PRIVACY RISK EVALUATION OF HUMAN MOBILITY DATA FOR URBAN TRANSPORTATION PLANNING

FINAL PROJECT REPORT

by

Jan Whittington, Feiyang Sun
University of Washington

Sponsorship
PacTrans and WSDOT

for

Pacific Northwest Transportation Consortium (PacTrans)
USDOT University Transportation Center for Federal Region 10
University of Washington
More Hall 112, Box 352700
Seattle, WA 98195-2700

In cooperation with U.S. Department of Transportation,
Office of the Assistant Secretary for Research and Technology (OST-R)

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The Pacific Northwest Transportation Consortium, the U.S. Government and matching sponsor assume no liability for the contents or use thereof.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
	01786452	

4. Title and Subtitle	5. Report Date
Privacy Risk Evaluation of Human Mobility Data For Urban Transportation Planning	July 2022
	6. Performing Organization Code

7. Author(s) and Affiliations	8. Performing Organization Report No.
Jan Whittington, University of Washington; 0000-0001-6444-2447 Feiyang Sun, University of Washington	2020-S-UW-3

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
PacTrans Pacific Northwest Transportation Consortium University Transportation Center for Federal Region 10 University of Washington More Hall 112 Seattle, WA 98195-2700	11. Contract or Grant No. 69A355174110

12. Sponsoring Organization Name and Address	13. Type of Report and Period Covered
United States Department of Transportation Research and Innovative Technology Administration 1200 New Jersey Avenue, SE Washington, DC 20590	Final Report 2019 - 2021
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This project examined the variations in re-identification risks of mobility trace data in different urban areas, characterized by residential population density, percentage of residential land use, and per capita income, and in different population segments, characterized by race, gender, and household income. The project used the 2017 Puget Sound Regional Travel Survey and estimated the uniqueness of the trip origins and destinations by using the method of *k*-anonymity. This project found that 42 percent of travelers could be re-identified by one trip origin or destination point aggregated at the census block group level and the one-hour time interval. This confirmed previous findings that mobility traces are highly unique. This project further estimated the associations between the built environment and sociodemographic variables and the *k*-values that measure the uniqueness of mobility traces in a data set. The results showed that trips to or from census block groups with a lower per capita income, higher residential population density, or higher percentage of residential land use were more likely to have a higher level of re-identifiability. Similarly, travelers whose mobility traces were more unique than others tended to have higher percentages of male, non-white, and lower income populations. The findings help to optimize the algorithmic solution to minimize the privacy risks and detect and mitigate algorithmic biases in current data practices.

17. Key Words	18. Distribution Statement
Data Privacy, Location Data, Built Environment, Social equity, Data sharing	

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified.	Unclassified.	34	N/A

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS				
Symbol	**When You Know**	**Multiply By**	**To Find**	**Symbol**
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
	NOTE: volumes greater than 1000 L shall be shown in m^3			
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
°F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	°C
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	$candela/m^2$	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS				
Symbol	**When You Know**	**Multiply By**	**To Find**	**Symbol**
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
°C	Celsius	1.8C+32	Fahrenheit	°F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	$candela/m^2$	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

*SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.
(Revised March 2003)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

K:	K-anonymity
BE:	Built environment
SE:	Sociodemographic
CBG:	Census block group
GPS:	Geographic Positioning System
NAPOC:	Non-Asian People of Color
PSRC:	Puget Sound Region Council
PacTrans:	Pacific Northwest Transportation Consortium
WSDOT:	Washington State Department of Transportation

EXECUTIVE SUMMARY

The emergence of new mobility options and location-based services has led to an unprecedented surge in the availability of mobility data sources. While the new data sources offer opportunities to discover new subject-level knowledge and open new fields of inquiry, they also raise privacy concerns, such as revealing intimate information about a person or allowing the re-identification of individuals in a database (Thompson and Warzel, 2019). A common practice that organizations use to mitigate such re-identification risks is to remove all explicit identifiers, such as name, address, and telephone number. Yet studies have shown that such practices can be insufficient because of the uniqueness of the combination of attributes from each individual in a dataset, especially for spatial trace data (De Montjoye et al., 2013; Gao et al., 2019).

Up to now, a number of studies have developed frameworks and algorithms intended to mitigate the associated re-identification risks while considering the tradeoffs between privacy protection and data quality (Sweeney, 2002; Dwork and Roth, 2014; Pellungrini et al., 2017). However, these studies have often lacked a general understanding of the heterogeneous nature of re-identification risks, as risks are associated with differences in urban areas and population segments. Three problems could emerge from such a knowledge gap. First, when the tradeoff between privacy protection and data quality is balanced, failing to consider heterogeneity within the dataset will leave some user groups more vulnerable to privacy attack while unnecessarily reducing data quality for others. Second, each time a dataset is collected from a new area or user group, privacy risks need to be reassessed. The computational complexity of privacy assessment algorithms required to exhaust all possibility of re-identification becomes a severe limitation for its practical application by municipal-level public agencies (Pellungrini et al., 2017). Third, the privacy implication of re-identification is contextually dependent (Nissenbaum, 2011), as being re-identified in a public space is different from being re-identified from a more private setting. Not understanding these differences in urban areas can impede the ability to transform the existing technical findings to support relevant policies and legal regulations.

