are expected to exhibit higher rates of collocation with other trips because limited access roads are often included in optimal routes.

In this study, median trip speed is treated as a proxy for trip mode and is tested for its prediction of collocation along with trip duration and road density. An ordinary least squares linear regression was run in both study regions to test the effect of median trip speed, trip duration, and road density on collocation in original unmasked GPS data sets. Table 8 demonstrates that overall the model is significant with a p-value of 0.000 and an adjusted R^2 of 0.216. Median trip speed is a significant predictor of collocation for the CMAP data with a p-value of 0.000. In contrast to the road density hypothesis for this study, the variable has a negative correlation with the collocation index as shown in Table 8. This result is likely mediated by the fact that waypoints from the same person falling in the same cell were added together in the density calculation. Points from the same person at low speeds are captured in the same grid cell and thus appear to have high collocation with other routes and high k-anonymity. An alternative explanation is that walking trips may fall in inner city business districts, where there is a higher probability of sharing a route with others due to higher daytime population density. Further analysis with a more precise collocation measure is needed to differentiate between these two patterns. Road density is also a significant predictor of route k-anonymity with a positive coefficient of 0.013. This result also runs contrary to the hypothesis that lower road density would be associated with higher k-anonymity. While the coefficient in this regression model is very low, a possible explanation is that low road density is correlated with low population density in remote areas. This would result in lower representation in the GPS data sets, and thus with greater route uniqueness. Trip duration did not reach significance in this model for the prediction of the route collocation index value for k-anonymity. An explanation for this is short trips occurred in both remote regions around homes as well as in central business districts, and there was thus a weak relationship with collocation with other routes. Long trips may have been particularly unique as travel far away is less common in daily travel patterns, which the GPS surveys intended to measure. On the other hand, longer trip durations may be equally associated with highway travel and thus collocated with other routes. These ideas should be explored further in addressing why trip duration is insignificant for these k-anonymity values.

Table 8. Results of a linear regression predicting route collocation index value, CMAP

Independent variables	Beta	p-value	
Median trip speed	-4.435	0.000	
Trip duration	0.002	0.264	
Road density	0.013	0.000	
Model statistics	Adjusted R^2	F-statistic	p-value
	0.216	521.768	0.000

The same linear regression was applied in the greater Atlanta region with similar results. Table 9 illustrates that the model is significant with a p-value of 0.000 and an adjusted R^2 of 0.193. All three predictor variables are significant in this regression with the same characteristics as for the Chicago region. Median trip speed is negatively correlated with route k-anonymity with a coefficient of -1.605. This supports the idea that either trips with low speeds are being erroneously collocated with points from the same trip or that trips with lower speeds are taking place along routes with greater foot-traffic and vehicle-traffic. Trip duration has a very low coefficient in the Atlanta model, but is positively correlated with route k-anonymity. While this weak correlation should not be ignored in giving this result undue weight, longer trips may be associated with highways and thus have more shared routes with other persons in the data set. Finally, road density is also a weak predictor positively correlated with route k-anonymity in the Atlanta data set. Just as for Chicago, this could be explained by lower road density exhibited in remote regions where there are fewer trips and greater route uniqueness. These tests should ideally be run on GPS point density rasters with only one point per person per grid cell to determine underlying relationships and remove confounding variables.

Table 9. Results of a linear regression predicting route collocation index value, ARC

Independent variables	Beta	p-value	
Median trip speed	-1.605	0.000	
Trip duration	0.003	0.002	
Road density	0.005	0.030	
Model statistics	**Adjusted R^2**	**F-statistic**	**p-value**
	0.193	222.478	0.000

Spatial Pattern Preservation

Spatial pattern in this study is measured by the Pearson correlation coefficient between the kernel density estimation (KDE) of the original GPS points and the KDE of the masked GPS points. The parameters for the kernel density estimation were a search radius of 100 meters and a cell size of 50 meters by 50 meters. These parameters were selected purposefully to be more restrictive than in other studies. Shi et al. (2009) test kernel density spatial pattern preservation with cell sizes the same size or larger than the distance threshold applied in random perturbation. The authors conclude that a cell size of five times the distance threshold is needed to maintain high correlation between the original and masked spatial patterns. This study instead aims to test correlation when the KDE cell size is relatively small and close in size to the distance thresholds. A cell size of 50 meters is within this range and commonly used in raster analyses.