This study examined the variations in re-identification risks of mobility trace data in different urban areas, characterized by residential population density, percentage of residential land use, and per capita income, and different population segments, characterized by race, gender, and household income. The project used the 2017 Puget Sound Regional Travel Survey

and estimated the uniqueness of the trip origins and destinations by using the method of *k*-anonymity. This report found that 42 percent of the travelers could be re-identified by one trip origin or destination point aggregated at the census block group level and the one-hour time interval. This confirmed previous findings that mobility traces are highly unique. This project further estimated the associations between the built environment and sociodemographic variables and the *k*-values that measure the uniqueness of mobility traces in a dataset. The results showed that trips to or from census block groups with a lower per capita income, higher residential population density, or higher percentage of residential land use were more likely to have a higher level of re-identifiability. Similarly, travelers whose mobility traces were more unique than others tended to have higher percentages of male, non-white, and lower income populations.

The results have two implications for data collection, processing, and publication practices. First, generalization or aggregation is a common strategy used by public agencies to de-identify individuals in a dataset. However, it is often tricky to determine the optimal level and the right fields for aggregation, as over-aggregation may lead to excessive information loss, and under-aggregation may be insufficient to achieve a desired level of *k*-anonymity. By showing the structural variations in re-identification risk, this project suggests the possibility of de-identifying data according to variation in urban areas and population segments, which could help to reduce information loss and offer greater potential protection of the privacy of individuals whose data records are more unique than others.

Second, the methodologies used in this study can help detect and mitigate algorithmic biases in current data practices. This project found that travelers whose spatiotemporal traces were unique from others consisted of higher percentages of non-white persons and lower income populations, which may reflect the symptom of existing geographical inequality. Because these population groups also tend to be underrepresented in the dataset, it makes them more vulnerable to the re-identification risk than others. When oversampling strategies are designed to account for these underrepresented population groups, it is important to consider the distribution of both residential and travel locations.

The study also pointed to two future directions to further expand this area of study. For future studies, public agencies will benefit from analyzing surveys of multiple years and larger geographical extent to confirm the findings. It would also be helpful to test the suggested

aggregation schema to evaluate the effectiveness of such schema in terms of the tradeoffs between information loss and re-identification risk reduction.

CHAPTER 1. INTRODUCTION

Human mobility analysis has attracted a growing interest in recent years from different disciplines because of its importance in a wide range of applications, ranging from urban planning and transportation (Chen et al., 2016) to public health (Chaix et al., 2013). These analyses generally rely on large datasets that store detailed information about the spatiotemporal points visited by individuals in an urban area, such as Global Positioning System (GPS) traces. The emergence of new mobility options and location-based services has led to an unprecedented surge in the availability of mobility data sources. While the new data sources offer opportunities to discover new subject-level knowledge and open new fields of inquiry, they also raise privacy concerns, such as revealing intimate information about a person or allowing the re-identification of individuals in a database (Thompson and Warzel, 2019).

It is a common practice for organizations to anonymize datasets by removing all explicit identifiers, such as name, address, and telephone number. Although the resulting data may look anonymous, in fact, the remaining data can be used to re-identify individuals (Sweeney, 2002). This is because the combination of attributes in a dataset for each observation, or tuple in short, can be unique and used as quasi-identifiers (Dalenius, 1986). These unique quasi-identifiers can be used to identify individuals in the released data or to link or match the data to other datasets. Previous studies have examined the privacy risks of GPS trajectories, mobile phone data, and other human mobility datasets (De Mulder et al., 2008; Kondor et al., 2018). For example, De Montjoye et al. (2013) showed, in their study of the hourly cell-phone tower tracking of 1.5 million devices by MAC address over 15 months, that only four spatial-temporal data points per day were needed to re-identify 95 percent of the owners of those devices. Similarly, Gao et al. (2019) measured the risk of license plate recognition data and found that five spatiotemporal records were enough to uniquely identify about 90 percent of individuals, even when the temporal granularity was set to be half of a day.

Up to now, a number of studies have developed frameworks and algorithms to mitigate the associated re-identification risks while considering the tradeoffs between privacy protection and data quality (Sweeney, 2002; Dwork and Roth, 2014; Pellungrini et al., 2017). However, those studies have often lacked a general understanding of the heterogeneous nature of re-identification risks, as risks are associated with differences in both urban areas and population segments. Three problems could emerge from such a knowledge gap. First, when the tradeoff

between privacy protection and data quality is balanced, failing to consider heterogeneity within the dataset will leave some user groups more vulnerable to privacy attack while unnecessarily reducing data quality for others. The limitation becomes more obvious in metropolitan areas such as Seattle, with a diversity of urban forms and population groups. Second, each time a dataset is collected from a new area or user group, privacy risks need to be reassessed. The computational complexity of privacy assessment algorithms required to exhaust all possibilities of reidentification becomes a severe limitation for its practical application by municipal-level public agencies (Pellungrini et al., 2017). Third, the privacy implication of re-identification is contextually dependent (Nissenbaum, 2011), as being re-identified in a public space is different from being re-identified in a more private setting. Not understanding these differences in urban areas can impede the ability to transform the existing technical findings to support relevant policies and legal regulations.

CHAPTER 2. RESEARCH OBJECTIVE AND HYPOTHESIS

This study furthered the findings and methods of previous studies to examine variations in re-identification risk associated with differing built environment (BE) and socioeconomic (SE) factors. This report tested the following hypothesis**:**

Due to differences in travel patterns, re-identification is heterogeneous across different urban areas measured by population density, per capita income, percentage of residential land use, and sociodemographic segments grouped by race, income, and gender of the travelers.

CHAPTER 3. DATA

This study used the 2017 Puget Sound Regional Travel Survey (PSRC household survey), which collected detailed trip information and sociodemographic characteristics from 6,254 individuals in 3,285 households in the central Puget Sound (Seattle, Washington) region (Puget Sound Regional Council, 2018). In 2017, the central Puget Sound region had a population of 4,063,700 with a density of 591 people per square mile, which was comparable to other metropolitan areas such as Boston or Houston (American Community Survey 2017). Covering 0.20 percent of households and 0.16 percent of the population in the region, the survey used a geographically proportional sampling plan based on the household distribution in each census block group, and it oversampled population segments that were traditionally difficult to reach, such as transit users and pedestrians (Puget Sound Regional Council, 2018). Data on 2,580 households and 5,019 individuals with 17,468 trips were collected via a one- to four-day travel diary filled in by the participant via telephone or a web-based interface, conducted from Tuesday through Thursday; and data on 705 households and 1,235 individuals with 35,024 trips were collected via a new app-based seven-day travel diary, conducted from Monday through Sunday. The travel diary included for each trip the participant identification (person ID), identification of trip origin and trip destination at the scale of the census block group (trip ID), trip start/end times, and trip duration.

In comparison to other data types used in previous studies that captured spatiotemporal points along a trip, including mobile phone tower data (De Montjoye et al., 2013) and license plate recognition data (Gao et al., 2019), the travel survey data only provided the spatiotemporal points at the origin and destination of a trip. Nevertheless, there were three advantages of using the PSRC travel survey over other data types. First, data sets used in previous studies have often only captured users from a single service or along a particular route. The PSRC survey provided a better representation of the region, with a sampling strategy based on the demographic distribution of the regional population. Second, the PSRC survey included detailed information about travelers that had not been available in previous studies, which allowed the analysis of variations of re-identification risks among different social groups. Third, given the prevalent use of travel survey data in public transportation agencies, using the travel survey data could provide more direct practical guidelines to preserve the privacy of the participants in future travel surveys.

As for built environment (BE) variables, we included residential population density, per capita income, and percentage of residential land use. All three variables are compounded variables that capture characteristics of urban areas and are available national-wide, which increased the scalability and reproducibility of this study. Population density and per capita income were obtained from the 2017 American Community Survey, and the percentage of residential land use was collected from the King County online data portal. The three variables were joined with the spatiotemporal points from the survey by census block group. Table 3.1 shows the descriptive statistics of BE variables at the census block group level.

We selected socioeconomic (SE) variables identified as factors associated with travel patterns and activity space, including gender, race, and household income, (Kwan, 2000; Stead, 2001). The data were obtained directly from the PSRC travel survey. To better interpret results, we further grouped race into three groups: white, Asian, and non-Asian people of color (NAPOC) and household income into two groups: household income less than $100,000 and household income above $100,000. The threshold of $100,000 was selected on the basis of the median income of the region. The SE variables were joined with the spatiotemporal points from the survey by person ID. Table 3.2 shows the descriptive statistics of SE variables at the individual traveler level.

Table 3.1 Descriptive statistics of built environment (BE) variables at the census block group level

	Overall (N=1422)
Residential Population density (per square mile)	
Mean (SD)	7510 (9160)
Median [Min, Max]	5320 [0, 161000]
Per capita income (thousand dollar)	
Mean (SD)	44200 (20100)
Median [Min, Max]	41400 [2890, 205000]
Percent residential land use	
Mean (SD)	0.501 (0.243)
Median [Min, Max]	0.549 [0, 0.959]

Table 3.2 Descriptive statistics of sociodemographic (SE) variables at the individual level

	Overall (N=5019)
Gender: male	
Mean (SD)	0.483 (0.500)
Median [Min, Max]	0 [0, 1.00]
Race: white	
Mean (SD)	0.604 (0.489)
Median [Min, Max]	1.00 [0, 1.00]
Race: asian	
Mean (SD)	0.150 (0.358)
Median [Min, Max]	0 [0, 1.00]
Race: non-asian people of color (napoc)	
Mean (SD)	0.256 (0.437)
Median [Min, Max]	0 [0, 1.00]
HH income_<100k	
Mean (SD)	0.466 (0.499)
Median [Min, Max]	0 [0, 1.00]
HH income_>=100k	
Mean (SD)	0.458 (0.498)
Median [Min, Max]	0 [0, 1.00]

CHAPTER 4. METHODOLOGY

The methodology is described in two steps: 1) estimating the probability of achieving *k*-anonymity by random sampling, and 2) estimating BE and SE variable means for each value of *k* by random sampling. The first step emulated existing methods for examining re-identification risk from mobility traces, while the second extended research to further differentiate risk according to BE and SE variables.

4.1. Estimate of Probability of Achieving *K*-Anonymity by Random Sampling

K-anonymity measures the risk of re-identification of an anonymized dataset. To achieve *k*-anonymity, a set of attributes—in this project the combination of census block group ID and timestamp of one or more spatiotemporal points—needs to be shared by *k* individuals. To estimate the probability of achieving *k*-anonymity and the factors associated with each *k* value, this study applied a random sampling approach. Similar methods were used by Sweeney (2002, p. 2) and De Montjoye et al. (2013) and tested for validity. In this study, we estimated values of *k* under five data aggregation scenarios: census block group (CBG) and a 1-minute interval, CBG and a 10-minute interval, CBG and a 15-minute interval, CBG and a half-hour interval, and CBG and a 1-hour interval. Previous studies found that the degree of re-identification risk increased with the number of spatiotemporal points randomly sampled from each individual each time (De Montjoye et al., 2013; Gao et al., 2019). In this study, we treated the origin and destination of each trip as separate spatiotemporal points and measured *k*-anonymity with one spatiotemporal point from each individual in each round of sampling in order to highlight the variability of *k*-anonymity among different subgroups in the dataset.