The results of the Pearson's correlation analysis for both the Chicago and Atlanta regions are shown in Table 10. As expected, as the distance threshold of masking increases, the correlation coefficient between the original and masked data sets decreases. This result is consistent for both study regions and obfuscation methods. The rate of decrease in the correlation coefficient as the distance threshold increases is faster for grid masking than for random perturbation. The 250-meter threshold for grid masking in CMAP produces a coefficient of 0.399, exhibiting a low correlation with the original data set. A reason for this is that the average distance a point is moved when snapped to the centroid of a grid cell is higher than the average distance a point is moved in random perturbation for this distance threshold. Based on this resolution of analysis with 50-meter grid cells, the 250-meter grid masking results would not be an acceptable replacement for the original data set, and incorrect conclusions would likely be drawn. The 30-meter and 100-meter perturbation thresholds exhibit close correlation to the original data set for CMAP with coefficients of 0.968 and 0.934 respectively. The highest level of correlation with the original data set for Chicago is found with the grid masking threshold of 30 meters at 0.985.

Table 10. Pearson's correlation coefficients between original and masked KDE

Region	Technique	30 meters	100 meters	250 meters
CMAP	*Grid*	0.985	0.789	0.399
	Perturbation	0.968	0.934	0.838
ARC	*Grid*	0.982	0.866	0.442
	Perturbation	0.984	0.944	0.819

The Pearson's results for the Atlanta region are similar to those in the Chicago region. At the 250-meter grid masking threshold, the correlation between the original and masked data sets is lowest with a coefficient of 0.442. The 30-meter distance thresholds for grid masking and random perturbation maintain high degrees of correlation with the original data set with coefficients of 0.982 for grid masking and 0.984 for random perturbation. Grid masking has the higher rate of decline for spatial pattern as the distance threshold increases. If a 50-meter resolution is needed for analysis of masked GPS data, it is recommended that the 30-meter thresholds or the 100-meter threshold for random perturbation be applied.

CHAPTER 6

CONCLUSIONS

A successful obfuscation method for volunteered geographic information (VGI) should be one that is robust, efficient, and easily implementable for any study area. It should balance between preserving identities as much as possible while minimizing disruption of spatial pattern. This is the first successful large-scale examination of GPS obfuscation results between two greater metropolitan regions. The application of the same masking distance thresholds between these regions provides a unique opportunity to compare how well privacy is preserved for home locations and routes between different settlement patterns. These results are especially valuable due to the large sample size.

This study confirms for grid masking and for random perturbation that as the distance threshold for obfuscation increases, the correlation between the original and masked spatial patterns decreases. In both the Chicago and Atlanta study areas, the correlation at the 250-meter threshold for grid masking is markedly weaker than any of the other thresholds tested. This suggests that for both regions, grid masking thresholds of lower than 250 meters are necessary for maintaining spatial patterns if a resolution of 50 meters in cell size is needed for analysis. Each of the other thresholds for both masking types reached a correlation coefficient of at least 0.8 when rounded.

The route k-anonymity results in this study demonstrate that there is a definite trade-off between spatial pattern and privacy for masked GPS data, particularly at the 250-meter grid masking threshold. While these results had the lowest Pearson's coefficients, making them inappropriate for small-scale analysis, they also exhibited the highest degree of route anonymity. The 250-meter grid masking GPS points demonstrated considerably higher average route collocation values than any other distance threshold and obfuscation technique. The lowest route k-anonymity values were recorded at the 250-meter thresholds for random perturbation, indicating that levels of route privacy decreased and route uniqueness appeared to increase at this random perturbation threshold. The 100-meter distance thresholds for both obfuscation techniques exhibited acceptable correlation with the original GPS data, but the route collocation values were relatively low.