As illustrated in figure 4.1, in each round, a sample of 5,028 spatio-temporal points was created by randomly selecting one point per individual. A value of *k* was assigned to each point by counting the total number of points sharing the same combination of block group trip ID and time stamp to that point. A value of one ($k = 1$) indicated that the point was unique in the sample and thus could be used to re-identify the person from the dataset, whereas a value of four ($k = 4$) meant that there were in total four points with the same combination, which reduced the risk of re-identification. The sample of points was divided into groups with the same value of *k*. The percentage of points in each group of the total 5,028 points was calculated as the probability of achieving *k*-anonymity in that round. After 1,000 rounds of sampling, we calculated the average probability of achieving each value of *k*.

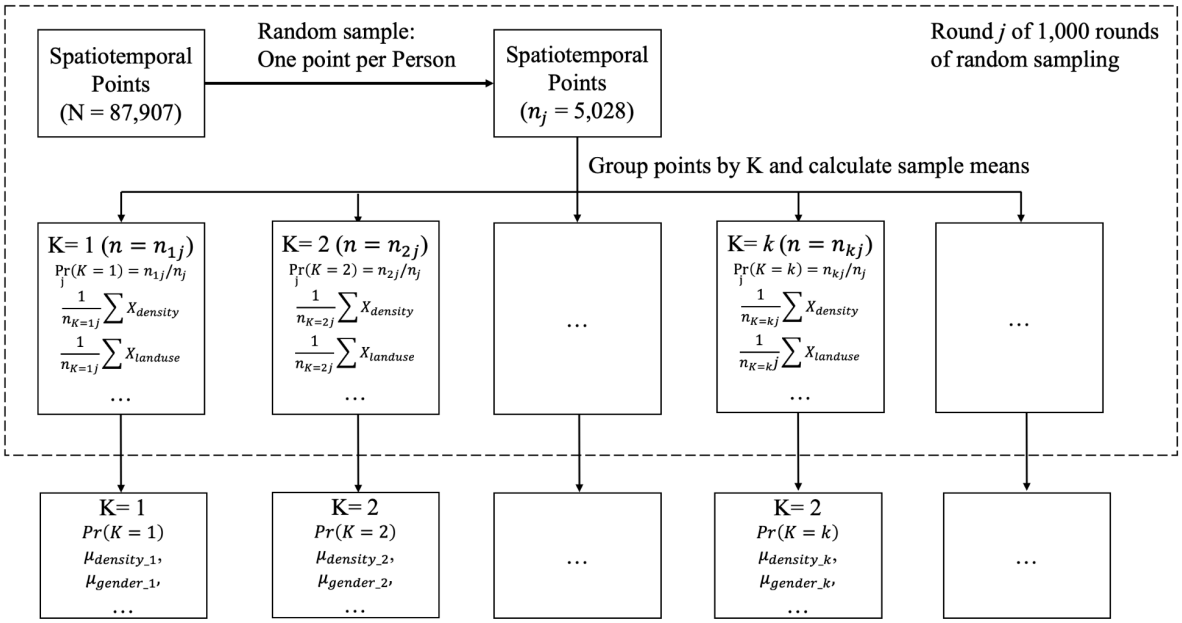

Figure 4.1 Workflow for estimating probabilities of achieving *k*-anonymity and associated BE and SE variable means for each *k* value from random sampling

4.2. Estimate of BE and SE Variable Means for Each Value of *K* by Random Sampling

In comparison to previous studies, this project took a further step by examining the associations between BE and SE variables and *k*-anonymity. As shown in equation (1), the sudy first calculated the sample mean of each variable for each *k* and then estimated the overall mean from the sample means.

$$\mu_{ik} = \frac{1}{N_k}\sum_{j=1}^{N_k}\frac{1}{n_{kj}}\sum_{m=1}^{n_{kj}} x_{im} \cdots \quad (1)$$

In equation (1), *k* is the number of individuals sharing the same spatiotemporal point that may identify the individual in the dataset. $\mu_{ik}$is the unbiased estimate of population mean for the *ith* variable when K=*k*. $N_k$is the total number of sample means from the random sampling for each value of *k*. $\frac{1}{n_{kj}}\sum_{m=1}^{n_{kj}} x_{im}$ calculates the sample means. $n_{kj}$is the number of spatiotemporal points from the *jth* round of random sampling and under the value of *k*. $x_{im}$is the value of the *ith* variable of the *mth* spatiotemporal points from that round of sampling.

To estimate the relationships between BE and SE variables and *k*-anonymity, we estimated a set of Poisson regression models with *k* as dependent variables and BE and SE sample means as predictors, using data from the 1,000 rounds of sampling under the aggregation

scenario of CBG and a 1-hour interval. The predictors were log-transformed to account for overdispersion.

CHAPTER 5. FINDINGS

Our findings relate to the two steps of the methods applied. First we examined the probability that a given individual's spatiotemporal point would be found unique and therefore would represent re-identification risk. Then we examined the extent to which variations in BE and SE were associated with this risk and to which some BE and SE groups might experience more exacerbated risks than others.