A limitation of the collocation component of this study is that it does not eliminate points from the same person falling within the same grid cell. This especially occurs for points recorded during walking trips, as well as points recorded at the origins and destinations of trips, where speeds are lower and the distance between waypoints is higher. Numerous points in the same grid cell can give the appearance of collocation with other trips for route anonymity, but this would be false where origins are unique. This study operates under the assumption that all trips have beginning and end points with high collocation values due to close proximity to themselves, compared to the middle portions of the trips. The focus for collocation of this study is how route uniqueness varies by mode, trip duration, and the distance threshold applied in obfuscation. Future studies that examine more closely the concept of collocation using density rasters should account for and resample for points from the same trip falling more than once in an index grid cell. This study also did not separate the collocation measure by time of day or day of week. These temporal considerations are important when addressing how unique a route truly is at a given day and time.

Future studies should focus on quantifying the tradeoffs between the level of privacy and cell size. A more quantitative measure of privacy protection is needed. It is clear from this study that a simple weighting function, such as the one used by Kwan et al. (2004) for perturbation according to population density, is too simplistic for effectively capturing the nuances of underlying settlement patterns. There is a need for mathematical models and a more sophisticated weighting scheme if settlement density is to be incorporated in obfuscation design. A vector of weights based on measures in landscape architecture and settlement morphology would be a more powerful tool to address underlying densities and balance more effectively between the preservation of privacy and spatial pattern in GPS data.

REFERENCES

AbdelMalik, P., M. N. K. Boulos, and R. Jones. 2008. The perceived impact of location privacy: A web-based survey of public health perspectives and requirements in the UK and Canada. *Bmc Public Health* 8.

Allshouse, W.B., Fitch, M.K., Hampton, K.H., Gesink, D.C., Doherty, I.A., Leone, P.A., Serre, M.L., Miller, W.C. 2010. Geomasking sensitive health data and privacy protection: an evaluation using an E911 database. *Geocarto International* 25(6): 443-452.

Andrienko, G., Andrienko, N., & Wrobel, S. 2007. Visual analytics tools for analysis of movement data. *ACM SIGKDD Explorations* 9(2): 38-46.

Baum, K., Catalano, S., Rand, M., Rose, K. 2009. *Stalking Victimization in the United States.* Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.

Bernheim Brush, A.J., J. Krumm, and J. Scott. 2010. Exploring end user preferences for location obfuscation, location-based services, and the value of location. UbiComp 2010, Copenhagen, Denmark.

Braga, R.B., Tahir, A., Bertolotto, M., and H. Martin. 2012. Clustering user trajectories to find patterns for social interaction applications. *Web and Wireless Geographical Information Systems* April 12-13, 2012. Naples, Italy. 82 -97.

Burns, M. 2014. Problem in Tinder dating app leaked user locations. *Tech Crunch* http://techcrunch.com/2014/02/20/problem-in-tinder-dating-app-leaked-user-locations/

Carr, J. 2013. Workshop notes – ethical and legal implications of geospatial privacy. NSF Workshop: Ethical and Legal Implications of Geospatial Privacy. May 13-15, 2013. Albuquerque, NM.

Census Bureau. 2012. Population Estimates, Metropolitan and Micropolitan Statistical Areas. http://www.census.gov/popest/data/metro/totals/2012/

CITI-SENSE 2014. Development of sensor-based Citizens' Observatory Community for improving quality of life in cities. http://www.citi-sense.eu/Portals/106/Documents/Dissemination%20material/CITI-SENSE_information_update_6.pdf

Crampton, J. 1995. The Ethics of GIS. *Cartography and Geographic Information Systems* 22(1): 84-89.

Crooks, A., Croitoru, A., Stefanidis, A., Radzikowski, J. 2013. #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS* 17(1): 124-147.

Curry, M.R. 1997. The digital individual and the private realm. *Annals of the Association of American Geographers* 87 (4): 681-699.

Curtis, A., Mills, J.W., Agustin, L., and Cockburn, M. 2011. Confidentiality risks in fine scale aggregations of health data. *Computers, Environment and Urban Systems* 35(1): 57-64.

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. http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html

Dobson, J. E., and P. F. Fisher. 2003. Geoslavery. *Ieee Technology and Society Magazine* 22 (1):47-52.

Dobson, J. E., and P. F. Fisher. 2007. The Panopticon's changing geography. *Geographical Review* 97 (3):307-323.

Duckham, M. and L. Kulik. 2007. Location privacy and location-aware computing. In *Dynamic & mobile GIS: Investigating change in space and time*, ed. J. Drummond, R. Billen, E. Joao, and D. Forrest. Boca Raton, FL: CRC Press 34-51.

Elwood, S., and A. Leszczynski. 2011. Privacy, reconsidered: New representations, data practices, and the geoweb. *Geoforum* 42 (1):6-15.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 226-231.

Friedland, G. and Sommer, R. 2010. Cybercasing the joint: On the privacy implications of geo-tagging. *5th Usenix Hot Topics in Security Workshop (HotSec2010)*, Washington DC.

Gaudin, S. 2009. Google Latitude lets you track friends, employees. *Computerworld*. http://www.computerworld.com/s/article/9127372/Google_Latitude_lets_you_track_friends_employees

Gambs, S., Killijian, M-O., Cortez, M.N.d.P. 2010. GEPETO: A GEoPrivacy-Enhancing TOolkit. *IEEE Computer Society AINA Workshops* 1071-1076.

Goodchild, M. F. 2011. Challenges in geographical information science. *Proceedings of the Royal Society a-Mathematical Physical and Engineering Sciences* 467 (2133): 2431-2443.

Goodchild, M.F. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4): 211-221.

Goodchild, M. F. and Hunter, G. J. 1997. A simple positional accuracy measure for linear features. *International Journal of Geographic Information Science* 11(3): 299-306.

Goss, J. 1995. We know who you are and we know where you live - the instrumental rationality of geodemographic systems. *Economic Geography* 71 (2):171-188.

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.

Harvey, D. 1990. Between space and time: reflections on the geographical imagination. *Annals of the Association of American Geographers* 80(3): 418-434.

Herzog, A. 2010. Why I deleted my Foursquare account. *Exploring Digital Media*. http://ariherzog.com/blog/why-i-deleted-my-foursquare-account/

Jacobson, S. 2012. Elderly couple abandon their home after address is posted on Twitter as that of George Zimmerman. *Orlando Sentinel* http://articles.orlandosentinel.com/2012-03-29/news/os-trayvon-martin-wrong-zimmerman-20120327_1_spike-lee-william-zimmerman-retweeted

Kar, B., R.C. Crowsey, and J.J. Zale. 2013. The Myth of location privacy in the United States: surveyed attitude versus current practices. *The Professional Geographer* 65(1): 47-64.

Kessler, F. 2011. Volunteered geographic information: a bicycling enthusiast perspective. *Cartography and Geographic Information Science* 38(3): 258-268.

Krumm, J. 2009. A survey of computational location privacy. *Personal and Ubiquitous Computing* 13 (6):391-399.

Krumm, J. 2007. Inference attacks on location tracks. *5th International Conference, Proceedings, PERVASIVE 2007* Toronto, Canada, May 13-16, 2007. 127-143.

Kwan, M. P.; Casas, I.; Schmitz, B. C. 2004. Protection of geoprivacy and accuracy of spatial information: how effective are geographical masks. *Cartographica* 39: 15-28.