5.1. *K*-Anonymity at Varying Time Intervals

As shown in table 5.1, with only one spatiotemporal point, there was a high probability that the point was unique in the dataset. At the CBG and 1-minute time interval, on average 85 percent of the time a point was unique. The probability of $k = 1$ diminished as the aggregation of time intervals increased, as more time allowed the aggregation of travelers within or across CBGs. However, at the 1-hour time interval, on average, 41.7 percent of the time a point was still unique in space and time. This indicated that 41.7 percent of the individuals in the dataset were the only travelers visiting the given CBG at the given hour, which was a mathematical expression of the vulnerability of the individuals to re-identification risk.

Table 5.1 Probability of achieving *k*=1, 2, 3, 4 at five levels of spatiotemporal aggregation

Spatial and Temporal Aggregation	Percentage of Points K = 1 (N=1000)	Percentage of Points K = 2 (N=1000)	Percentage of Points K = 3 (N=1000)	Percentage of Points K = 4 (N=1000)
CBG & 1-minute				
Mean (SD)	0.851 (0.00583)	0.115 (0.00574)	0.0241 (0.00347)	0.00669 (0.00220)
Median [Min, Max]	0.851 [0.833, 0.870]	0.115 [0.0947, 0.133]	0.0239 [0.0131, 0.0358]	0.00636 [0.000796, 0.0143]
CBG & 10-minute				
Mean (SD)	0.725 (0.00699)	0.176 (0.00717)	0.0538 (0.00517)	0.0219 (0.00406)
Median [Min, Max]	0.725 [0.702, 0.746]	0.177 [0.155, 0.201]	0.0537 [0.0382, 0.0698]	0.0223 [0.00955, 0.0342]

Spatial and Temporal Aggregation	Percentage of Points K = 1 (N=1000)	Percentage of Points K = 2 (N=1000)	Percentage of Points K = 3 (N=1000)	Percentage of Points K = 4 (N=1000)
CBG & 15-minute				
Mean (SD)	0.679 (0.00722)	0.193 (0.00729)	0.0649 (0.00564)	0.0289 (0.00438)
Median [Min, Max]	0.679 [0.658, 0.699]	0.193 [0.167, 0.214]	0.0650 [0.0495, 0.0853]	0.0286 [0.0127, 0.0414]
CBG & 30-minute				
Mean (SD)	0.551 (0.00719)	0.217 (0.00778)	0.0923 (0.00653)	0.0459 (0.00547)
Median [Min, Max]	0.551 [0.529, 0.573]	0.217 [0.189, 0.243]	0.0925 [0.0740, 0.115]	0.0453 [0.0278, 0.0636]
CBG & 60-minute				
Mean (SD)	0.417 (0.00683)	0.218 (0.00799)	0.118 (0.00747)	0.0675 (0.00658)
Median [Min, Max]	0.417 [0.396, 0.440]	0.218 [0.196, 0.243]	0.118 [0.0961, 0.144]	0.0668 [0.0493, 0.0883]

5.2. Associations between *K*-Anonymity and BE and SE Variables

Figure 5.1 illustrates the relationships between BE variables and values of *k* achieved at the CBG and a 1-hour time interval. In each plot, the x-axis is the value of *k,* and the y-axis is the overall mean of the variable of interest for each *k* value estimated from the random sampling. The further to the right of the x-axis, the higher the number of data subjects sharing the same attributes, and therefore the lower the likelihood of being re-identified from the dataset. Thus, a positive slope suggests that a higher value of the measured BE variable is associated with a lower risk of being re-identified, while a negative slope indicates that a higher value of the measured BE variable is associated with a higher risk of being re-identified from the crowd.

According to figure 5.1, spatiotemporal points were more likely to achieve a lower *k*-value and were therefore more unique in the dataset in CBGs with higher population densities,

higher percentages of residential land use, and lower per capita income. Similar patterns were also observed with other time intervals.

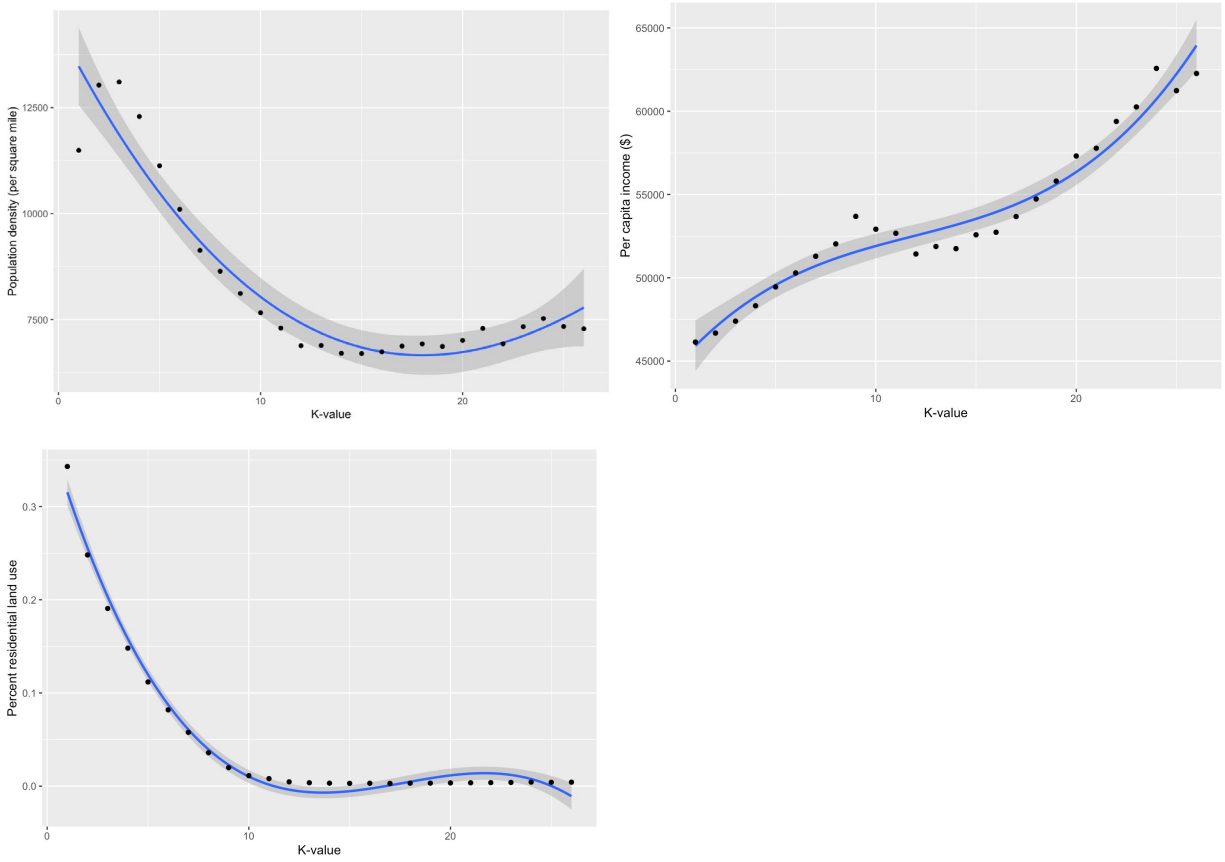

Figure 5.1 Relationship between BE and *k* value achieved at CBG and 1-hour intervals

Figure 5.2 shows the relationships between SE variables and values of *k* achieved at the CBG and a 1-hour time interval. The x-axis is the *k*-value, and the y-axis is the percentage of a demographic group whose points achieved the corresponding *k*-value. The further to the right end of the x-axis, the higher the number of data subjects sharing the same attributes, and therefore the lower the likelihood of being re-identified from the dataset. Individuals whose points had a low *k* value and were more unique in space and time came from demographic groups that had a lower percentage of male individuals, a higher percentage of people with income of less than $100,000, a lower percentage of Asian population, and a higher percentage of non-Asian people of color. In comparison, there was not a clear trend between the percentage of white population and *k* value. Similar patterns were also observed when *k* was estimated at other time intervals.