Leitner, M., and A. Curtis. 2006. A first step towards a framework for presenting the location of confidential point data on maps - results of an empirical perceptual study. *International Journal of Geographical Information Science* 20 (7):813-822.

Li, L. and Goodchild, M.F. 2013. Is privacy still an issue in the era of big data? – Location disclosure in spatial footprints. *Geoinformatics* 1-4.

Lin, B. and Su, J. 2008. One way distance: for shape based similarity search of moving object trajectories. *Geoinformatica* 12(2): 117-142.

Malheiros, M. 2009. Medical data secondary use issues. *Privacy Value Networks* http://www.pvnets.org/2009/06/medical-data-secondary-use-issues/

Meratnia, N. and de By, R.A. 2002. Aggregation and comparison of trajectories. *Proceedings of the ACM Symposium on Advances in Geographic Information Systems* 49–54.

Monmonier, M. 2003. The Internet, cartographic surveillance, and locational privacy. In Peterson, M. (ed.) *Maps and the Internet*. Amsterdam, Cambridge: Elsevier Press.

National Renewable Energy Laboratory. 2013. Real-World Data for Planning, Modeling and Analysis Fact Sheet. http://www.nrel.gov/docs/fy13osti/57559.pdf

National Research Council. 2007. Putting people on the map: protecting confidentiality with linked social-spatial data. *Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data*. M.P. Gutmann and P.C. Stern, Eds. Committee on the Human Dimensions of Global Change. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nergiz, M.E., Atzori, M., Saygin, Y. and Güç, B. 2009. Towards trajectory anonymization: a generalization-based approach. *Transactions on Data Privacy* 2: 47-75.

Nustats. 2008. Chicago Regional Household Travel Inventory Data User's Manual. *Chicago Metropolitan Agency for Planning* 1-63. http://www.cmap.illinois.gov/c/document_library/get_file?uuid=c4e21cbc-a18d-41a5-a7c1-3b5dbcf9cf66&groupId=14401

Nouwt, S. 2008. Reasonable expectations of geo-privacy? *SCRIPTed* 375 5(2) http://www.law.ed.ac.uk/ahrc/script-ed/vol5-2/nouwt.pdf

Onsrud, H. J., J. P. Johnson, and X. R. Lopez. 1994. Protecting personal privacy in using geographic information-systems. *Photogrammetric Engineering and Remote Sensing* 60 (9):1083-1095.

Pickles, J. ed. 1995. *Ground Truth: The social implications of geographic information systems*. New York: Guilford.

Pingley, A., W. Yu, N. Zhang, X. Fu, and W. Zhao. 2009. CAP – A context-aware privacy protection system for location-based services. *Distributed Computing Systems Conference*, June 22-26, 2009, 49-57.

PTV Nustats. 2011. Atlanta Regional Commission Regional Travel Survey Final Report. *Atlanta Regional Commission* http://www.atlantaregional.com/File%20Library/Transportation/Travel%20Demand%20Model/tp_2011regionaltravelsurvey_030712.pdf 1-207.

Richardson, D. and van Oosterom P., ed. 2002. *Advances in Spatial Data Handling*. Springer: New York.

Riordan Seville, L. 2010. Foursquare's stalker problem. *The Daily Beast* http://www.thedailybeast.com/articles/2010/08/08/foursquare-and-stalking-is-geotagging-dangerous.html

Robison, P. 2014. Tracking every move you make—for a $5 gift card. *Bloomberg Businessweek* http://www.businessweek.com/articles/2014-02-14/tracking-every-move-you-make-by-gps-for-a-5-gift-card-without-that-nsa-taint?

Shi, X., Alford-Teaster, J., and Onega, T. 2009. Kernel density estimation with geographically masked points. *Proceedings of the 17th International Conference on Geoinformatics* August 2009.