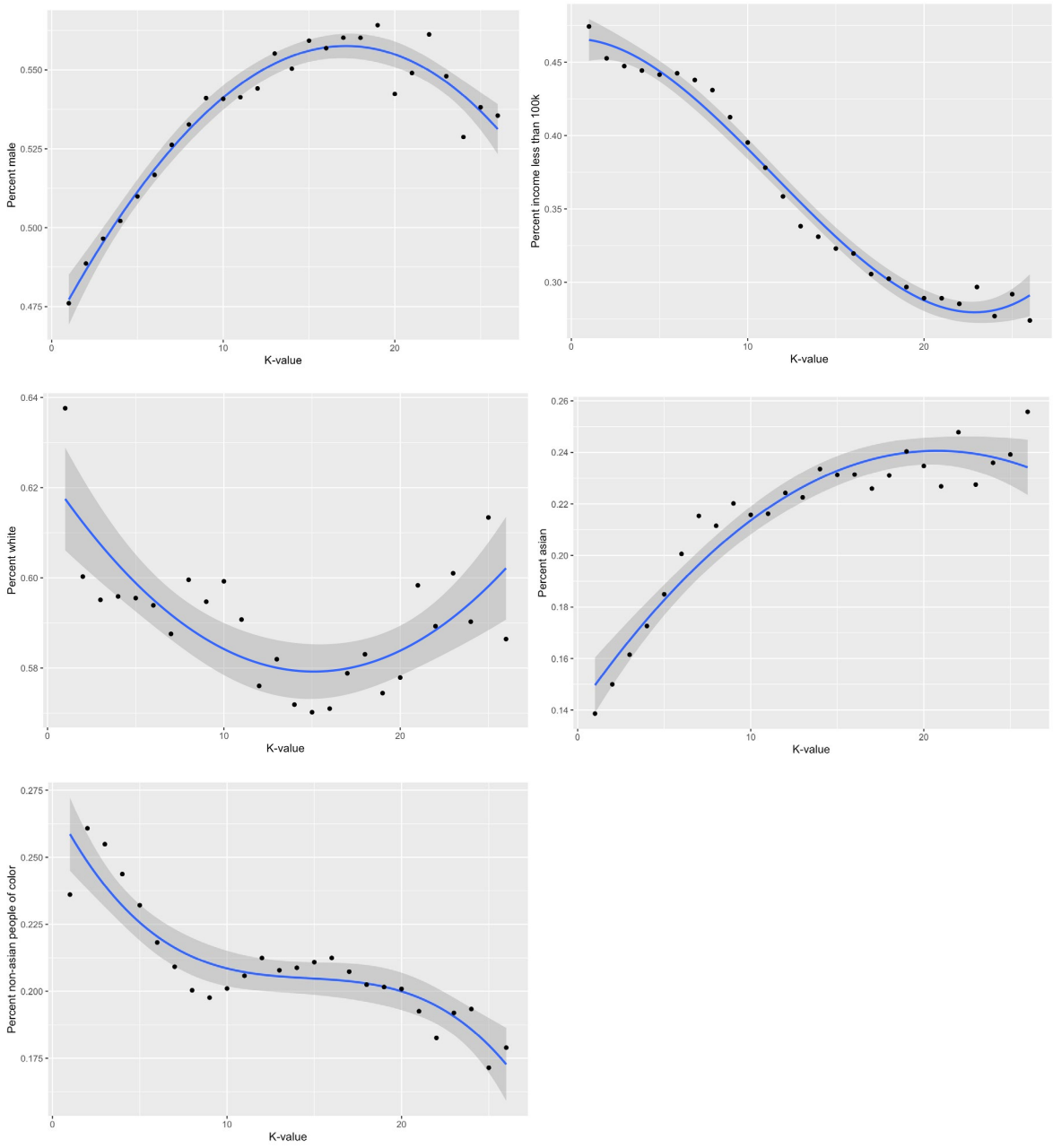

Figure 5.2 Relationship between SE and *k* value achieved at CBG and a 1-hour interval

Overall, the results from the regression models confirmed the patterns observed in figure 5.1 and 5.2. Table 5.2 shows that Model 1 estimated the associations between *k* and BE variables. The results were consistent with the observations shown in figure 5.1. A 1 percent increase in population density was associated with a 0.12 unit decrease in *k*. A 1 percent increase in per capita income was associated with a 0.34 unit increase in *k*. A 1 percent increase in the

percentage of residential land use was associated with a 8.78 unit decrease in *k*. Similarly, Model 2 estimated the associations between *k* and SE variables. The results were also mostly consistent with the patterns shown in figure 5.2 except that a 1 percent increase in the percentage of males was associated with a 0.18 unit decrease in *k* instead of an increase. Model 3 estimated the associations using both BE and SE variables. The only estimate that differed from previous model results was that the percentage of Asians was negatively associated with *k* after other confounding factors were controlled.

Table 5.2 Associations between *k* and BE and SE variables from Poisson regression models

	Model 1		**Model 2**		**Model 3**	
	Est.	*Sig.*	*Est.*	*Sig.*	*Est.*	*Sig.*
(Intercept)	0.06		1.16	***	0.07	
log(density)	-0.12	***			-0.06	***
log(per capita income)	0.34	***			0.24	***
log(percent residential)	-8.78	***			-8.51	***
log(percent male)			-0.18	***	-0.14	***
log(percent asian)			0.06	***	-0.05	***
log(percent napoc)			-0.28	***	-0.03	***
log(percent income < 100k)			-0.68	***	-0.25	***
AIC	93179		131924		90320	
Observations	19079		18809		18809	

Note 1: 0 '***' 0.001 '**' 0.01 '*'

CHAPTER 6. DISCUSSION

Previous studies using GPS data and surveillance data have demonstrated the uniqueness of human mobility traces at high spatiotemporal resolution. This study showed that even at relatively low spatial and temporal resolution—points at the CBG level and a 1-hour time interval from a large regional travel survey—traces of individuals' movements remain highly unique. While the traces alone may not be enough to re-identify people and reveal sensitive information, they can be used to link or match the data to other datasets. For example, Sweeney (1997) used ZIP code, sex, and birth date information and linked the 1997 voter's list for Cambridge, Massachusetts, and local public medical records.

This study further extended previous studies by examining the variations in re-identification risk of mobility data that accrues as a result of variations in urban areas and population segments. The results showed that travelers in more residential neighborhoods, i.e., neighborhoods with higher percentages of residential land use or higher residential population densities, and neighborhoods with lower per capita income are more likely to be re-identified. The relationship between the uniqueness of traces and percentage of residential land use point to the role of one's home as an origin and destination of mobility traces, and to the spatial predominance of a relatively suburban distribution of residents in single family housing within Seattle. Similarly, areas of relatively low per capita income are not likely to provide destinations that attract travelers and therefore are more distinctive than other areas for mobility traces.

The finding that uniqueness rises with population density of residents, however, could be counterintuitive, as one would expect high residential population density to confer a measure of anonymity to travelers. There could be two possible explanations. First, the result may reflect Seattle's unique distribution of population density across CBGs, where CBGs in the downtown area are mostly commercial lands and have very low numbers of residents and residential population density. Thus, the built environment characteristics captured by the residential population density were similar to those captured by the percentage of residential land use and may not have reflected the housing or building density of the city. Second, as shown in figure 5.2, there could be a nonlinear relationship between *k*-values and residential population densities that was not fully captured by the current model specification.

Besides BE variables, the study also found that travelers whose mobility traces were more unique than others tended to have higher percentages of male, non-white, and lower

income populations. For male travelers, Kwan (2000) found that they have fewer space-time constraints and therefore larger activity space than female travelers, which could lead to more trips to places that are less visited by others. For non-white and lower income travelers, no studies have explained why their mobility traces are more unique than others. One possible explanation may be that white and higher income travelers aggregate origins and destinations such as downtown work and retail locations in mobility traces more than their non-white and lower income counterparts, and they may be less likely to visit places visited by non-white and lower income travelers or may be more likely to visit those places at different times than those two groups. In addition, places visited by white and higher income travelers are also frequently visited by non-white and lower income travelers.

The results have two implications for data collection, processing, and publication practices. First, generalization or aggregation is a common strategy used by public agencies to de-identify individuals in a dataset. However, it is often tricky to determine the optimal level and the right fields for aggregation, as over-aggregation may lead to excessive information loss, and under-aggregation may be insufficient to achieve a desired level of *k*-anonymity. By showing the structural variations in re-identification risk, this study suggests the possibility of de-identifying data according to variation in urban areas and population segments, which could help to reduce information loss and offer greater potential protection of the privacy of individuals whose data records are more unique than others.