Shilton, K. 2012. Participatory personal data: An emerging research challenge for the information sciences. *J. Am. Soc. Inf. Sci. Technol.* 63 (10):1905-1915.

Solove, D.J. 2007. 'I've got nothing to hide' and other misunderstandings of privacy. *San Diego Law Review 44*: 745-772.

Sui, D. 2007. Paradoxical consequences of location-based services (LBS): A tetradic analysis using McLuhan's laws of media. In *Societies and Cities in the Age of Instant Access* (ed. Miller, H.J.) Springer: Netherlands. 255-271.

Sui, D. 2004. The media and messages of location-based services (LBS): death of distance or the revenge of geography? *Journal of Geographical Information Sciences* 10(2): 166-174.

Sweeney, L. 2002. k-Anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5): 557-570.

Vicente, C.R., Freni, D., Bettini, C. and C.S. Jensen. 2011. Location-related privacy in geo-social networks. *Internet Computing* IEEE 20-27.

Vlachos, M., Gunopoulos, D. and G. Kollios. 2002. Discovering similar multidimensional trajectories. *Proceedings of the 18th International Conference on Very Large Databases*. Aug 25-29, 1997. 186-195.

Warren, S.D. and Brandeis, L. 1890. The Right to privacy. *Harvard Law Review* 4(5): 193-220.

Wieland, S.C., Cassa, C.A., Mandl, K.D., and Berger, B. 2008. Revealing the spatial distribution of a disease while preserving privacy. *Proceedings of the National Academy of Science of the United States* 105(46): 17608-17613.

Wolf, J., Guensler, R., and Bachman, W. 2001. Elimination of the travel diary: an experiment to derive trip purpose from GPS travel data. *Transportation Research Board 80th Annual Meeting*, Jan 7-11, 2001. Washington, D.C.

Zhong, G., Goldberg, I., and U. Hengartner. 2007.Louis, Lester and Pierre: Three protocols for location privacy. *Privacy Enhancing Technologies* 62-76.

APPENDIX

IRB APPROVAL DOCUMENT

SAN DIEGO STATE UNIVERSITY

Graduate and Research Affairs
Division of Research Affairs
San Diego State University
5500 Campanile Drive
San Diego CA 92182-8220
Phone: 619-594-5938
Fax: 619-594-4109

Emailed Questionnaire

December 23, 2013

SDSU Student Researcher: Dara Seidl
SDSU Faculty Researcher: Dr. Piotr Jankowski
Department: Geography
Title: *Route Obfuscation for Privacy Preservation in Volunteered Trajectory Data*

Dear Dara Seidl:

Your responses on the screening questionnaire have been received and evaluated for your project. The screening questionnaire is a tool that allows the IRB to determine whether your project requires IRB review. Your responses indicate that you are conducting research that involves human subjects. The IRB is required to review activities that involve the participation of human subjects in research as defined in the code of federal regulations (45 CFR 46).

According to the Code of Federal Regulations, research is defined as, "A systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge." As described in the Belmont Report, "...the term 'research' designates an activity designed to test a hypothesis, permit conclusions to be drawn, and thereby to develop or contribute to generalizable knowledge. A human subject is defined as "a living individual about whom an investigator (whether professional or student) conducting research obtains (1) data through intervention or interaction, or (2) identifiable private information."

In applying these definitions to the questionnaire responses submitted *and your emailed response to the SDSU IRB on 12/19/13* it appears that **your project does not involve research and human subjects.** Given that the information you are receiving is not personally identifiable or specific to any individual.

The project must receive IRB approval in advance of initiation. You may submit your IRB application materials through the vIRB which is accessible through the SDSU Web Portal available at http://www.sdsu.edu/webportal. If you have any questions regarding the use of human subjects in research or the appropriate review type for your research, please contact me at (619) 594-6622 or by E-mail at irb@mail.sdsu.edu.

Sincerely,

Choya Washington
Regulatory Compliance Analyst