Second, the methodologies applied in this study can help detect and mitigate algorithmic biases in current data practices. This study found that travelers whose spatiotemporal traces were unique from others consisted of higher percentages of non-white persons and lower income populations, which may reflect the symptom of existing geographical inequality. Because these population groups also tend to be underrepresented in datasets, it makes them more vulnerable to re-identification risk than others. When oversampling strategies are designed to account for these underrepresented population groups, it is important to consider the distribution of both residential and travel locations.

CHAPTER 7. CONCLUSION AND FUTURE DIRECTIONS

Using a large regional travel survey, this project examined the variations in re-identification risk measured by *k*-anonymity in different urban areas and population segments. The results showed that even at relatively low spatiotemporal resolution, such as CBG level and a 1-hour interval, individuals' mobility traces through time and space were still highly unique. The study found that travelers in neighborhoods with higher percentages of residential land use, higher residential population densities, and lower per capita income were more likely to be re-identified. Travelers whose mobility traces were more unique than others also consisted of higher percentages of male, Asian, NAPOC, and lower income populations.

The findings can be applied in several ways in the different stages of the data cycle. First, in the early stages of technology procurement, installation, and collection, the findings can be integrated into the current surveillance ordinance act and the privacy impact assessment in the City of Seattle and used to identify and avoid urban areas that are vulnerable to privacy attacks. Second, in the data processing stage, the findings can be used to determine optimal strategies of de-identifying data by different urban areas and population segments, as well as to detect and mitigate potential algorithmic biases in the data collection process. Finally, in the data publishing stage, the findings can be used to guide restrictions on publishing data collected from urban areas or social groups with higher re-identification risks.

One limitation of the study is that the findings reflect the travel behavior patterns of only the surveyed participants and cannot be directly generalized to the entire region or other metropolitan areas. For future studies, public agencies will benefit from analyzing surveys of multiple years and larger geographical extent to confirm the findings. It would also be helpful to test the suggested aggregation schema to evaluate the effectiveness of such schema in terms of the tradeoffs between information loss and re-identification risk reduction. Currently, there are three common privacy-preserving approaches: generalization, suppression, and differential privacy with synthetic data. Generalization replaces a value with a more general value (e.g., aggregating survey results to gender or race). Suppression removes a sensitive value from the original dataset (e.g., replacing race with "xxx") (Sweeney, 2002). Differential privacy describes a family of methodologies, such as the laplace mechanism, the exponential mechanism, and the sparse vector technique, that protect individuals from any additional harm that they might receive as a result of data being in a private dataset that they would not have received had the data not

been part of the dataset (Dwork and Roth, 2014). The next step of this study will test each approach for their capacity to address the urban and social heterogeneity of re-identification risks while minimizing the tradeoffs of information loss in the process.

CHAPTER 8. REFERENCES

American Community Survey 2017 United State Census Bureau. https://www.census.gov/programs-surveys/acs

Chaix, B., Meline, J., Duncan, S., Merrien, C., Karusisi, N., Perchoux, C., Lewin, A., Labadi, K., Kestens, Y., 2013. GPS tracking in neighborhood and health studies: a step forward for environmental exposure assessment, a step backward for causal inference? Health Place 21, 46–51.

Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp. Res. Part C Emerg. Technol. 68, 285–299.

Dalenius, T., 1986. Finding a needle in a haystack or identifying anonymous census records. J. Off. Stat. 2, 329.

De Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., Blondel, V.D., 2013. Unique in the crowd: The privacy bounds of human mobility. Sci. Rep. 3, 1376.

De Mulder, Y., Danezis, G., Batina, L., Preneel, B., 2008. Identification via location-profiling in GSM networks, in: Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society. pp. 23–32.

Dwork, C., Roth, A., 2014. The algorithmic foundations of differential privacy. Found. Trends® Theor. Comput. Sci. 9, 211–407.

Gao, J., Sun, L., Cai, M., 2019. Quantifying privacy vulnerability of individual mobility traces: A case study of license plate recognition data. Transp. Res. Part C Emerg. Technol. 104, 78–94. https://doi.org/10.1016/j.trc.2019.04.022

Kondor, D., Hashemian, B., de Montjoye, Y.-A., Ratti, C., 2018. Towards matching user mobility traces in large-scale datasets. IEEE Trans. Big Data.

Kwan, M.-P., 2000. Gender differences in space-time constraints. Area 32, 145–156. https://doi.org/10.1111/j.1475-4762.2000.tb00125.x

Nissenbaum, H., 2011. A contextual approach to privacy online. Daedalus 140, 32–48.

Pellungrini, R., Pappalardo, L., Pratesi, F., Monreale, A., 2017. A data mining approach to assess privacy risk in human mobility data. ACM Trans. Intell. Syst. Technol. TIST 9, 1–27.

Puget Sound Regional Council, 2018 Regional Travel study. https://www.psrc.org/media/3631

Stead, D., 2001. Relationships between land use, socioeconomic factors, and travel patterns in Britain. Environ. Plan. B Plan. Des. 28, 499–528.

Sweeney, L., 2002. k-anonymity: A model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10, 557–570.

Sweeney, L., 1997. Guaranteeing anonymity when sharing medical data, the Datafly System. Proc. AMIA Annu. Fall Symp. 51–55.

Thompson, S., Warzel, C., 2019. Twelve Million Americans Were Tracked Through Their Phones. N. Y. Times.